June 1, 2020

Project Status
Colorado OER Council Grant (AY 2019-20)

Craig Trumbo, Journalism and Media Communication
JTC270 Quantitative Analysis in Journalism and Media

This deposit provides a status report and materials collection for this OER project. Work toward the OER resource was started in fall 2019 and continued until March 2020. At that time effort was diverted to revising the course to online delivery for the balance of the semester. The faculty member was awarded a sabbatical for the 2020-21 AY. The OER will be brought to full completion in fall 2021. The following materials are provided here:

- OER Proposal

- Lecture slides and lab exercises. There were remade from scratch to entirely disconnect any previously used textbook material. Exams were also remade (not provided here). The Canvas site for the course was remade to include external video links to support lecture topics and software used in lab.

- A described in the proposal, a collection of news articles pertaining to scientific studies was started. These are a partial set of the collection for demonstration purposes. In many cases the original scientific reports behind the news articles were also collected. Within this material resides the examples of statistical principles and tools that will be integrated into the online text.

Name: Craig Trumbo, PhD
Department/program: Journalism and Media Communication
Date: 4/22/19

Course information:	**Information on the current textbook:**
Course: JTC270 Quantitative Analysis in Journalism and Media (and others)	Title: Social Statistics for a Diverse Society (7th Ed), Frankfort-Nachmias and Leon-Guerrero.
Total enrollment/semester: 30	Cost: $99 new on Amazon
Semester/s offered: F/S	Estimated savings: Up to $3,000 per semester, total to date of $30,000 for sections taught and scheduled next AY.

Would you be interested in working with an instructional designer and/or librarian? **Yes**

Project Overview

This OER application is based on a course that was established three years ago, JTC270 Quantitative Analysis in Journalism and Media. The mission of this course is to introduce early career undergraduates in various areas of communication (typically liberal arts) to fundamental quantitative concepts and basic statistics. This is done in a contextual manner that serves to enhance their interest and engagement. A statistics course is required for the major. Most majors take this one, as well as some non-majors.

Broadly speaking, the structural approach for the OER will be to introduce each major concept within the content of a contemporary issue. To do so, typically a quality news story or analysis is traced back through a press release, and then to an original study. Material is selected to specifically highlight a given concept (e.g., sampling). Learning flows from "how was this reported" to "how was this information provided" to "how was knowledge generated." Working in this manner can greatly enhance the ability of students to engage in thoughtful critique in a step-wise manner. The material, and process, are both very well suited to an OER approach.

Having been already partially engaged in developing these materials I presently have a collection of some 25 contemporary issue cases. Space prohibits elaboration, but there are of course several in the political domain, in health, and a range of social and environmental issues. The present text does a good job of contextualizing broadly within the social sciences, but the examples are dated and not directly of current interest to students. This OER support will allow me to consolidate and integrate these materials into open resource content in quantitative reasoning for this population of students.

This OER can be used to serve as the primary text in this course, but may also be used as supplementary material in a wide range of other courses in Journalism (advanced reporting, data journalism, in the professional master's program), and across the College of Liberal Arts. As such, the cost savings to students will be expanded beyond this single course. Demand for the course is also sufficient to potentially consider additional sections.

OER Plan Specifics

How do you plan to go about replacing the textbook in this course?

This OER textbook will be created. As noted above in the project overview there will be open access and fair use materials drawn into the course from outside sources. These will be uniquely incorporated, as the contextual framework, into my own narrative on the basic subject matter.

What format(s) and/or platform/s will be used (e.g., PDF/A, e-book, video, website, etc.)?

The platform will be based online using Canvas, to incorporate a variety of formats to include PDF, video, discussion, dynamic links to contextualizing material, and common learning support tools such as flash cards, practice quizzes, etc. In particular, brief lecture “outtakes” will be recorded, and I will produce a series of short, tightly focused demonstrations of formula and calculations. I would like to work with a course designer in order to optimize the presentation in Canvas (as opposed to the standard template). I would also like to take part in relevant campus training (e.g., Course Design and Development at TILT). As noted, this OER should be of interest to other CSU courses. Canvas as the platform will facilitate such sharing and collaboration.

How will students access the content? Will it follow universally accessible design principles?

The primary mode of access will be through a desktop browser. To the degree provided by the Canvas Student app, content will also be available on mobile devices. Standard design principles will be followed, with respect to content and structure. Any involvement by an instructional designer would be most welcome.

What problems do you anticipate?

There can be issues encountered with respect to the use of some material under fair use. However, many quality publications are explicit in their support of educational use (e.g., *The New York Times*, which provides free digital access for campus). Given the rich variety of such sources as contextualizing material, this should be manageable. Otherwise it’s a common experience that tasks take longer than planned. Since I am teaching this class next academic year, the additional instructional time provided by this award should allow efficient dovetailing.

How will relevant copyright issues be addressed? Will you be using a Creative Commons license?

Contextualizing material can be found that will be available under educational fair use. The basic concepts and equations will be originally made by the author, based on long-standing material that is in the public domain. Also, I will collaborate as provided with library professionals to address copyright concerns, and to implement the project under a Creative Commons license.

What are your anticipated outcomes and how will you know they were achieved?

The desired outcome is to make this topic, on this introductory level, more compelling and accessible specifically to students with a Liberal Arts background who are pursuing careers in the various areas of public communication (from journalism to marketing, etc.). Having taught this material for this audience over several years I am familiar with the barriers students face. With this perspective, it will be possible going forward with the new material to make comparisons based on my experience and student assessment/feedback to gauge effectiveness. Another outcome goal is for this OER to be adopted as supplementary material in other CSU courses, which can be promoted and observed.

Level of grant requested and plan for using funds.

$4000 will provide a one-course teaching load reduction during the 2019-20 AY.

What is your plan for sustaining the use of OER adapted or created beyond initial use?

As mentioned in the overview, the OER will provide the foundation for an established on-going course. I would like to thereafter expand its use through development of a CSU Online course, and promotion of use as supplementary material in other courses. The content of the contextualizing material will be a living component of the OER, as this can be updated with contemporary material going forward.

When do you plan to implement the content adopted, adapted, or created via this proposal?

The OER will be developed concurrently with the course in fall 2019 and at least partially implemented in the spring 2020 semester. Full implementation thereafter, and possible expansion online.

JTC270 **Quantitative Analysis in Journalism and Media,** Spring 2020

Class Hours **Section 1** Lecture Tuesday 11:00 - 12:15 Clark C363
Lab Thursday 11:00 - 12:15 Clark C141

Section 2 Lecture Tuesday 9:30 - 10:45 Clark A07
Lab Thursday 9:30 - 10:45 Clark C141

Dr. Craig Trumbo Office – Before/after class open, Hartshorn 117 (by appointment)
Phone – (970) 491-2077 email – via Canvas

Intended Learning Outcomes – The course is based on three objectives. **First**, students will acquire an understanding of the fundamental concepts in basic statistics, especially approaches to data acquisition, sampling, data description, data visualization, estimation, and the testing of differences and associations. This objective will be met through lectures, demonstrations, and hands-on exercises using computational software rather than hand calculation. **Second**, students will gain a contextualizing perspective on the way that quantitative thinking and statistical analyses are used in the various domains of their professional practice. will especially emphasize the use of statistical analysis in public relations work (e.g., campaign planning and evaluation) and in journalistic reporting (e.g., in health and politics). **Third**, students will learn how to integrate these skills and contextual insights into their own work.

Required Text

Frankfort-Nachmias, Chava and Anna Leon-Guerrero. 2015. *Social Statistics for a Diverse Society.* 7th ed. Thousand Oaks, CA: Pine Forge Press. ISBN-13: 978-1483333540

Policy

Academic Integrity – This course will adhere to the CSU Academic Integrity Policy as found in the General Catalog - 1.6, pages 7-9 and the Student Conduct Code . At a minimum, violations will result in a grading penalty in this course and a report to the Office of Conflict Resolution and Student Conduct Services. In addition to the material just referenced, academic integrity in this course will include: not receiving or providing assistance from others on exams or individual homework assignments, full participation in group activities, and appropriate citation of other's work in print or on-line. Also, support materials provided for this class (e.g., lecture notes, handouts, auxiliary readings) may not be shared or sold outside of the course as this is a violation of the Academic Integrity Policy as well as copyright law.

Federal Credit Hour Rule – You should anticipate spending on average a minimum of six hours per week on text readings, assignments, and auxiliary materials.

Attendance – The material in this class is ***strongly cumulative.*** It is not possible to "catch up" on this material if one falls behind. The only excuse to reschedule an exam that will be accepted prior to the exam is for a university sanctioned event. A letter must be provided. The only excuse to reschedule a missed exam will be for an extenuating circumstance with verifiable documentation. A letter must be provided from Student Case Management. No exceptions.

Grading – There will be three exams that are comprised of an objective component and a lab component. There will also be lab quizzes. Assignment specifics will be provided in class and on Canvas. Final grades will use plus-minus, with the exclusion of the C- as per university policy. Note that since this is a JTC class majors may not receive credit for graduation for grades less than C. The statistics requirement may also be satisfied by non-JTC classes in which this policy does not apply (e.g. STAT 100, 201, 205, 301 and others in various departments).

Points	Element
900	Exam 1-3 = 300 pts each (150 objective/150 lab)
100	Lab quizzes (9 at 10-15 pts each)

Total Percentage	Final Grade
95.0 - 100.0	A
90.0 - 94.9	A-
85.0 - 89.9	B+
80.0 - 84.9	B
75.0 - 79.9	B-
70.0 - 74.9	C+
65.0 - 69.9	C
60.0 - 64.9	D
00.0 - 59.9	F

Subject to revision if necessary

Week	DATE	TOPIC	READINGS
1	T 1/21	Course intro	
	R 1/23	No Lab: Complete math review, Stata videos	Review videos on Canvas
2	T 1/28	Variables, measurement, frequencies, graphing	Frankfort Ch 1-3
	R 1/30	Lab 1: Measurement, freq tables	Review videos on Canvas
3	T 2/4	Central Tendency, Variability	Frankfort Ch 4-5
	R 2/6	Lab 2: Reporting and plotting summary statistics	Review videos on Canvas
4	T 2/11	The Normal Distribution, Areas	Frankfort Ch 6
	R 2/13	Lab 3: Findings areas under the curve	Review videos on Canvas
5	T 2/18	Lecture: Concept Review, Q&A	
	R 2/20	Lab: Practice Exercises	
6	T 2/25	Exam 1	
	R 2/27	Exam 1	
7	T 3/3	Probability, Estimation, CIs	Frankfort Ch 7-8
	R 3/5	Lab 4 Finding CIs	Review videos on Canvas
8	T 3/10	Testing hypotheses about proportion	Frankfort Ch 8-9
	R 3/12	Lab 5 Testing proportions	Review videos on Canvas
9	T 3/24	Testing hypotheses about means	Frankfort Ch 9
	R 3/26	Lab 6 Testing means	Review videos on Canvas
10	T 3/31	Lecture: Concept Review, Q&A	
	R 4/2	Lab: Practice Exercises	
11	T 4/7	Exam 2	
	R 4/9	Exam 2	
12	T 4/14	Chi-Square Test	Frankfort Ch 10-11
	R 4/16	Lab 7 Chi-square	Review videos on Canvas
13	T 4/21	ANOVA	Frankfort Ch 12
	R 4/23	Lab 8, ANOVA	Review videos on Canvas
14	T 4/28	Correlation	Frankfort Ch 13
	R 4/30	Lab 9: Correlation	Review videos on Canvas
15	T 5/5	Lecture: Concept Review, Q&A	
	R 5/7	Lab: Practice Exercises	
16	R 5/14	Final Thursday 8:20 pm	

JTC270: Quantitative thinking

JTC270 approach:

Accessible and contextually relevant

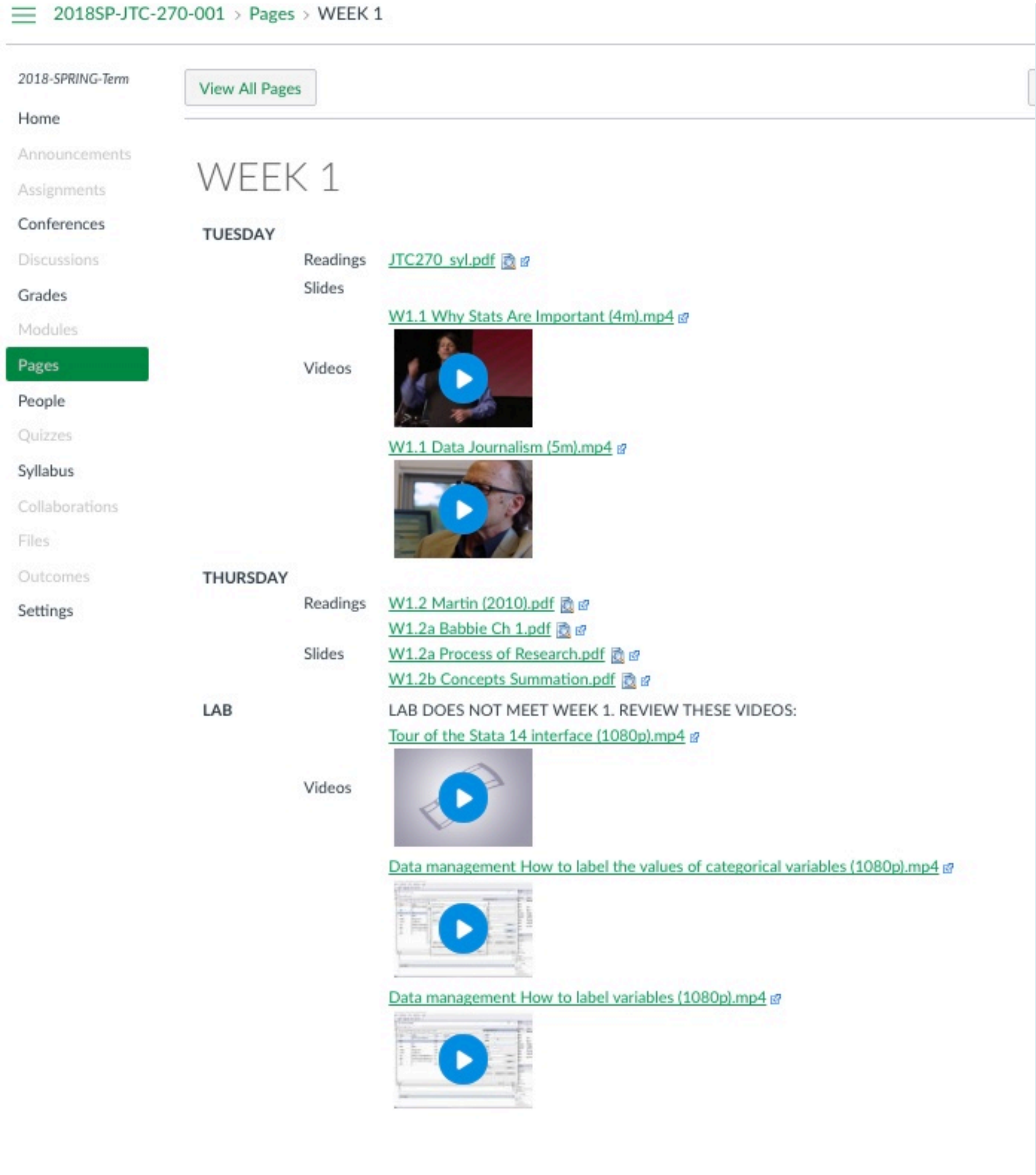
* Concepts over calculations
* Use of software
* Text using contemporary social issues
* News items used for examples
* Examples/exercises from actual research

JTC270: Semester Plan and Resources

Canvas: Organized by week

> Lecture:
> New topics introduced on Tuesday
> readings, handouts, slides, concept videos

> Lab:
> Matching Lab Thursday (w/quiz due Friday)
> exercises, instructional videos (required)



2018SP-JTC-270-001 > Pages > WEEK 1

2018-SPRING-Term

Home
Announcements
Assignments
Conferences
Discussions
Grades
Modules
Pages
People
Quizzes
Syllabus
Collaborations
Files
Outcomes
Settings

View All Pages

WEEK 1

TUESDAY

Readings JTC270_syl.pdf

Slides

Videos W1.1 Why Stats Are Important (4m).mp4

W1.1 Data Journalism (5m).mp4

THURSDAY

Readings W1.2 Martin (2010).pdf

W1.2a Babbie Ch 1.pdf

Slides W1.2a Process of Research.pdf

W1.2b Concepts Summation.pdf

LAB

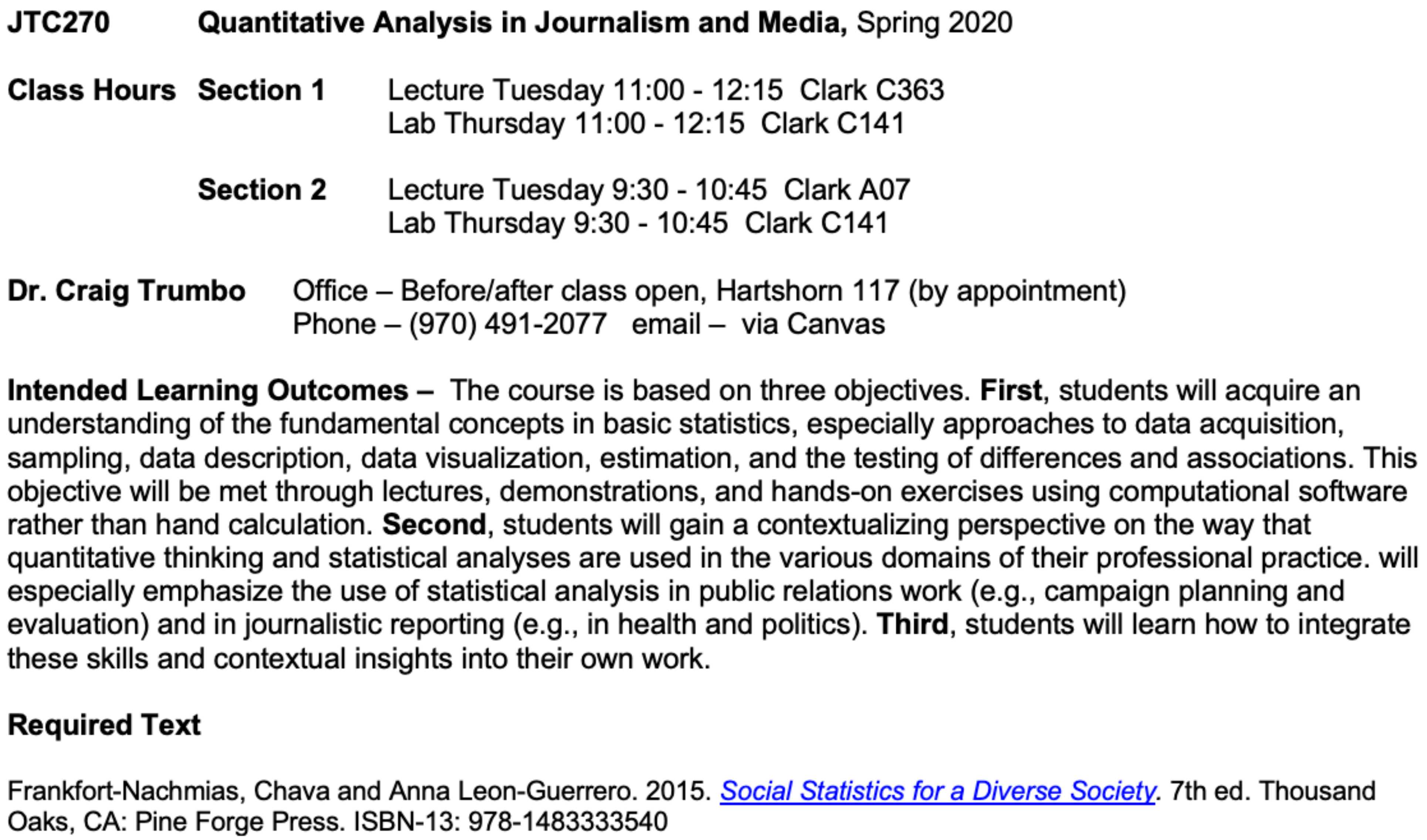
LAB DOES NOT MEET WEEK 1. REVIEW THESE VIDEOS:

Videos Tour of the Stata 14 interface (1080p).mp4

Data management How to label the values of categorical variables (1080p).mp4

Data management How to label variables (1080p).mp4

JTC270: Syllabus…



JTC270 **Quantitative Analysis in Journalism and Media,** Spring 2020

Class Hours **Section 1** Lecture Tuesday 11:00 - 12:15 Clark C363
Lab Thursday 11:00 - 12:15 Clark C141

Section 2 Lecture Tuesday 9:30 - 10:45 Clark A07
Lab Thursday 9:30 - 10:45 Clark C141

Dr. Craig Trumbo Office – Before/after class open, Hartshorn 117 (by appointment)
Phone – (970) 491-2077 email – via Canvas

Intended Learning Outcomes – The course is based on three objectives. **First**, students will acquire an understanding of the fundamental concepts in basic statistics, especially approaches to data acquisition, sampling, data description, data visualization, estimation, and the testing of differences and associations. This objective will be met through lectures, demonstrations, and hands-on exercises using computational software rather than hand calculation. **Second**, students will gain a contextualizing perspective on the way that quantitative thinking and statistical analyses are used in the various domains of their professional practice. will especially emphasize the use of statistical analysis in public relations work (e.g., campaign planning and evaluation) and in journalistic reporting (e.g., in health and politics). **Third**, students will learn how to integrate these skills and contextual insights into their own work.

Required Text

Frankfort-Nachmias, Chava and Anna Leon-Guerrero. 2015. *Social Statistics for a Diverse Society*. 7th ed. Thousand Oaks, CA: Pine Forge Press. ISBN-13: 978-1483333540

Policy

Academic Integrity – This course will adhere to the CSU Academic Integrity Policy as found in the General Catalog - 1.6, pages 7-9 and the Student Conduct Code . At a minimum, violations will result in a grading penalty in this course and a report to the Office of Conflict Resolution and Student Conduct Services. In addition to the material just referenced, academic integrity in this course will include: not receiving or providing assistance from others on exams or individual homework assignments, full participation in group activities, and appropriate citation of other's work in print or on-line. Also, support materials provided for this class (e.g., lecture notes, handouts, auxiliary readings) may not be shared or sold outside of the course as this is a violation of the Academic Integrity Policy as well as copyright law.

Federal Credit Hour Rule – You should anticipate spending on average a minimum of six hours per week on text readings, assignments, and auxiliary materials.

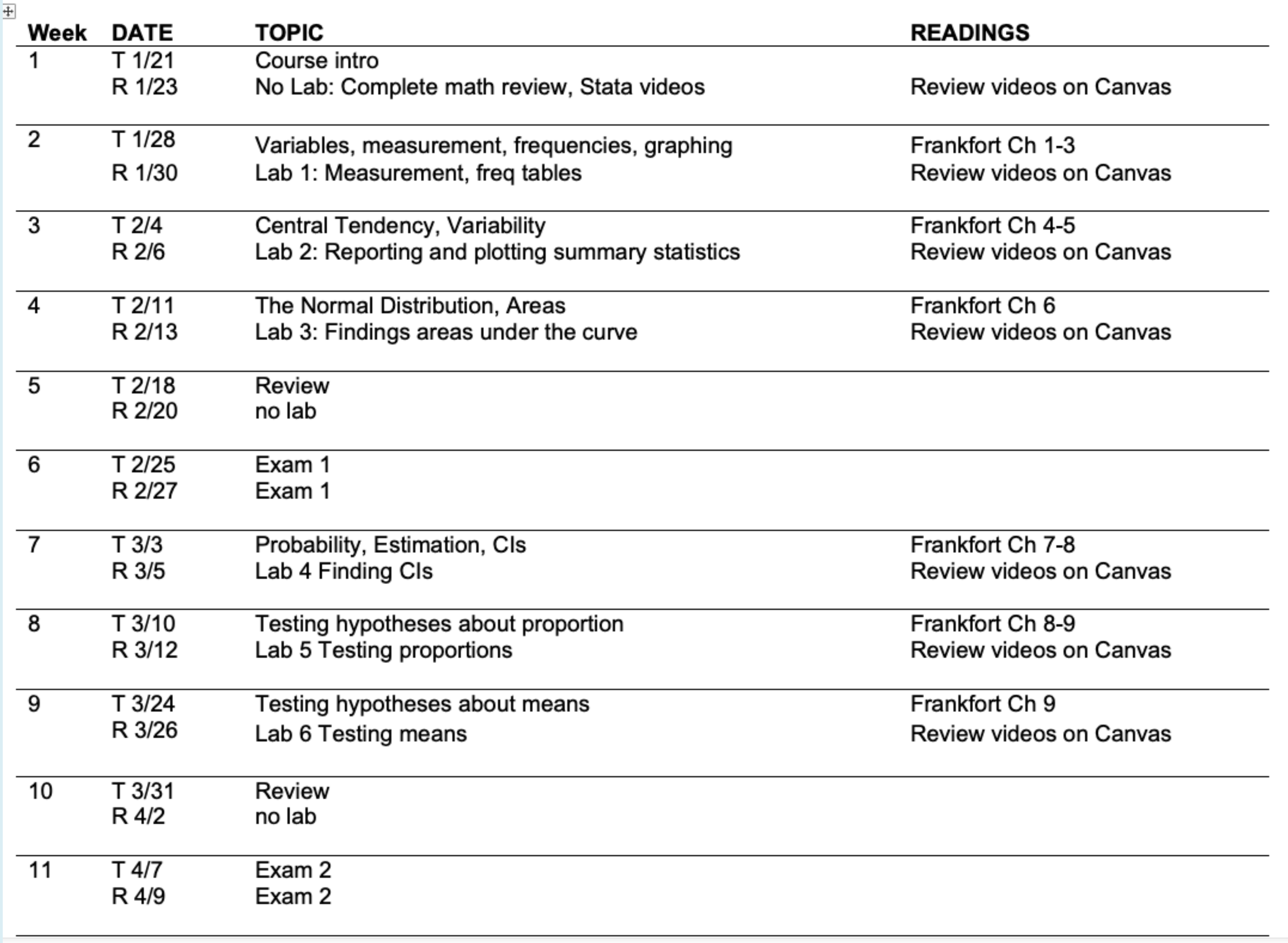
Attendance – The material in this class is ***strongly cumulative.*** It is not possible to "catch up" on this material if one falls behind. The only excuse to reschedule an exam that will be accepted prior to the exam is for a university sanctioned event. A letter must be provided. The only excuse to reschedule a missed exam will be for an extenuating circumstance with verifiable documentation. A letter must be provided from Student Case Management. No exceptions.

Grading – There will be three exams that are comprised of an objective component and a lab component. There will also be lab quizzes. Assignment specifics will be provided in class and on Canvas. Final grades will use plus-minus, with the exclusion of the C- as per university policy. Note that since this is a JTC class majors may not receive credit for graduation for grades less than C. The statistics requirement may also be satisfied by non-JTC classes in which this policy does not apply (e.g. STAT 100, 201, 205, 301 and others in various departments).

Points	Element
900	Exam 1-3 = 300 pts each (150 objective/150 lab)
100	Lab quizzes (9 at 10-15 pts each)

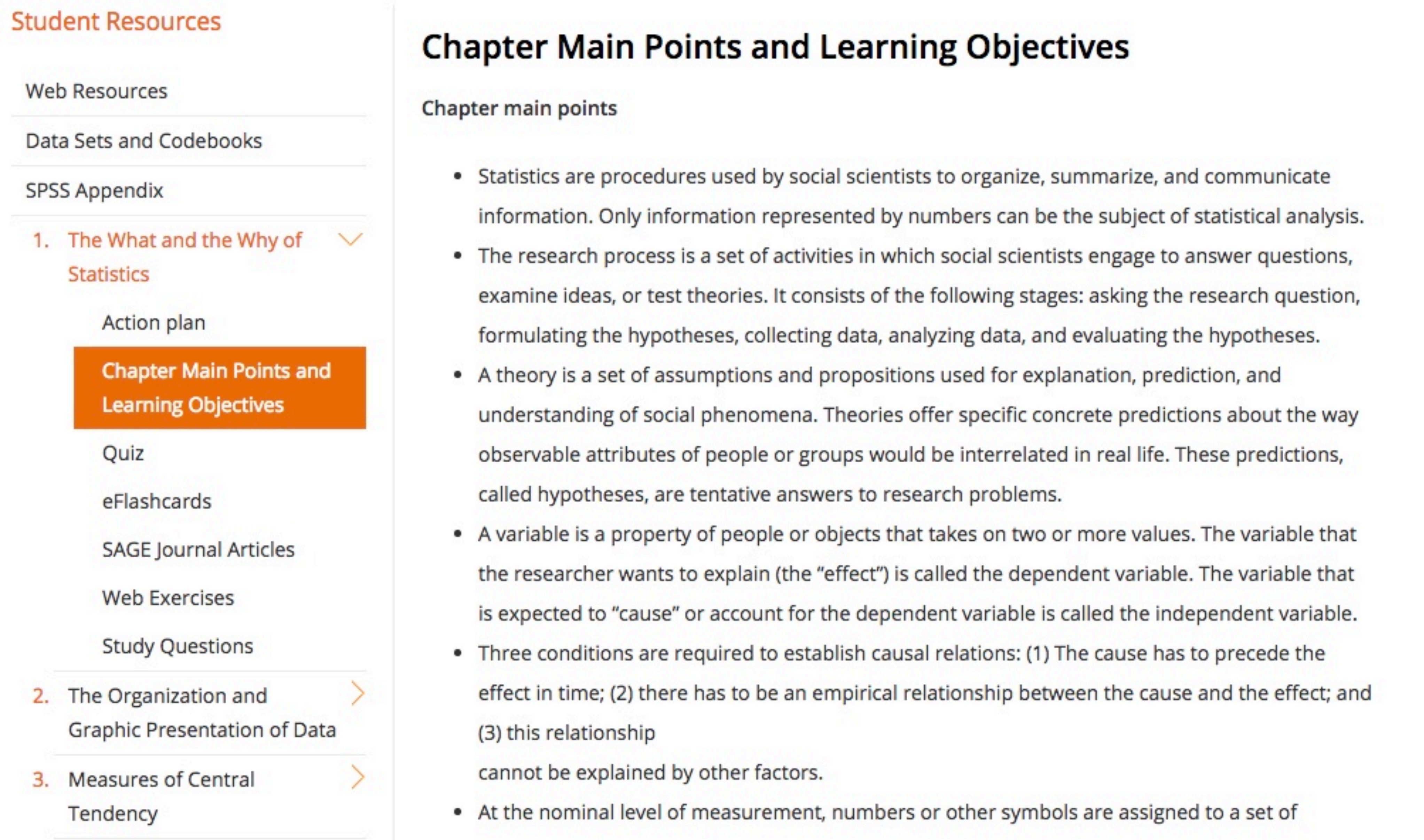
Total Percentage	Final Grade
95.0 - 100.0	A
90.0 - 94.9	A-
85.0 - 89.9	B+
80.0 - 84.9	B
75.0 - 79.9	B-
70.0 - 74.9	C+
65.0 - 69.9	C
60.0 - 64.9	D
00.0 - 59.9	F

JTC270: Syllabus…



Week	DATE	TOPIC	READINGS
1	T 1/21 R 1/23	Course intro No Lab: Complete math review, Stata videos	Review videos on Canvas
2	T 1/28 R 1/30	Variables, measurement, frequencies, graphing Lab 1: Measurement, freq tables	Frankfort Ch 1-3 Review videos on Canvas
3	T 2/4 R 2/6	Central Tendency, Variability Lab 2: Reporting and plotting summary statistics	Frankfort Ch 4-5 Review videos on Canvas
4	T 2/11 R 2/13	The Normal Distribution, Areas Lab 3: Findings areas under the curve	Frankfort Ch 6 Review videos on Canvas
5	T 2/18 R 2/20	Review no lab	
6	T 2/25 R 2/27	Exam 1 Exam 1	
7	T 3/3 R 3/5	Probability, Estimation, CIs Lab 4 Finding CIs	Frankfort Ch 7-8 Review videos on Canvas
8	T 3/10 R 3/12	Testing hypotheses about proportion Lab 5 Testing proportions	Frankfort Ch 8-9 Review videos on Canvas
9	T 3/24 R 3/26	Testing hypotheses about means Lab 6 Testing means	Frankfort Ch 9 Review videos on Canvas
10	T 3/31 R 4/2	Review no lab	
11	T 4/7 R 4/9	Exam 2 Exam 2	

JTC270 Resources: Online Text



Student Resources

- Web Resources
- Data Sets and Codebooks
- SPSS Appendix
- 1. The What and the Why of Statistics
  - Action plan
  - Chapter Main Points and Learning Objectives
  - Quiz
  - eFlashcards
  - SAGE Journal Articles
  - Web Exercises
  - Study Questions
- 2. The Organization and Graphic Presentation of Data
- 3. Measures of Central Tendency

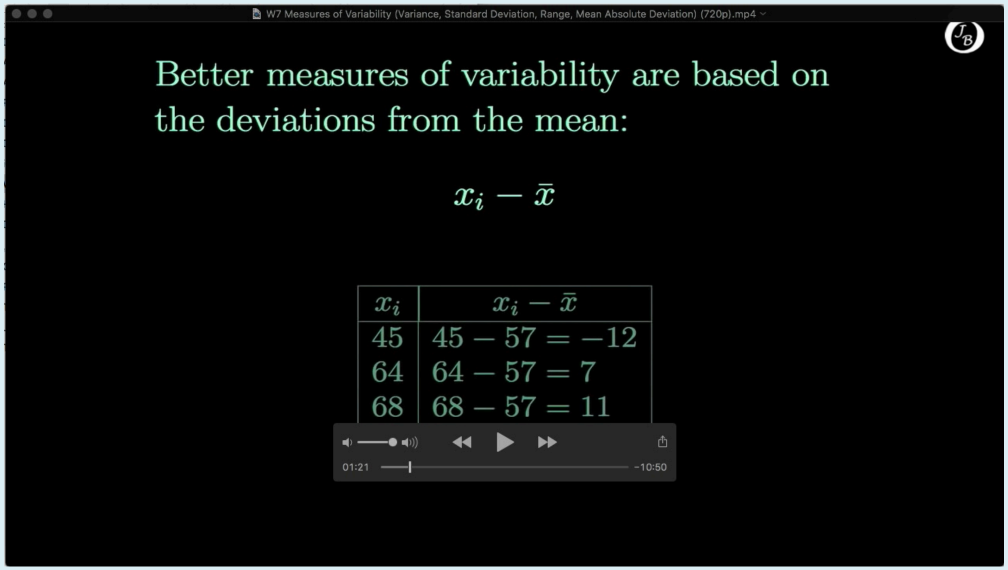
Chapter Main Points and Learning Objectives

Chapter main points

- Statistics are procedures used by social scientists to organize, summarize, and communicate information. Only information represented by numbers can be the subject of statistical analysis.
- The research process is a set of activities in which social scientists engage to answer questions, examine ideas, or test theories. It consists of the following stages: asking the research question, formulating the hypotheses, collecting data, analyzing data, and evaluating the hypotheses.
- A theory is a set of assumptions and propositions used for explanation, prediction, and understanding of social phenomena. Theories offer specific concrete predictions about the way observable attributes of people or groups would be interrelated in real life. These predictions, called hypotheses, are tentative answers to research problems.
- A variable is a property of people or objects that takes on two or more values. The variable that the researcher wants to explain (the "effect") is called the dependent variable. The variable that is expected to "cause" or account for the dependent variable is called the independent variable.
- Three conditions are required to establish causal relations: (1) The cause has to precede the effect in time; (2) there has to be an empirical relationship between the cause and the effect; and (3) this relationship

  cannot be explained by other factors.
- At the nominal level of measurement, numbers or other symbols are assigned to a set of

JTC270 Resources: Supplementary Videos

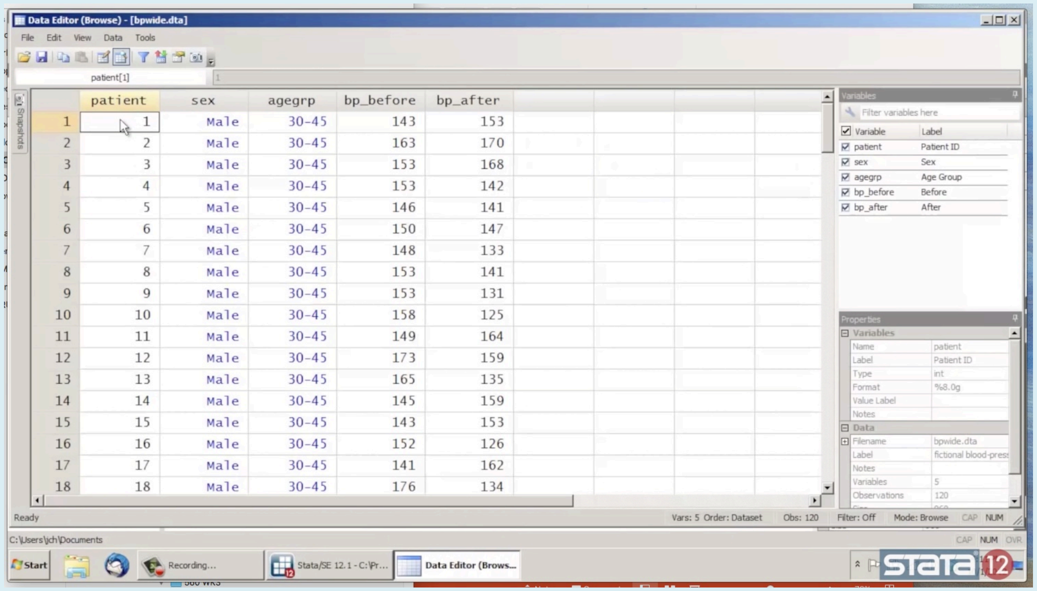

W7 Measures of Variability (Variance, Standard Deviation, Range, Mean Absolute Deviation) (720p).mp4
01:21
-10:50

Better measures of variability are based on the deviations from the mean:

$$x_i - \bar{x}$$

x_i	$x_i - \bar{x}$
45	$45 - 57 = -12$
64	$64 - 57 = 7$
68	$68 - 57 = 11$

Concepts



	patient	sex	agegrp	bp_before	bp_after
1	1	Male	30-45	143	153
2	2	Male	30-45	163	170
3	3	Male	30-45	153	168
4	4	Male	30-45	153	142
5	5	Male	30-45	146	141
6	6	Male	30-45	150	147
7	7	Male	30-45	148	133
8	8	Male	30-45	153	141
9	9	Male	30-45	153	131
10	10	Male	30-45	158	125
11	11	Male	30-45	149	164
12	12	Male	30-45	173	159
13	13	Male	30-45	165	135
14	14	Male	30-45	145	159
15	15	Male	30-45	143	153
16	16	Male	30-45	152	126
17	17	Male	30-45	141	162
18	18	Male	30-45	176	134

Variables

Filter variables here

☑ Variable	Label
☑ patient	Patient ID
☑ sex	Sex
☑ agegrp	Age Group
☑ bp_before	Before
☑ bp_after	After

Properties

Variables	
Name	patient
Label	Patient ID
Type	int
Format	%8.0g
Value Label	
Notes	
Data	
Filename	bpwide.dta
Label	fictional blood-press
Notes	
Variables	5
Observations	120

Ready

Vars: 5 Order: Dataset Obs: 120 Filter: Off Mode: Browse CAP NUM

C:\Users\jch\Documents

Software

JTC270: Semester at a glance

The nature of data and variables

“levels of measurement” continuous vs. discrete, variables, units of observation

Tools of basic description

Organizing descriptive statistics, “frequency tables”

Graphing data

Identification of the most common characteristic

“central tendency” mean – median – mode

Characterization of the diversity of data

“dispersion” variance – standard deviation – standard error

JTC270: Semester at a glance

The shape of data

distributions … and the normal distribution

The magic of sampling

Estimation and Inference

what it's all about, from sample to population

Testing hypotheses

Are two percentages really different?
Are two averages really different?
Are there really differences among 3+ averages?
Are two variables really associated?
Does one variable really predict another? Multiple predictors?

Reporting results

Lab will not meet this Thursday.

Use time to review all Canvas material for Week 1:

- WK1 Lecture:
  - Textbook chapter 1
  - Videos on statistics and data journalism
  - Semester Datasets:
    - Background on WNv dataset to be used for lecture examples
    - Background on Uranium Mine dataset for use in lab exercises

- WK1 Lab:
  - Textbook math review
  - Slideset on summation notation
  - 3 videos on summation notation
  - 3 videos introducing the software we’ll use (Stata)

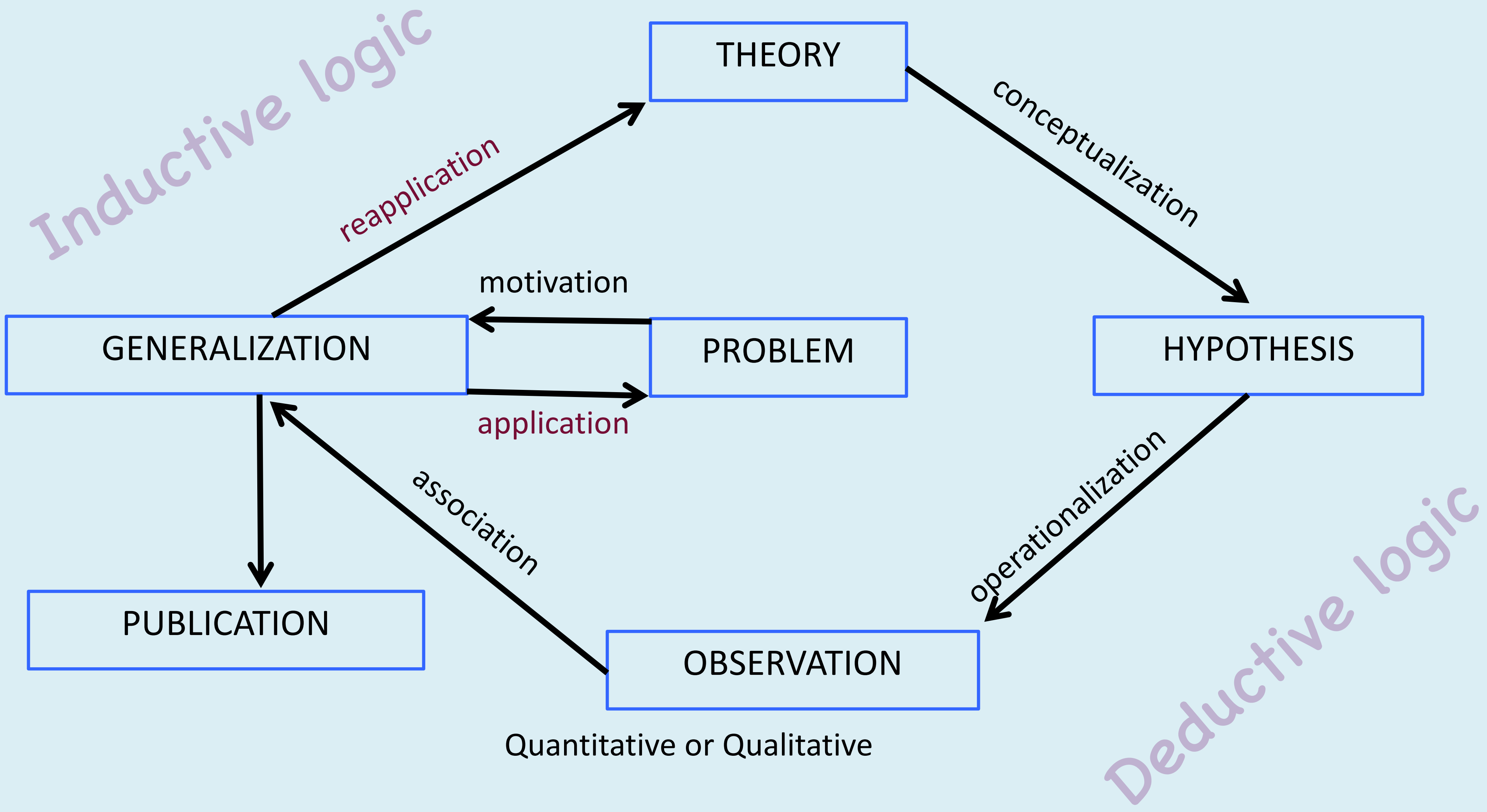
The Process of Research and the Role of Statistics

Research is:

- A *process of steps* used to *collect and analyze information* in order to *increase our knowledge* about a topic or issue

- Most research involves at least these steps:

  1. Posing a question
  2. Collecting data about the question
  3. Analyzing the data to answer the question
  4. Reporting the results



Empiricism: knowledge comes only or primarily from sensory experience.

Two General Types: Descriptive & Explanatory

– Descriptive research presents a profile of a group or describes a process, mechanism or relationship or presents basic background information or a context.

– Explanatory goes beyond simple description to model empirically the social phenomena under investigation. It involves theory testing or elaboration of a theory.

– Can be observational or Experimental

– To be causal, must satisfy three criteria:
  cause before effect
  an empirical relationship
  cannot be explained by other factors

Variables, Measurement, Graphing

Variable: a property

A property of people or objects that takes on two or more values. If categorical must be both exhaustive and mutually exclusive.

UOA = What we want to analyze, typically the rows in a data matrix. Variables: the columns.

Variable roles. Dependent, the variable to be explained
Independent, the variable that accounts for the dependent variable.

Variables come in three levels of measurement.

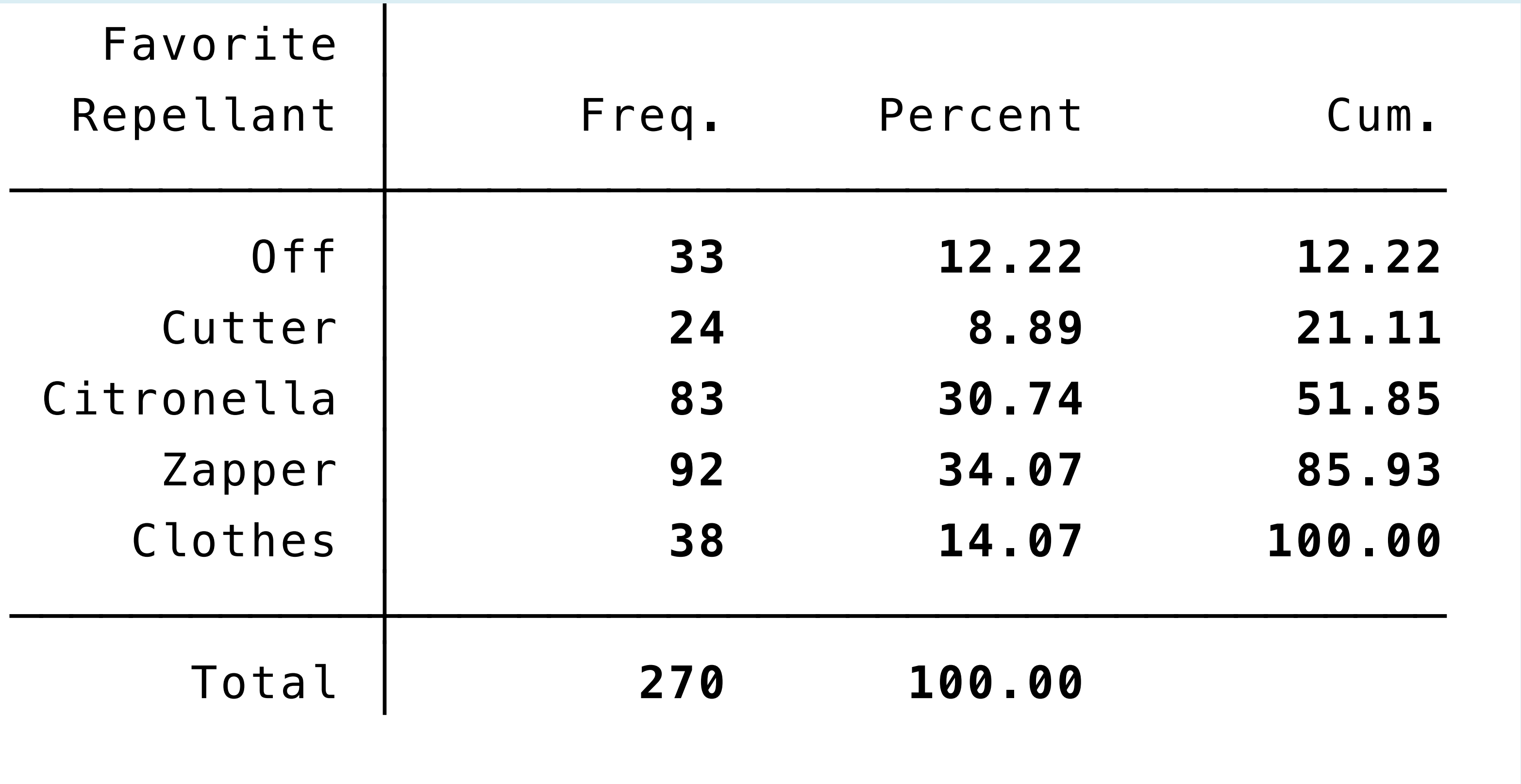
Levels of Measurement: Nominal

Mutually exclusive, exhaustive, no order.

What's your favorite mosquito repellant?

Respondent must pick just one (mutually exclusive)

Assume that these are the only possibilities (exhaustive)



```
   Favorite |
  Repellant |      Freq.     Percent        Cum.
------------+-----------------------------------
        Off |         33       12.22       12.22
     Cutter |         24        8.89       21.11
 Citronella |         83       30.74       51.85
     Zapper |         92       34.07       85.93
    Clothes |         38       14.07      100.00
------------+-----------------------------------
      Total |        270      100.00
```

Note: since this is not ordinal, the cumulative column is not used.

Why?

From a study of what influences people to self-protect from mosquitoes.

Levels of Measurement: Ordinal

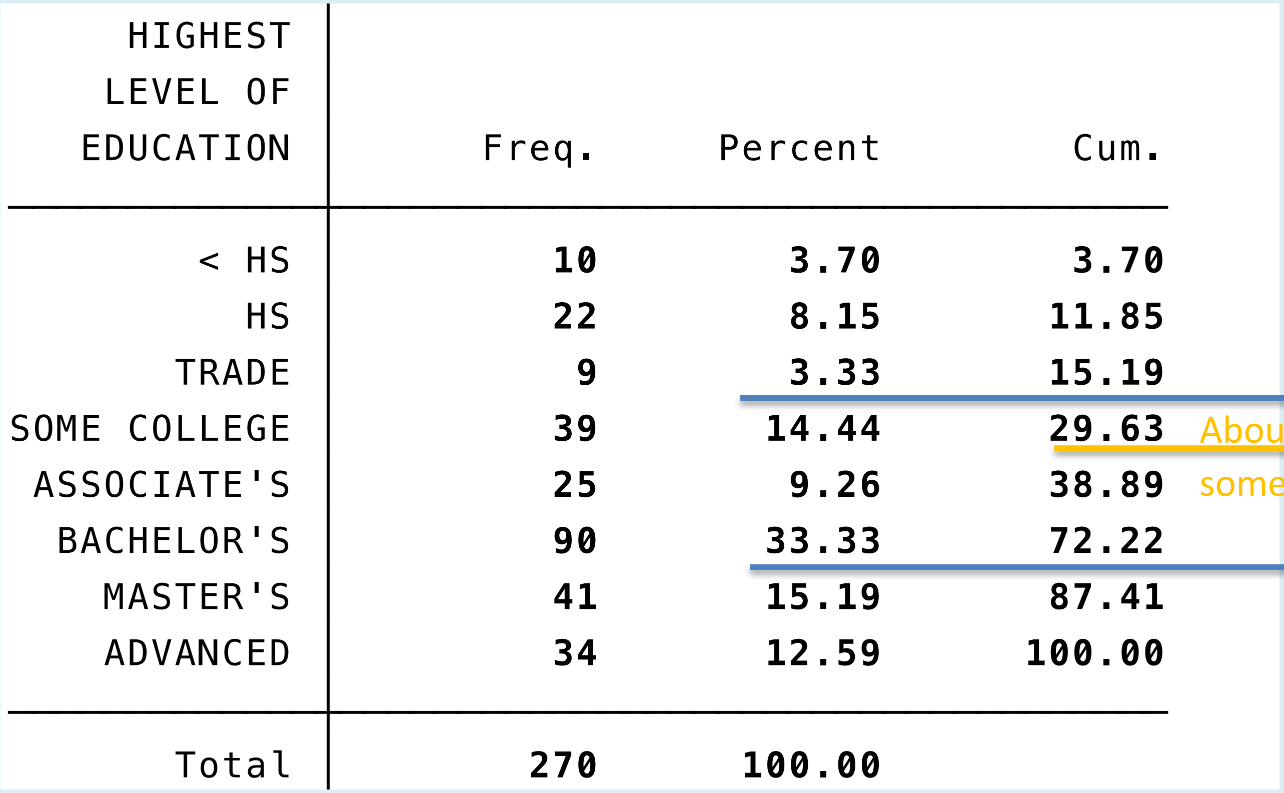
Same as Nominal but with an order. Intervals not equal.

What’s your highest level of educational attainment?

Respondent must pick just one (mutually exclusive)

Assume that these are the only possibilities (exhaustive)

Note that the categories increase in the years needed to complete (ordinality)



```
    HIGHEST |
   LEVEL OF |
  EDUCATION |      Freq.     Percent        Cum.
------------+-----------------------------------
       < HS |         10        3.70        3.70
         HS |         22        8.15       11.85
      TRADE |          9        3.33       15.19
SOME COLLEGE|         39       14.44       29.63
ASSOCIATE'S |         25        9.26       38.89
 BACHELOR'S |         90       33.33       72.22
   MASTER'S |         41       15.19       87.41
   ADVANCED |         34       12.59      100.00
------------+-----------------------------------
      Total |        270      100.00
```

About 30% have at least some college.

About 57% have at least college, but not grad level.

Cumulative: Percentage of cases at that level or below.

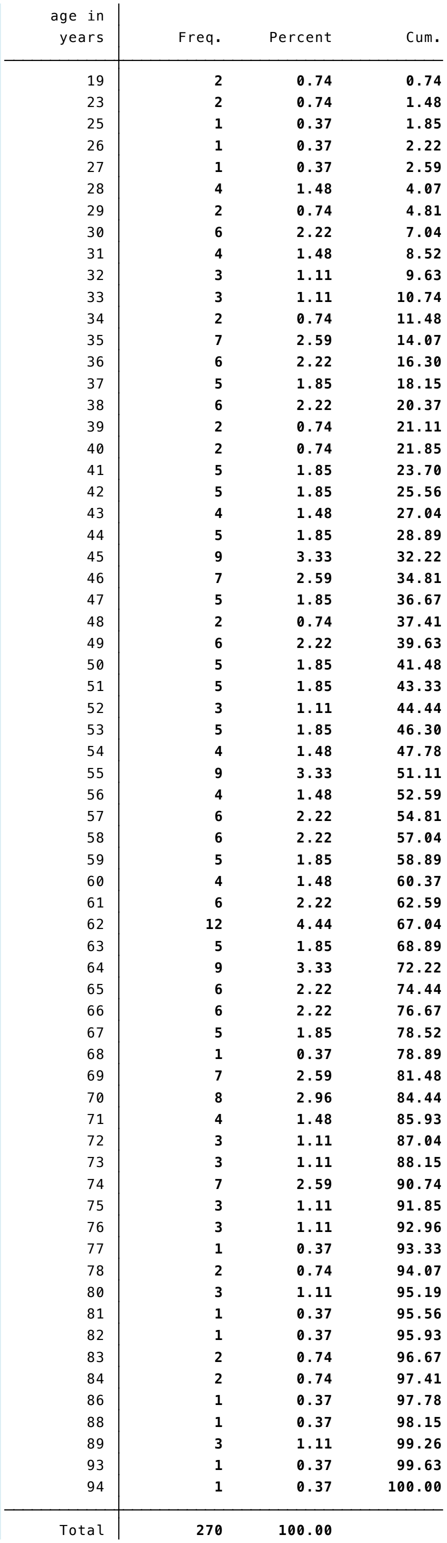
Levels of Measurement: Interval

All cases are expressed in the same units, equal intervals with a natural zero point.

Two forms:

Discrete: minimum-sized unit of measurement, cannot be sub-divided. e.g., The number of children per family.

Continious: can theoretically can take on all possible values in a given interval. e.g., Length



```
     age in |
      years |      Freq.     Percent        Cum.
------------+-----------------------------------
         19 |          2        0.74        0.74
         23 |          2        0.74        1.48
         25 |          1        0.37        1.85
         26 |          1        0.37        2.22
         27 |          1        0.37        2.59
         28 |          4        1.48        4.07
         29 |          2        0.74        4.81
         30 |          6        2.22        7.04
         31 |          4        1.48        8.52
         32 |          3        1.11        9.63
         33 |          3        1.11       10.74
         34 |          2        0.74       11.48
         35 |          7        2.59       14.07
         36 |          6        2.22       16.30
         37 |          5        1.85       18.15
         38 |          6        2.22       20.37
         39 |          2        0.74       21.11
         40 |          2        0.74       21.85
         41 |          5        1.85       23.70
         42 |          5        1.85       25.56
         43 |          4        1.48       27.04
         44 |          5        1.85       28.89
         45 |          9        3.33       32.22
         46 |          7        2.59       34.81
         47 |          5        1.85       36.67
         48 |          2        0.74       37.41
         49 |          6        2.22       39.63
         50 |          5        1.85       41.48
         51 |          5        1.85       43.33
         52 |          3        1.11       44.44
         53 |          5        1.85       46.30
         54 |          4        1.48       47.78
         55 |          9        3.33       51.11
         56 |          4        1.48       52.59
         57 |          6        2.22       54.81
         58 |          6        2.22       57.04
         59 |          5        1.85       58.89
         60 |          4        1.48       60.37
         61 |          6        2.22       62.59
         62 |         12        4.44       67.04
         63 |          5        1.85       68.89
         64 |          9        3.33       72.22
         65 |          6        2.22       74.44
         66 |          6        2.22       76.67
         67 |          5        1.85       78.52
         68 |          1        0.37       78.89
         69 |          7        2.59       81.48
         70 |          8        2.96       84.44
         71 |          4        1.48       85.93
         72 |          3        1.11       87.04
         73 |          3        1.11       88.15
         74 |          7        2.59       90.74
         75 |          3        1.11       91.85
         76 |          3        1.11       92.96
         77 |          1        0.37       93.33
         78 |          2        0.74       94.07
         80 |          3        1.11       95.19
         81 |          1        0.37       95.56
         82 |          1        0.37       95.93
         83 |          2        0.74       96.67
         84 |          2        0.74       97.41
         86 |          1        0.37       97.78
         88 |          1        0.37       98.15
         89 |          3        1.11       99.26
         93 |          1        0.37       99.63
         94 |          1        0.37      100.00
------------+-----------------------------------
      Total |        270      100.00
```

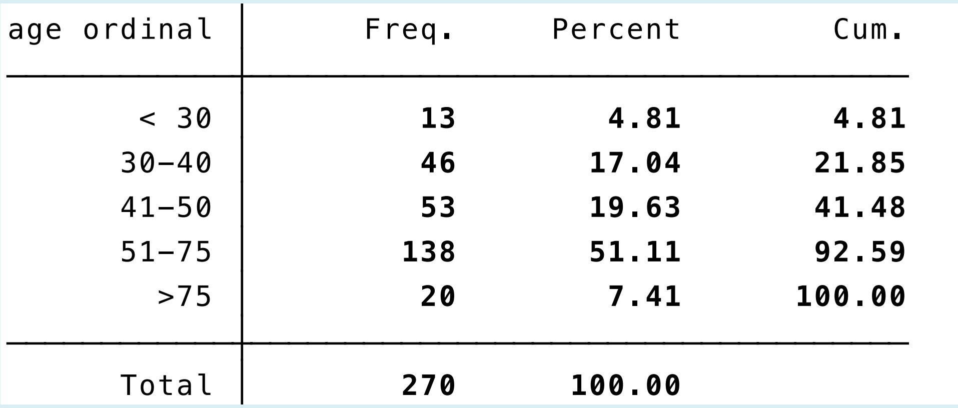
Levels of Measurement: Cumulative Property

- Variables that can be measured at the interval-ratio level of measurement can also be measured at the ordinal and nominal levels
- Variables that are measured at the nominal and ordinal levels can't be measured at higher levels

Levels of Measurement: Cumulative Property

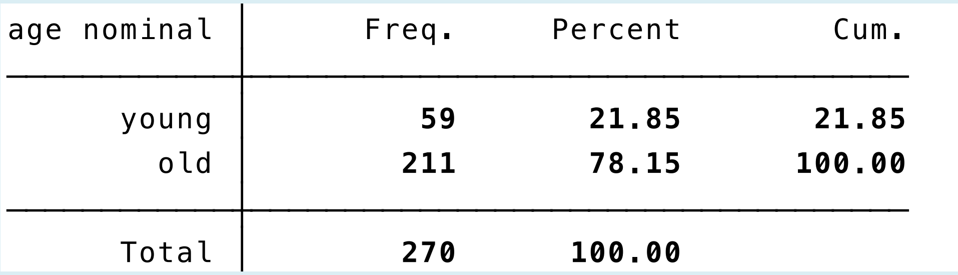
What's your age?

We can reduce the level of measurement of an interval variable down to ordinal (5 categories)

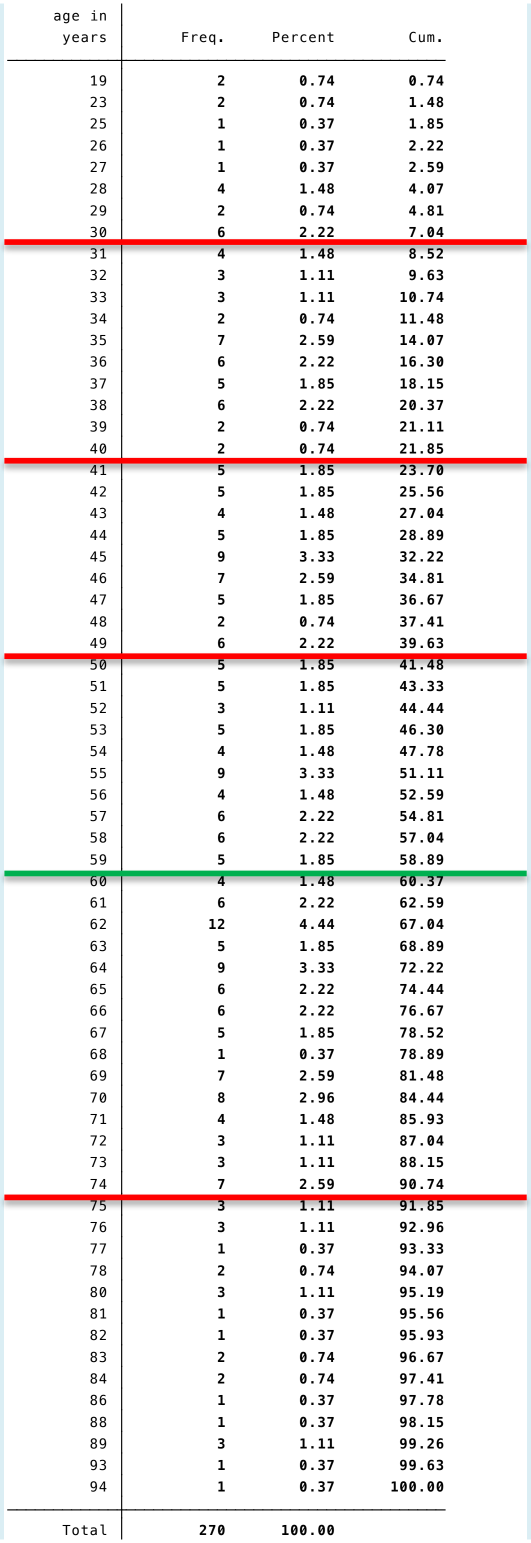


```
age ordinal │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
       < 30 │         13        4.81        4.81
      30-40 │         46       17.04       21.85
      41-50 │         53       19.63       41.48
      51-75 │        138       51.11       92.59
        >75 │         20        7.41      100.00
────────────┼───────────────────────────────────
      Total │        270      100.00
```

or nominal-dichotomous (2 levels)



```
age nominal │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
      young │         59       21.85       21.85
        old │        211       78.15      100.00
────────────┼───────────────────────────────────
      Total │        270      100.00
```



```
     age in │
      years │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
         19 │          2        0.74        0.74
         23 │          2        0.74        1.48
         25 │          1        0.37        1.85
         26 │          1        0.37        2.22
         27 │          1        0.37        2.59
         28 │          4        1.48        4.07
         29 │          2        0.74        4.81
         30 │          6        2.22        7.04
         31 │          4        1.48        8.52
         32 │          3        1.11        9.63
         33 │          3        1.11       10.74
         34 │          2        0.74       11.48
         35 │          7        2.59       14.07
         36 │          6        2.22       16.30
         37 │          5        1.85       18.15
         38 │          6        2.22       20.37
         39 │          2        0.74       21.11
         40 │          2        0.74       21.85
         41 │          5        1.85       23.70
         42 │          5        1.85       25.56
         43 │          4        1.48       27.04
         44 │          5        1.85       28.89
         45 │          9        3.33       32.22
         46 │          7        2.59       34.81
         47 │          5        1.85       36.67
         48 │          2        0.74       37.41
         49 │          6        2.22       39.63
         50 │          5        1.85       41.48
         51 │          5        1.85       43.33
         52 │          3        1.11       44.44
         53 │          5        1.85       46.30
         54 │          4        1.48       47.78
         55 │          9        3.33       51.11
         56 │          4        1.48       52.59
         57 │          6        2.22       54.81
         58 │          6        2.22       57.04
         59 │          5        1.85       58.89
         60 │          4        1.48       60.37
         61 │          6        2.22       62.59
         62 │         12        4.44       67.04
         63 │          5        1.85       68.89
         64 │          9        3.33       72.22
         65 │          6        2.22       74.44
         66 │          6        2.22       76.67
         67 │          5        1.85       78.52
         68 │          1        0.37       78.89
         69 │          7        2.59       81.48
         70 │          8        2.96       84.44
         71 │          4        1.48       85.93
         72 │          3        1.11       87.04
         73 │          3        1.11       88.15
         74 │          7        2.59       90.74
         75 │          3        1.11       91.85
         76 │          3        1.11       92.96
         77 │          1        0.37       93.33
         78 │          2        0.74       94.07
         80 │          3        1.11       95.19
         81 │          1        0.37       95.56
         82 │          1        0.37       95.93
         83 │          2        0.74       96.67
         84 │          2        0.74       97.41
         86 │          1        0.37       97.78
         88 │          1        0.37       98.15
         89 │          3        1.11       99.26
         93 │          1        0.37       99.63
         94 │          1        0.37      100.00
────────────┼───────────────────────────────────
      Total │        270      100.00
```

The Organization of Information: Frequency Distributions, Graphing

Tables: Frequency Distribution

A table reporting the number of observations falling into each category of the variable. Obviously requires categorical or ordinal data. Tables are made of Frequencies and Proportions, (row or column).

A relative frequency obtained by dividing the frequency (f) *in each category* by the total number of cases.

Provides a decimal figure ranging from 0 - 1 $P = \frac{f}{N}$

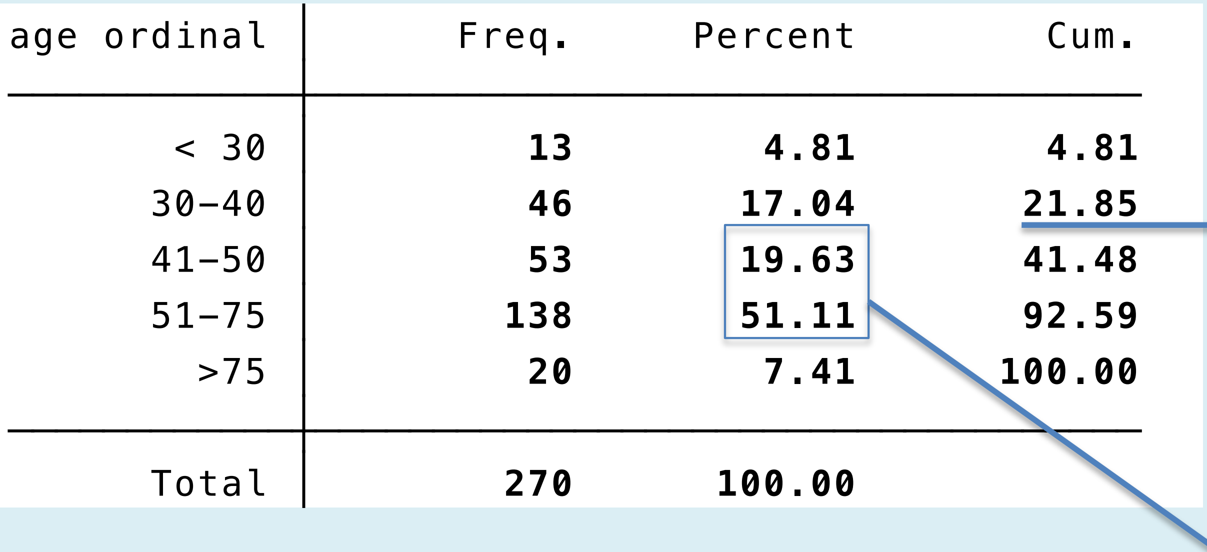
This is the Proportion

Row/Column Percentage: $(\%) = P(100)$

Cumulative Distributions

A distribution showing the frequency and/or percentage at or below each category of the variable

Useful for some interpretation



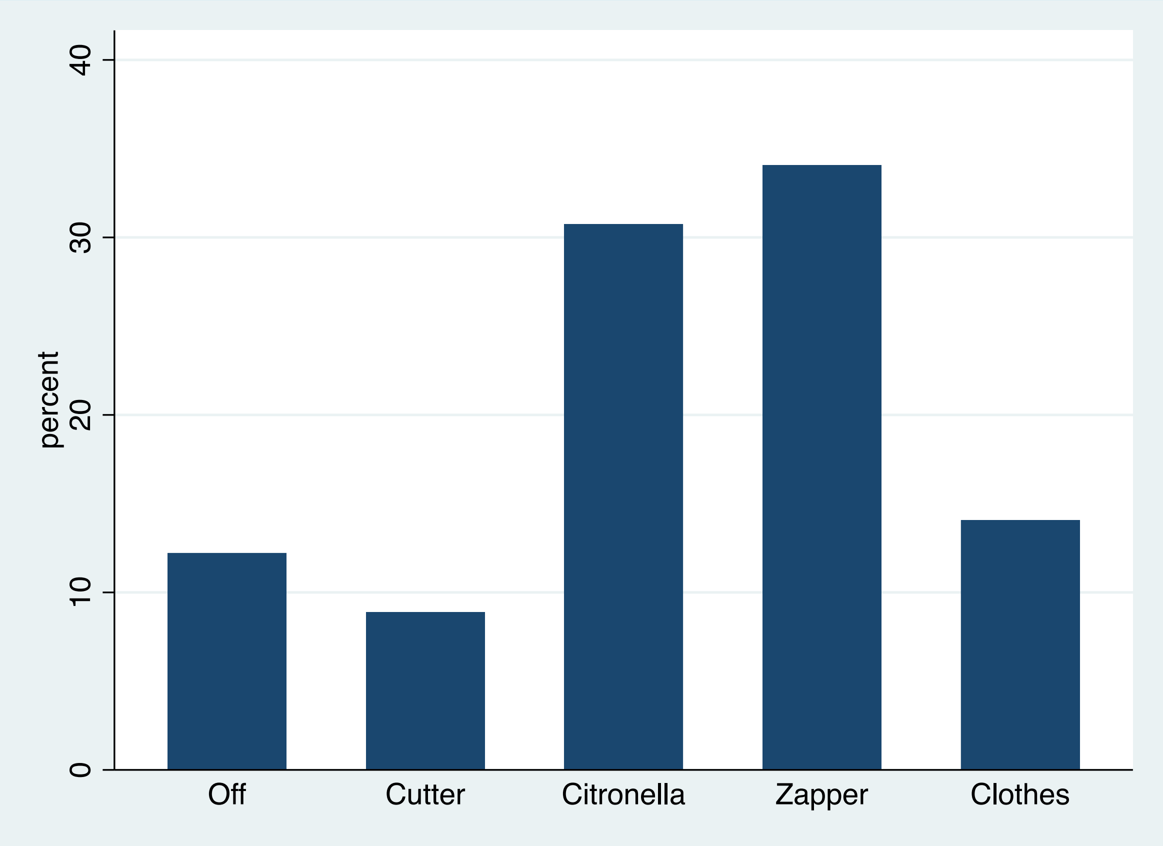
```
age ordinal |      Freq.     Percent        Cum.
------------+-----------------------------------
       < 30 |         13        4.81        4.81
      30-40 |         46       17.04       21.85
      41-50 |         53       19.63       41.48
      51-75 |        138       51.11       92.59
        >75 |         20        7.41      100.00
------------+-----------------------------------
      Total |        270      100.00
```

21.85% of the cases are 40 years old or younger

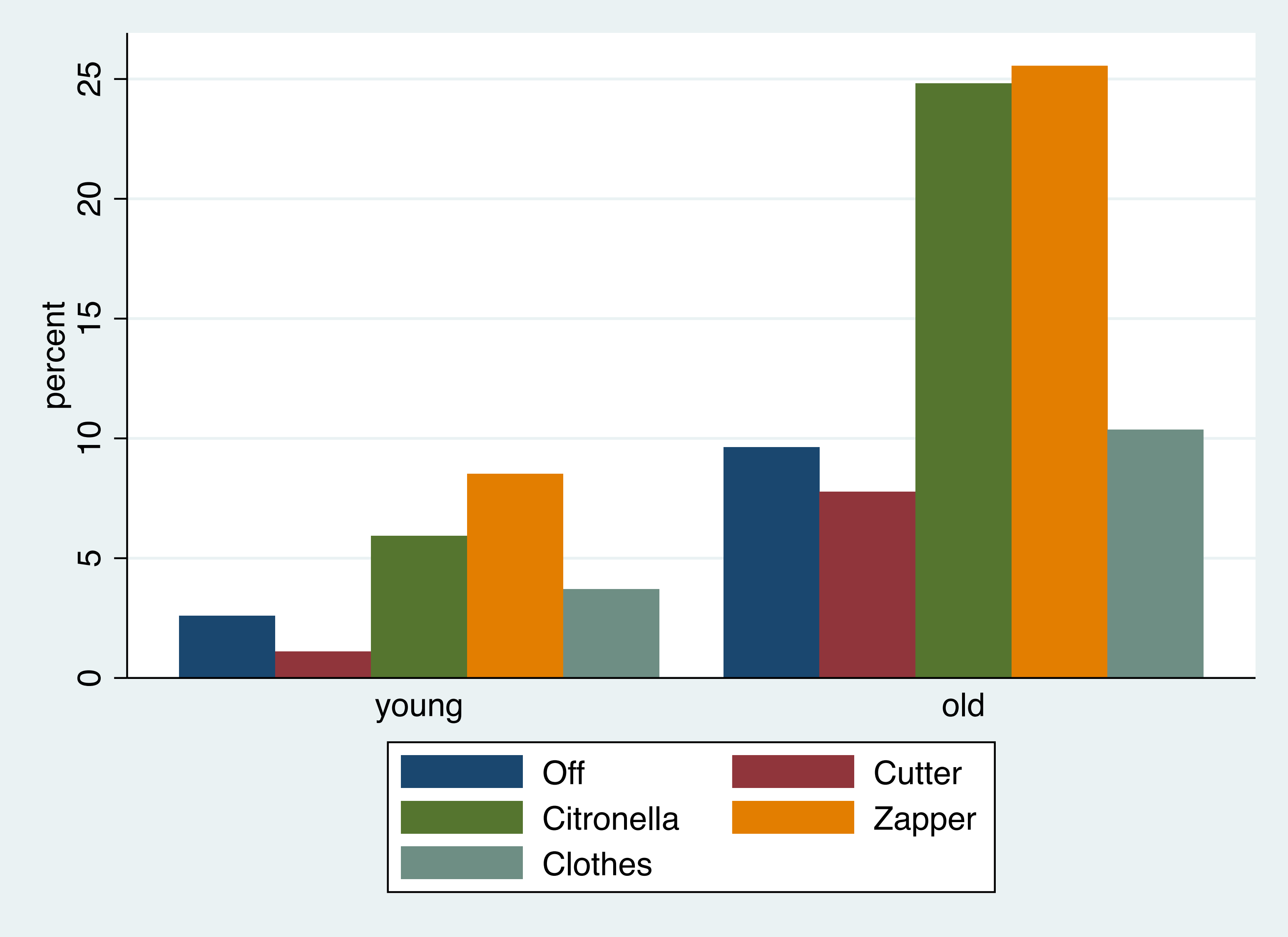
70.74% of the cases are between 40 and 75 years

The Bar Graph

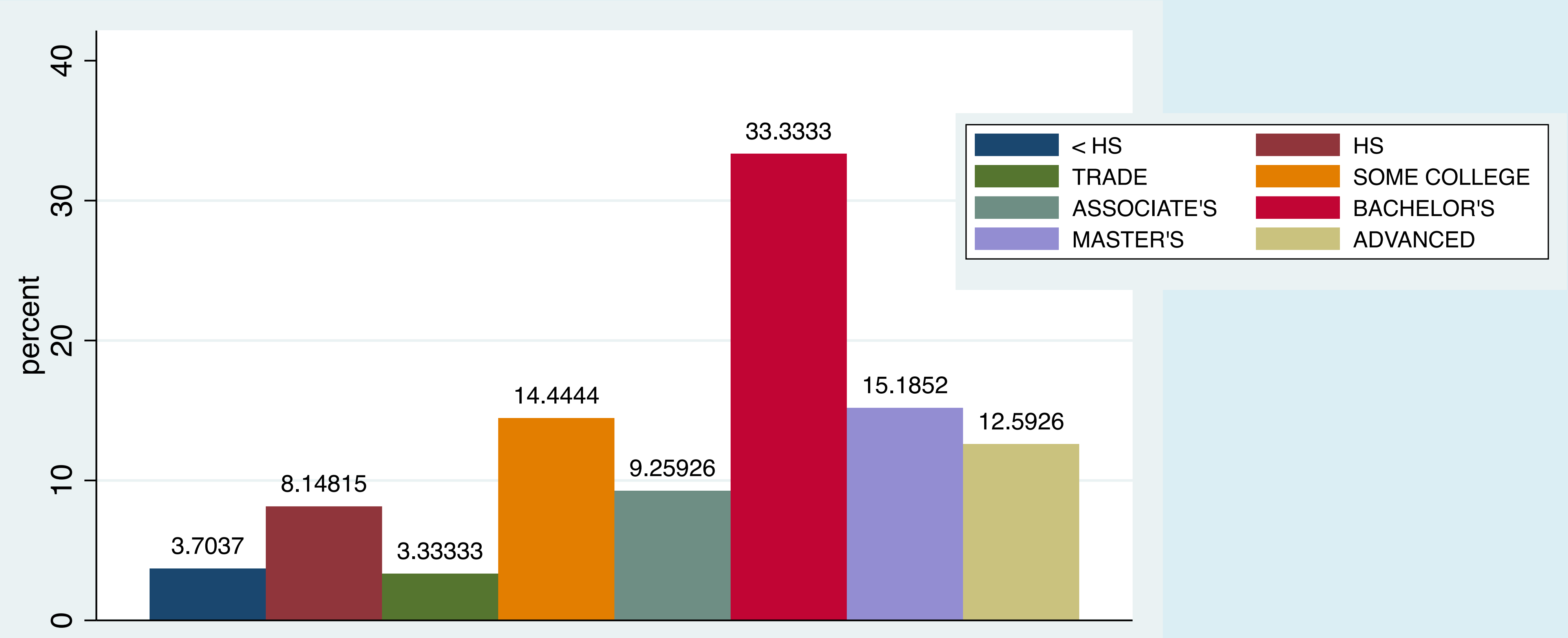
Differences in frequencies or percentages among categories: nominal or ordinal. Displayed as rectangles of equal width with height proportional to the frequency or percentage. Bars are not adjoining.

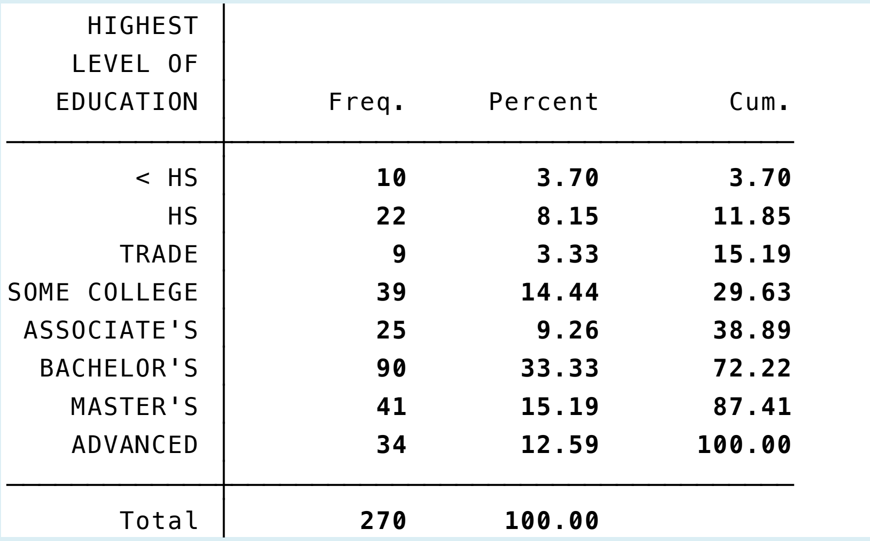


The Bar Graph (compare groups)



The Bar Graph (ordinal)

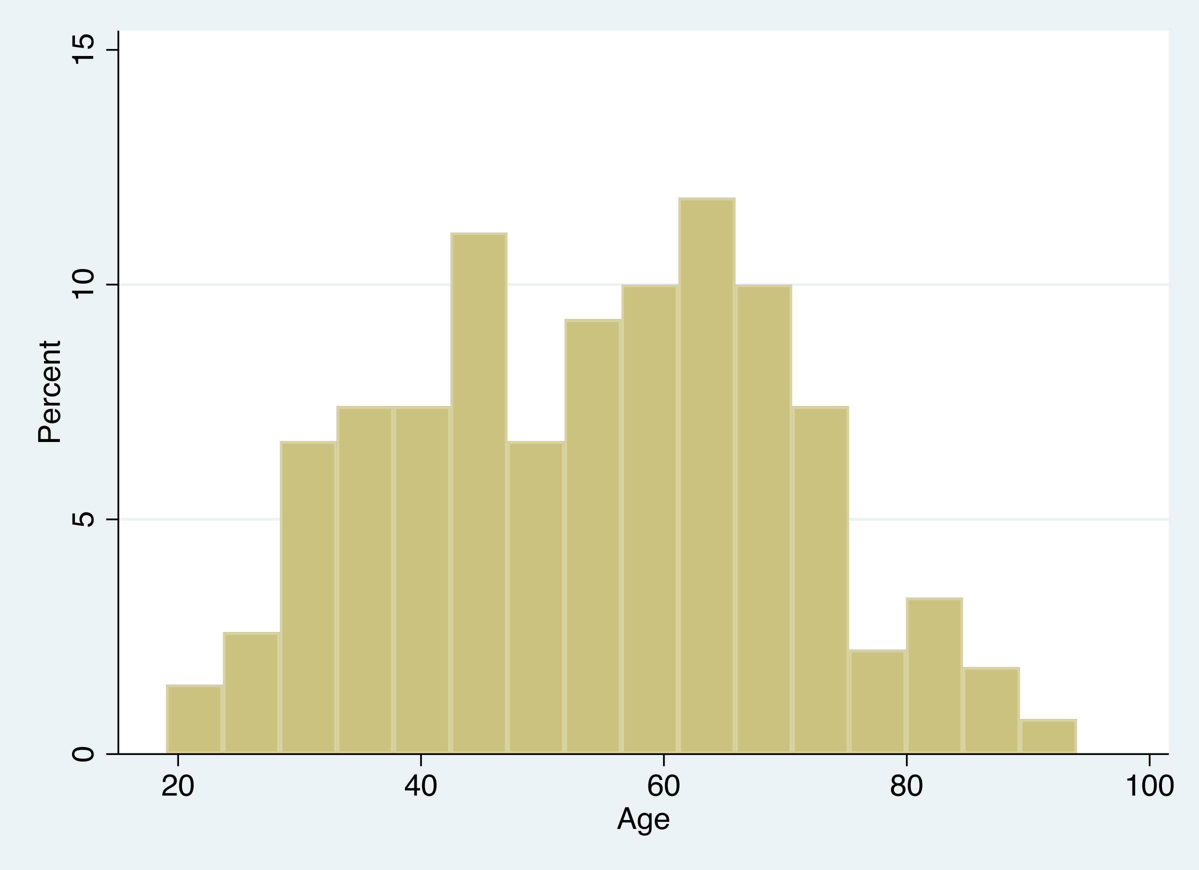




```
     HIGHEST |
    LEVEL OF |
   EDUCATION |      Freq.     Percent        Cum.
-------------+-----------------------------------
        < HS |         10        3.70        3.70
          HS |         22        8.15       11.85
       TRADE |          9        3.33       15.19
SOME COLLEGE |         39       14.44       29.63
 ASSOCIATE'S |         25        9.26       38.89
  BACHELOR'S |         90       33.33       72.22
    MASTER'S |         41       15.19       87.41
    ADVANCED |         34       12.59      100.00
-------------+-----------------------------------
       Total |        270      100.00
```

The Histogram

Shows differences in frequencies or percentages among bins of an interval-ratio variable. Bins displayed as contiguous bars, with width proportional to the width of the bin and height proportional to the frequency or percentage.



Why Use Charts and Graphs?

What do you gain?

Direct attention to one aspect of the evidence
Reach readers who might be intimidated by tabular format
Focus on bigger picture rather than details

What do you lose?

Ability to examine detail offered by a table
Ability to see additional relationships within data
Fluff: use of unnecessary graphing is poor practice

Key Terms

Research Process
Empirical Research
Hypothesis
Data
Variable
Unit of Analysis
Nominal Measurement
Dichotomous Variable
Ordinal Measurement
Interval-ratio
Dependent Variable
Independent Variable
Population
Sample
Percentage

Proportion
Percentage Distribution
Rate
Cumulative Frequency
Distribution
Cumulative Percentage
Distribution
Frequency Distribution
Bar Graph
Histogram
Line Graph
Pie Chart
Time-Series Chart

JTC270 Lab 1 Frequency tables and graphics <JTC270 Lab Dataset.dta>

Frequency Tables. We have two categorical variables that can be displayed as frequency tables: educ and income. Run frequency tables on both **(must be done separately):**

Statistics --> Summaries, tables --> Frequency tables --> One-way table
Categorical variable = educ

Repeat for Categorical variable = inc

Answer these questions about the distributions of the variables:

1) What percent of the distribution of Education is below college grad?
2) What percent of the distribution of Education is between high school and masters?
3) What percentage of the distribution of income is 50-74K?
4) What percentage of the distribution of income is 50K or greater?

Because there are relatively few discrete values, we might examine the variable Age as well. Run a frequency table on age and answer these questions:

5) What percentage of participants are younger that 50?
6) What percentage of participants are older than 80?

Bar Chart, one variable. Often we want a bar chart for just one variable.

Graphics --> Bar Chart -->
Main Tab = check Graph of percent of frequencies within categories, Orientation = Vertical
Categories Tab = check group 1, select income

7) Judging from the graph, which is the smallest group?

Bar Chart, two variables. Sometimes we want a bar chart that compares one variable by another.

Graphics --> Bar Chart -->
Main Tab = check Graph of percent of frequencies within categories, Orientation = Horizontal
Categories Tab = check group 1, select income | check group 2, select town

8) Judging from the graph, which area has a greater proportion of households with income >100K?

Histograms, one variable. Interval-level variables are examined with histograms.

Graphics --> Histogram
Main: check data are continuous, select age, Y axis = percent

9) Histograms group cases into equal bins. About what percent of the cases fall in the largest bin?

Histograms, one variable across two groups. Histograms can be compared across different groups.

Graphics --> Histogram
Main: check data are continuous, select age, Y axis = percent
By: check draw subgroups, select town. (the BY tab is far right in the dialog box)

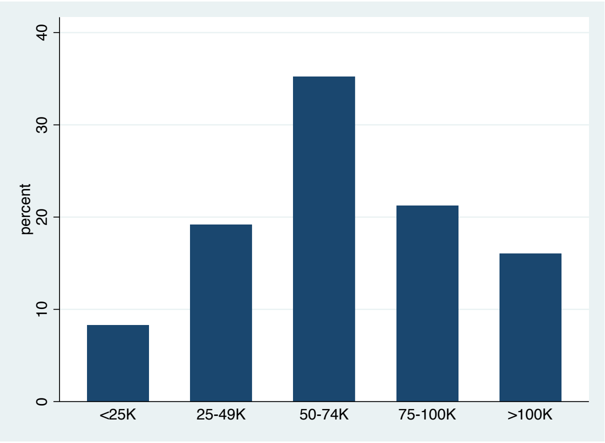
10) Which area appears to have a greater proportion of individuals over 80 years of age?

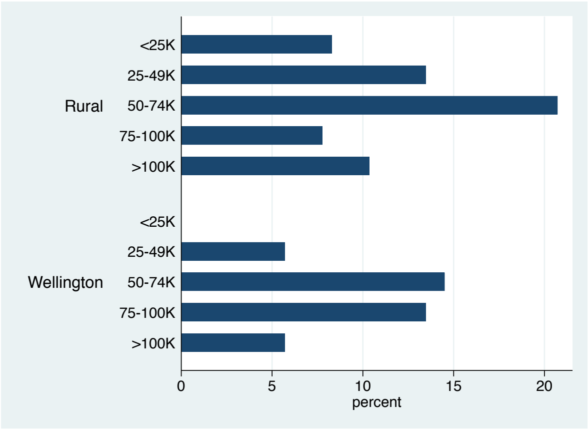
LAB1 OUTPUT

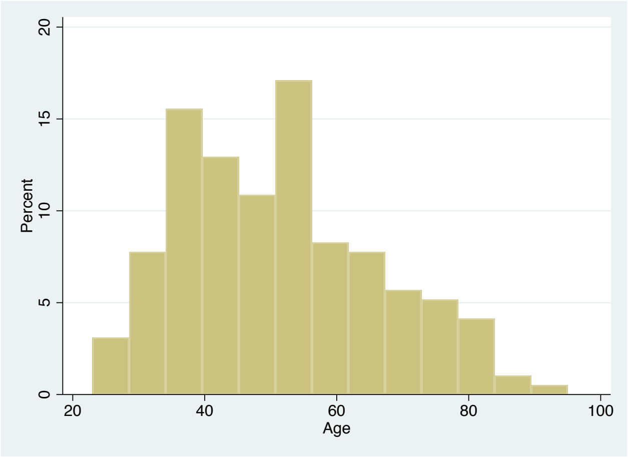
```
         Education |      Freq.     Percent        Cum.
-------------------+-----------------------------------
      less than HS |          5        2.59        2.59
           HS grad |         45       23.32       25.91
 some college/tech |         54       27.98       53.89
         tech grad |         19        9.84       63.73
      college grad |         42       21.76       85.49
          master's |         21       10.88       96.37
          doc/prof |          7        3.63      100.00
-------------------+-----------------------------------
             Total |        193      100.00
```

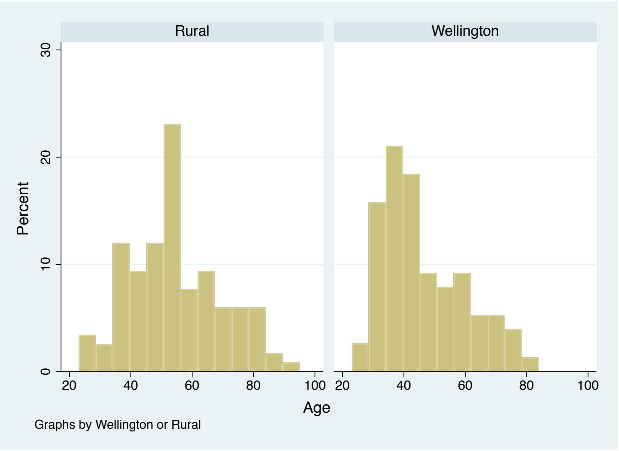
```
     Annual |
     Income |      Freq.     Percent        Cum.
------------+-----------------------------------
       <25K |         16        8.29        8.29
     25-49K |         37       19.17       27.46
     50-74K |         68       35.23       62.69
    75-100K |         41       21.24       83.94
      >100K |         31       16.06      100.00
------------+-----------------------------------
      Total |        193      100.00
```

```
        Age |      Freq.     Percent        Cum.
------------+-----------------------------------
         23 |          2        1.04        1.04
         24 |          1        0.52        1.55
         25 |          1        0.52        2.07
 - - - - - - - - - - - - - - - - - - - - - - - - 
         48 |          2        1.04       45.60
         49 |          6        3.11       48.70
         50 |          3        1.55       50.26
         51 |          5        2.59       52.85
         52 |         10        5.18       58.03
         53 |          4        2.07       60.10
         54 |          7        3.63       63.73
         55 |          3        1.55       65.28
         56 |          4        2.07       67.36
         57 |          4        2.07       69.43
         58 |          4        2.07       71.50
         60 |          4        2.07       73.58
         61 |          4        2.07       75.65
         62 |          2        1.04       76.68
         63 |          1        0.52       77.20
         64 |          3        1.55       78.76
         65 |          1        0.52       79.27
         66 |          3        1.55       80.83
         67 |          5        2.59       83.42
         68 |          2        1.04       84.46
         69 |          5        2.59       87.05
         70 |          1        0.52       87.56
         71 |          3        1.55       89.12
         73 |          2        1.04       90.16
         74 |          1        0.52       90.67
         75 |          4        2.07       92.75
         76 |          1        0.52       93.26
         78 |          2        1.04       94.30
         79 |          2        1.04       95.34
         80 |          2        1.04       96.37
         81 |          2        1.04       97.41
         82 |          1        0.52       97.93
         83 |          1        0.52       98.45
         88 |          1        0.52       98.96
         89 |          1        0.52       99.48
         95 |          1        0.52      100.00
------------+-----------------------------------
      Total |        193      100.00
```









Measures of Central Tendency

Measures of Central Tendency

- Numbers that describe what is average or typical of the distribution

- Think of this value as where the middle of a distribution lies

Measures of Central Tendency

- The choice of an appropriate measure of central tendency for representing a distribution depends on three factors
  - Level of measurement
  - The shape of the distribution
  - The purpose of the research

The Mode

- The category or score with the largest frequency in the distribution

- Can be calculated for:

  – Nominal level variables

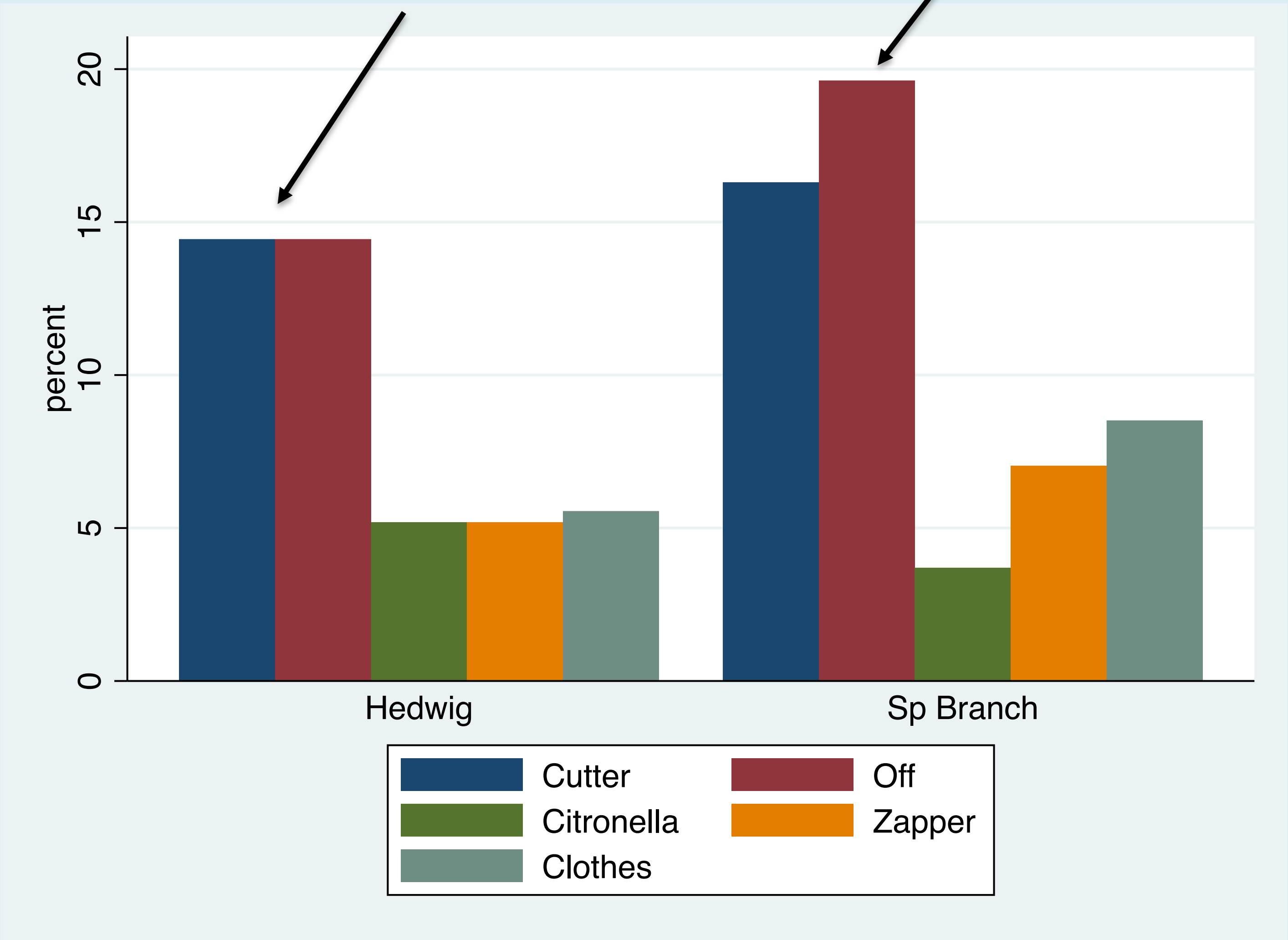
  – Ordinal level variables

  – Interval-ratio level variables (sometimes)

The Mode

Fairly strong mode in Spring Branch.

Bi-modal in Hedwig



Repellent preferences in Hedwig Village and East Spring Creek.

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it
- The middle score, or average of middle scores in a distribution
- At least ordinal level (values must be sorted)

The Median

Data: Mosquitoes with WNv per trap, Spring Branch:

1, 3, 9, 3, 4, 6, 7, 10, 2, 5 N = 10

→ sort 1, 2, 3, 3, 4, 5, 6, 7, 9, 10

Find mid-point, and if even N average: (4 + 5) ÷ 2 = 4.5

Note that the mode is 3

Data: Data: Mosquitoes with WNv per trap, Hedwig:

0, 6, 3, 2, 1, 4, 2 N = 7

→ sort 0, 1, 2, 2, 3, 4, 6

Find natural mid-point if odd N
Note that the mode is 2

The Mean

$$\bar{X} = \frac{\sum x_i}{N}$$

Important properties:

Interval-ratio level of measurement

"Center of gravity"

Sensitivity to extremes

The Mean

Data: Mosquitoes with WNv per trap, Spring Branch:

1, 3, 9, 3, 4, 6, 7, 10, 2, 5 **N = 10**

$\bar{X} = \frac{\sum x_i}{N}$ **(1 + 3 + 9 + 3 + 4 + 6 + 7 + 10 + 2 + 5) ÷ 10**

50 ÷ 10 = 5

median still 4.5

mode is still 3

Data: Mosquitoes with WNv per trap, Hedwig:

0, 6, 3, 2, 1, 4, 2 **N = 7**

$\bar{X} = \frac{\sum x_i}{N}$ **(0 + 6 + 3 + 2+ 1 + 4 + 2) ÷ 7**

18 ÷ 7 = 2.57

median still 2

mode is still 2

Note: Spring Branch data more affected by “outlier” high values, which move the mean away from the median and the mode. This is less so for Hedwig data. Also note that the mean takes into account the different number of traps in the two communities.

Measures of Variability

Measuring Variability

Central Tendency

Numbers that describe what is typical or average (central) in a distribution

Measures of Variability

Numbers that describe diversity or variability in the distribution

The two types of measures together help us to understand a distribution of scores without looking at each and every score!

Index of Qualitative Variation

A measure of variability for **nominal/ordinal** variables.

The IQV is a number that expresses the diversity of a distribution.

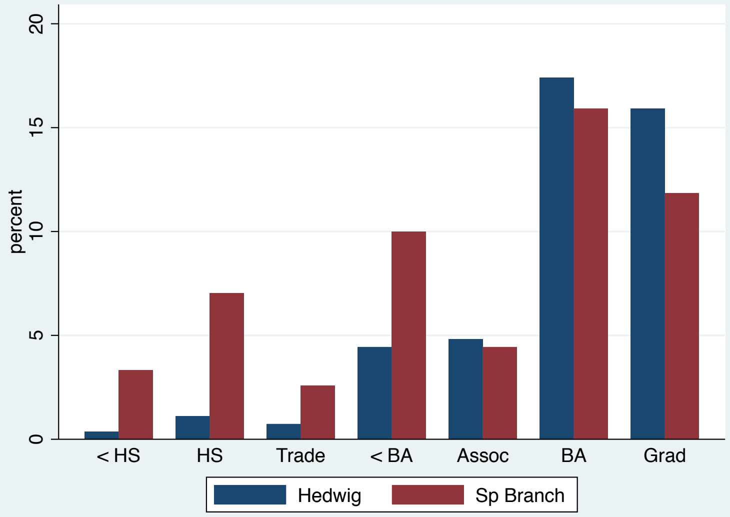
The IQV ranges from 0 to 1

An IQV of 0 would indicate that the distribution has NO diversity at all.

An IQV of 1 would indicate that the distribution is maximally diverse.

Where *k* = the number of categories:

$$\mathrm{IQV} = \frac{k\,(100^2 - \sum \%^2)}{100^2\,(k-1)}$$



IQV Comparison

How does the variance in educational attainment differ across the two areas?

Strong variance in each area, slightly more in Hedwig

Education	Sp Branch	$\%^2$	Hedwig	$\%^2$
<HS	6.04	36.4816	0.83	0.6889
HS	12.75	162.563	2.48	6.1504
Trade	4.7	22.09	1.65	2.7225
<BA	18.12	328.334	9.92	98.4064
Assoc	8.05	64.8025	10.74	115.348
BA	28.86	832.9	38.84	1508.55
Grad	21.48	461.39	35.54	1263.09
	$\sum \%^2$	1909	$\sum \%^2$	2995
	$\frac{7\,(10{,}000\ -1909)}{10{,}000\ (6)}$	0.944	$\frac{7\,(10{,}000\ -2995)}{10{,}000\ (6)}$	0.817

The Range

A measure of variation in interval-ratio variables. It is the difference between the highest (maximum) and the lowest (minimum) scores.

Range = Highest score – Lowest score

Can be calculated on percentages

Not very useful

Interquartile Range

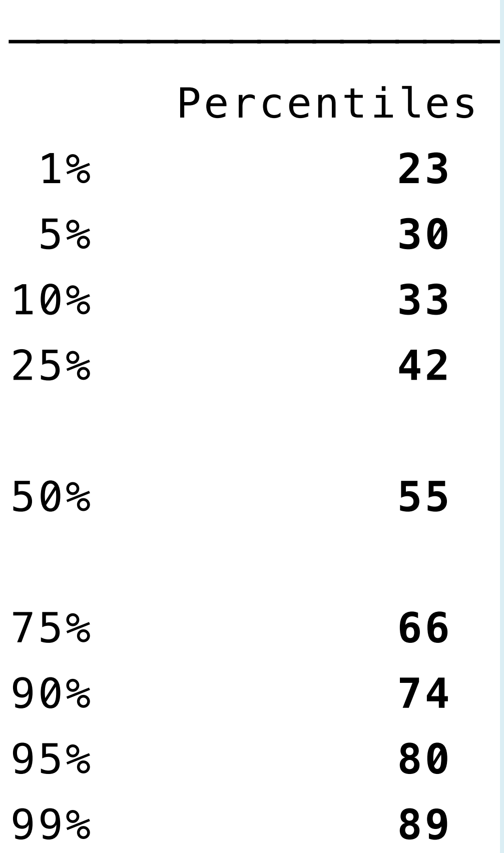
A measure of variation for interval-ratio data

Width of the middle 50% of the distribution

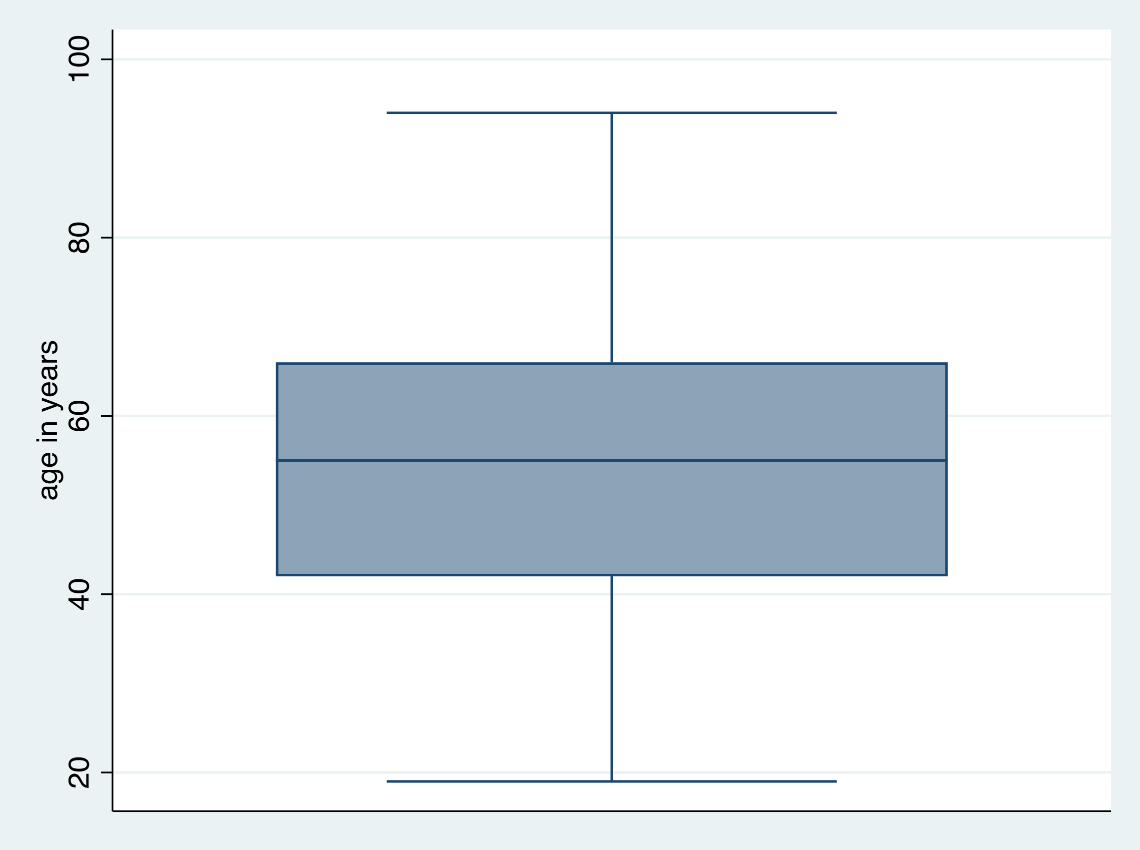
IQR = 75th percentile – 25th percentile

Age in both areas combined (N = 270, mean = 54.4, median = 55, IQR = 24)

Calculate with software (66 - 42 = 24) , graph as a boxplot



```
--------------------
      Percentiles
 1%              23
 5%              30
10%              33
25%              42

50%              55

75%              66
90%              74
95%              80
99%              89
```



Standard Deviation

A measure of variation for interval variables based on the average squared difference from the mean.

$$s_y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{N - 1}}$$

Summed
Squared deviations from mean

Averaged

This value, called the variance, is in “squared units” (e.g., squared years). Not useful

We take the square root, which changes the units back to their original measure (e.g., years)

Returning to our WNv mosquitoes per trap data . . .

traps	Sp Branch	$Yi - \bar{Y}$	$(Yi - \bar{Y})_2$		Hedwig	$Yi - \bar{Y}$	$(Yi - \bar{Y})_2$
1	1	1 - 5 = -4	16		0	0 - 2.6 = -2.6	6.76
2	3	3 - 5 = -2	4		6	6 - 2.6 = 3.4	11.6
3	9	9 - 5 = 4	16		3	3 - 2.6 = 0.4	0.16
4	3	3 - 5 = -2	4		2	2 - 2.6 = - 0.6	0.36
5	4	4 - 5 = -1	1		1	1 - 2.6 = - 1.6	2.56
6	6	6 - 5 = 1	1		4	4 - 2.6 = 1.4	1.96
7	7	7 - 5 = 2	4		2	2 - 2.6 = - 0.6	0.36
8	10	10 - 5 = 5	25	$\bar{Y}$	2.6		23.76
9	2	2 - 5 = -3	9				
10	5	5 - 5 = 0	0				
$\bar{Y}$	5		80				

$$\sqrt{\frac{80}{9}} = 2.98 \qquad \sqrt{\frac{23.76}{6}} = 1.99$$

So, to summarize . . . Spring Branch (5, 2.98) has on average almost twice as WNv detections per trap than does Hedwig (2.6, 1.99), as well as a greater degree of variance.

Choosing Descriptive Statistics

Level of Measurement	***Central Tendency***	***Variance***	***Plot***
Nominal	Mode	IQV	Bar
Ordinal	Mode	IQV	Bar
For more detailed studies, large N	Median	Range/IQR	Box
Interval	Median	Range/IQR	Box
For normally distributed data	Mean	Std Dev	Histogram

Key Terms

Measures of Central Tendency

Mean

Median

Mode

Measures of Variability

Index of Qualitative Variation

Range

Interquartile Range

Variance

Standard Deviation

JTC270 Lab 2 Summary Statistics <JTC270 Lab Dataset.dta>

$$\bar{X} = \frac{\sum x_i}{N} \qquad s_y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{N-1}} \qquad \text{IQV} = \frac{k\,(100^2 - \sum \%^2)}{100^2\,(k-1)}$$

1. To investigate community activism we looked at the number of individuals who attended 6 different public meetings on the mine issue, and noted whether they were from Wellington of the Rural area.
Calculate the mean and standard deviation for each measure. The data are:

Wellington	Xi - mean	(Xi - mean)^2
61		
61		
58		
55		
57		
60		

Rural	Xi - mean	(Xi - mean)^2
67		
72		
70		
70		
72		
69		

Round to 1 decimal point
1) Mean Wellington
2) Std Dev Wellington
3) Mean Rural
4) Std Dev Rural

2. One typical demographic comparison made is the distribution of income. Here we have the percentage of individuals in each of 5 income categories, separately for each community. Values are rounded to whole percents. NOTE: Empty categories don't count, so for Wellington *k* = 4.

Income	% in catagory Rural	%^{2}
<25K	14	
25-49K	22	
50-74K	34	
75-100K	13	
>100K	17	
Σ		

% in category Wellington	%^{2}
14	
37	
34	
15	

Σ

Report at 2 decimal points
5) IQV Rural
6) IQV Wellington
7) Interpretation: Income is more evenly distributed in which area?

Use the semester lab data on a survey of attitudes toward uranium mining in rural Colorado.
We want to obtain summery statistics on some variables and make a couple comparisons.

Statistics --> Summaries, tables, tests --> Summary and descriptive --> Summary statistics
    enter variable: age
    by/if/in: check repeat commands by group, variable = town

8) select the correct statement in Canvas
    Rural area is on average older but with less variance
    Wellington is on average younger with less variance
    Rural area is on average younger but with more variance
    Wellington is on average younger with greater variance

Statistics --> Summaries, tables, tests --> Frequency tables
        --> One-way tables
        --> Multiple one-way tables
            Main: categorical variable: water
            by/if/in: check repeat commands by group, variable = town

9) Which area has the greater percentage of household on well water?

Statistics --> Summaries, tables, tests --> Frequency tables --> One-way tables
    Main: categorical variable: educ

10) What is the mode for education?

The Normal Distribution

Identifying Distribution Shape

When we plot data we can observe its shape, or distribution.

Different measurement have different forms of distribution,

e.g., distribution of heads/tails in a coin toss: bar chart.

Recall that we want to measure at the interval level.

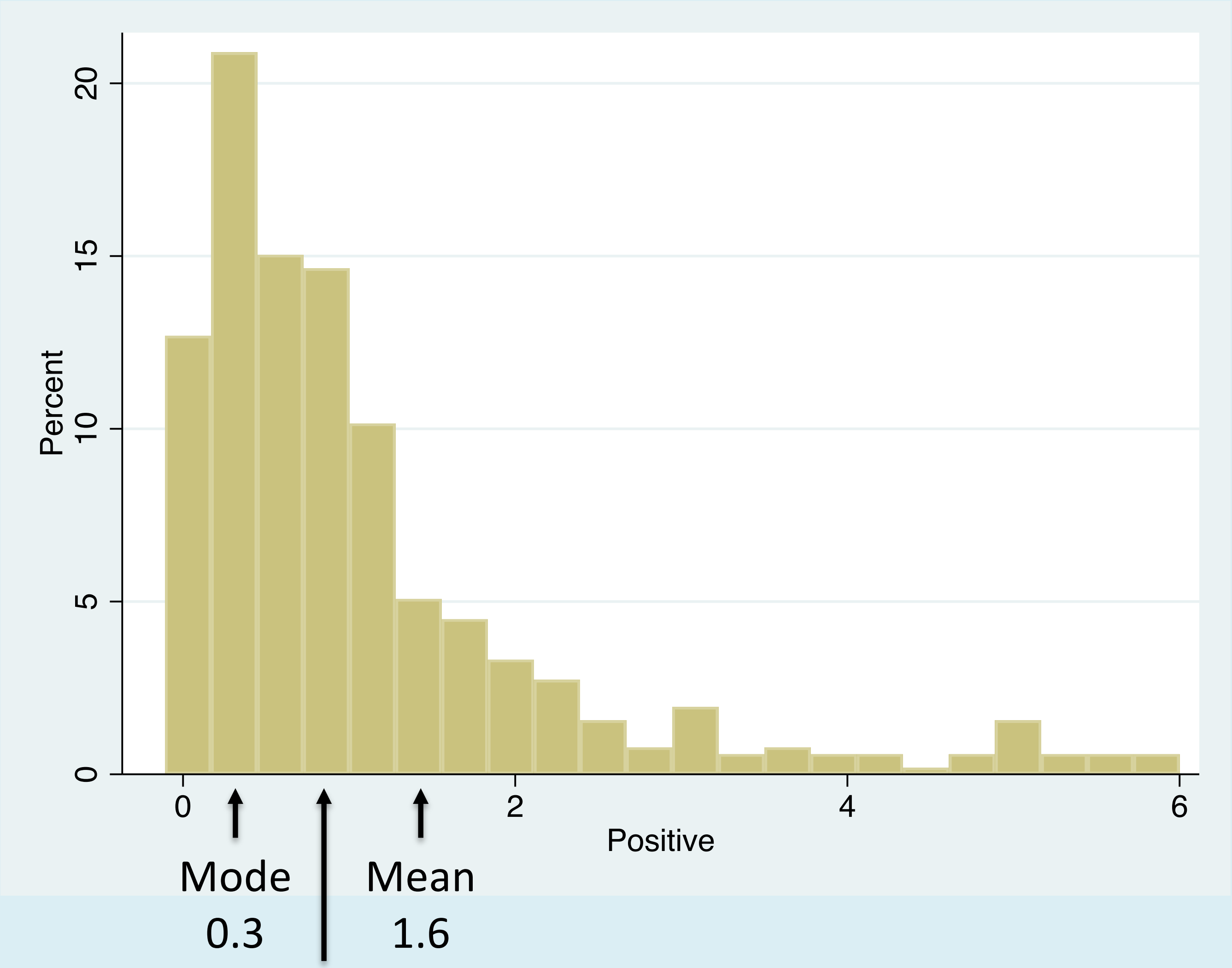
Here we focus on the shape of interval-level data, and assess it for departure from normality, or skew.

A key tool is the histogram.

Identifying Distribution Shape

Positive Skew
(tail points right)

The mean is greater than the median. A relatively few cases in the data with very high values distorts the mean.



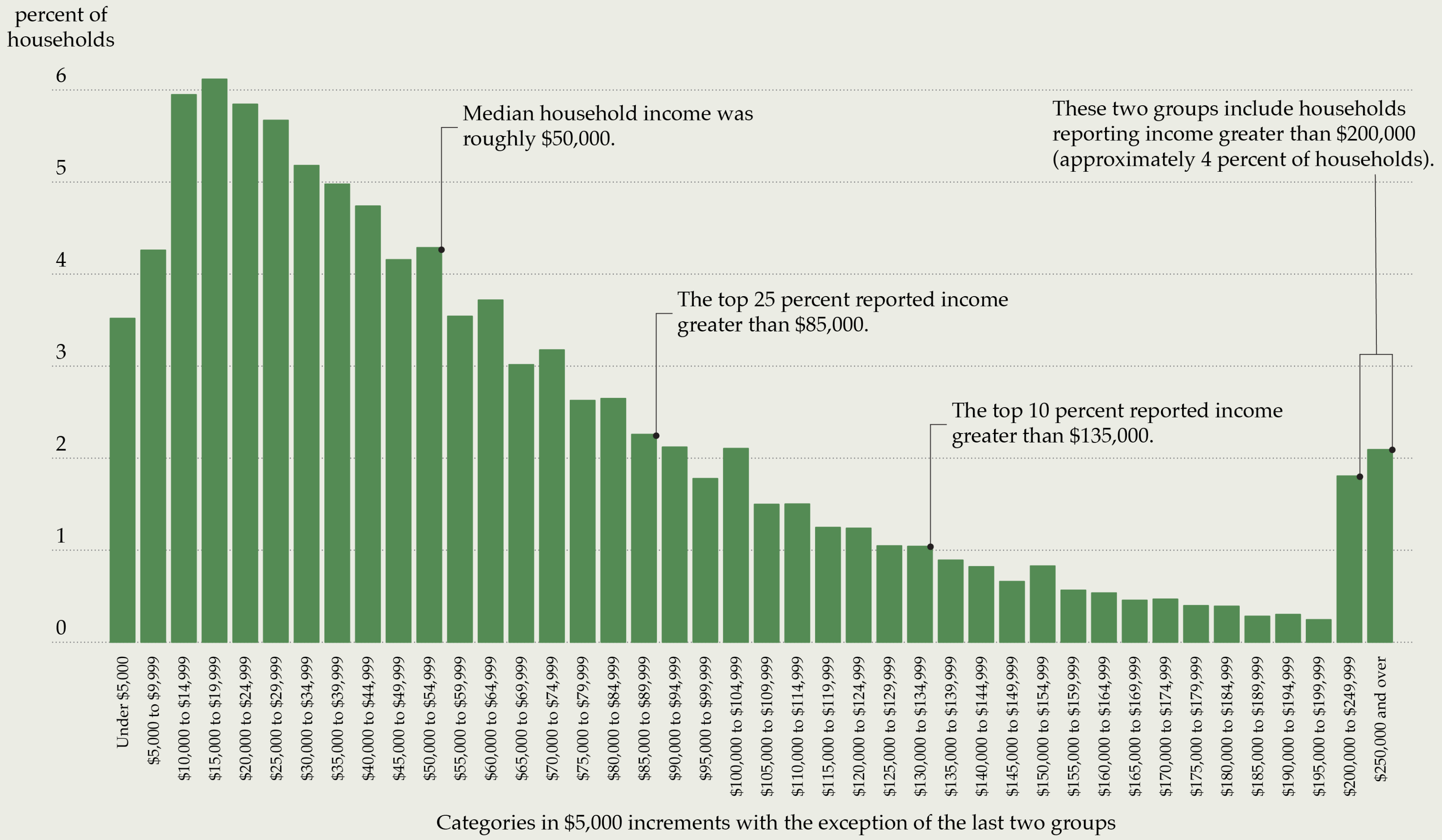
Mode
0.3

Mean
1.6

Median
0.75

Identifying Distribution Shape

Positive Skew



Distribution of annual household income in the United States

2010 estimate

Median household income was roughly $50,000.

The top 25 percent reported income greater than $85,000.

The top 10 percent reported income greater than $135,000.

These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).

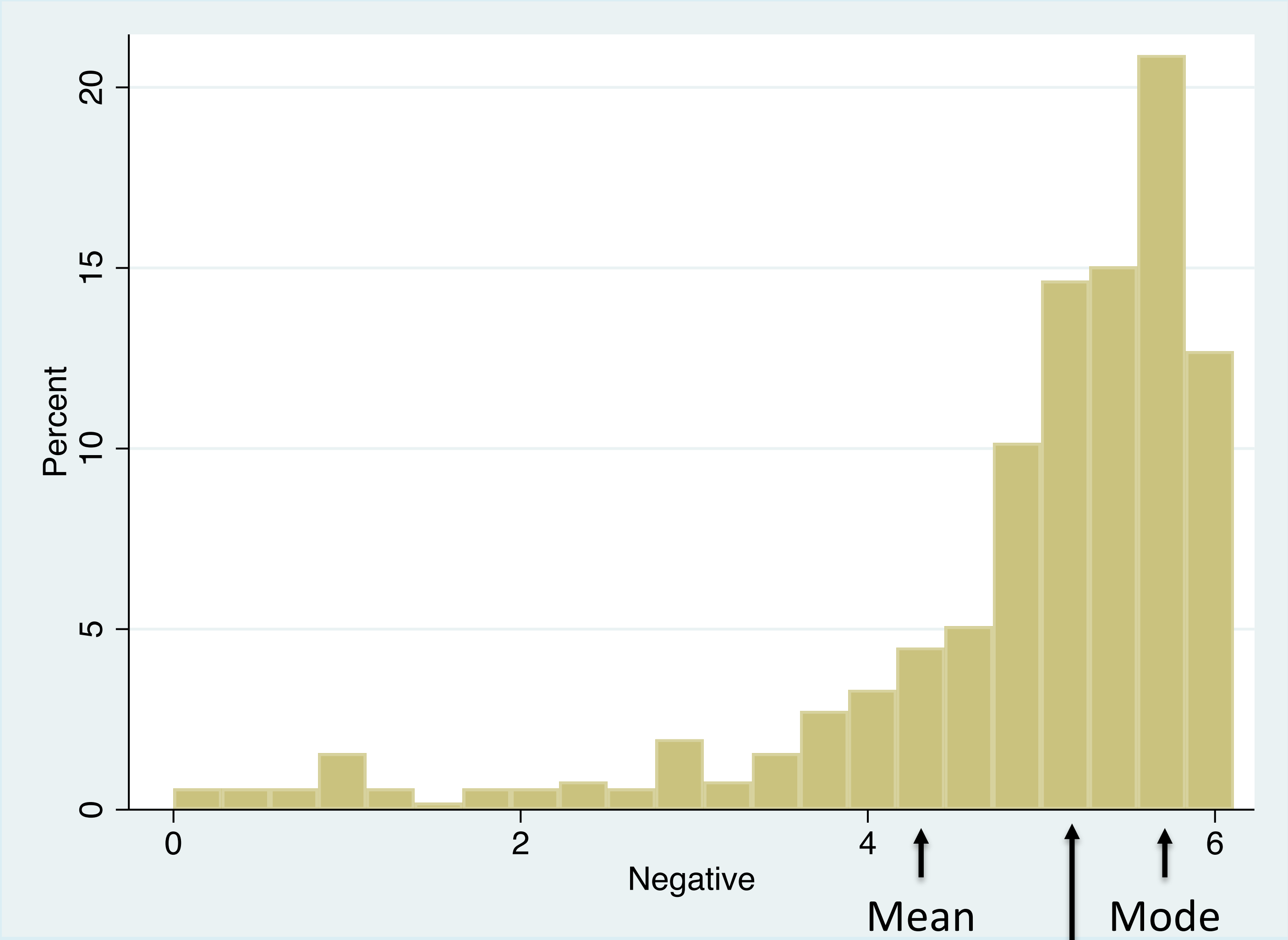
Categories in $5,000 increments with the exception of the last two groups

Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

Identifying Distribution Shape

Negative Skew (tail points left)

The mean is lesser than the median. A relatively few cases in the data with very low values distorts the mean.



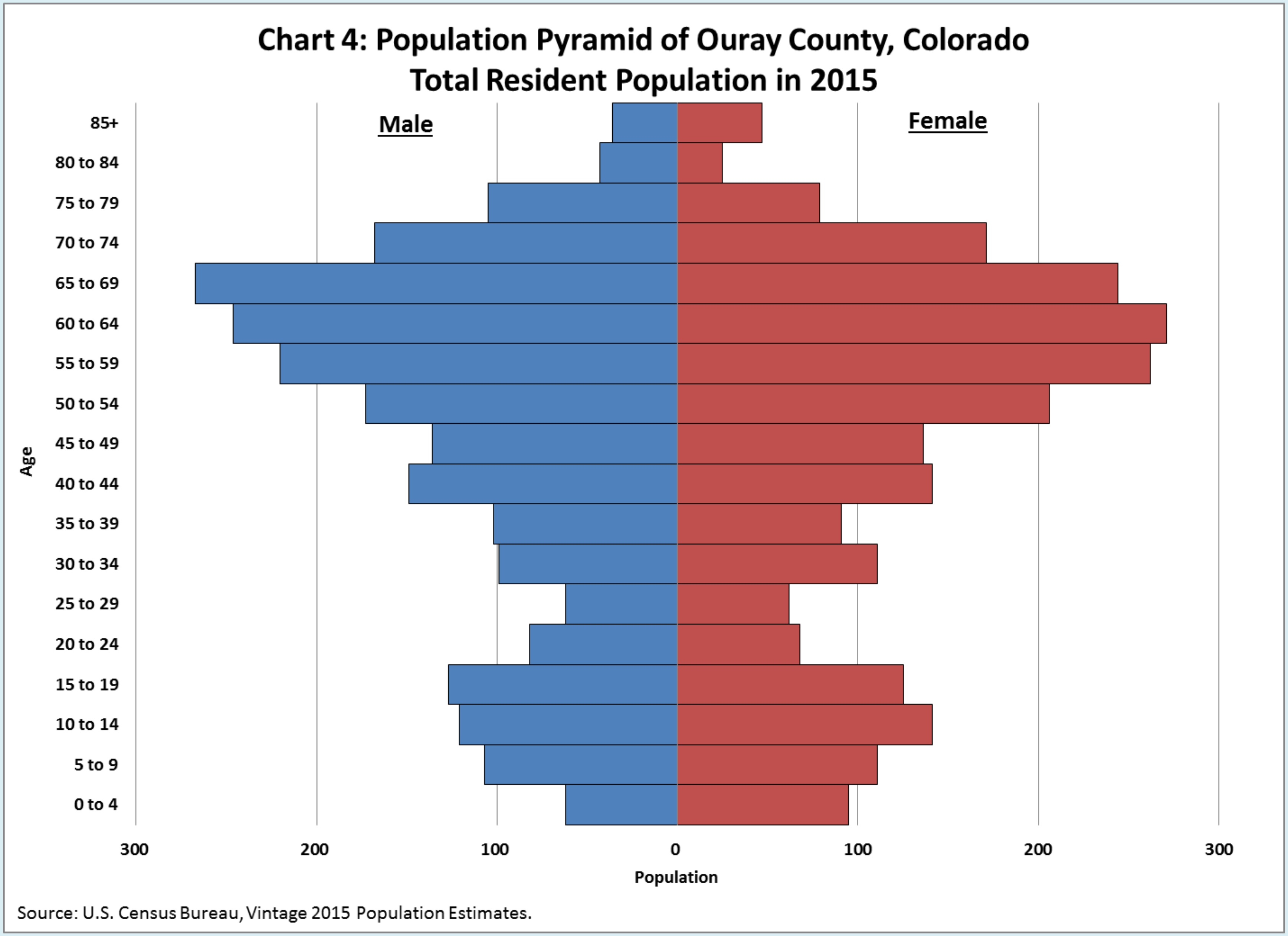
Mean
4.2

Mode
5.8

Median
5.2

Identifying Distribution Shape

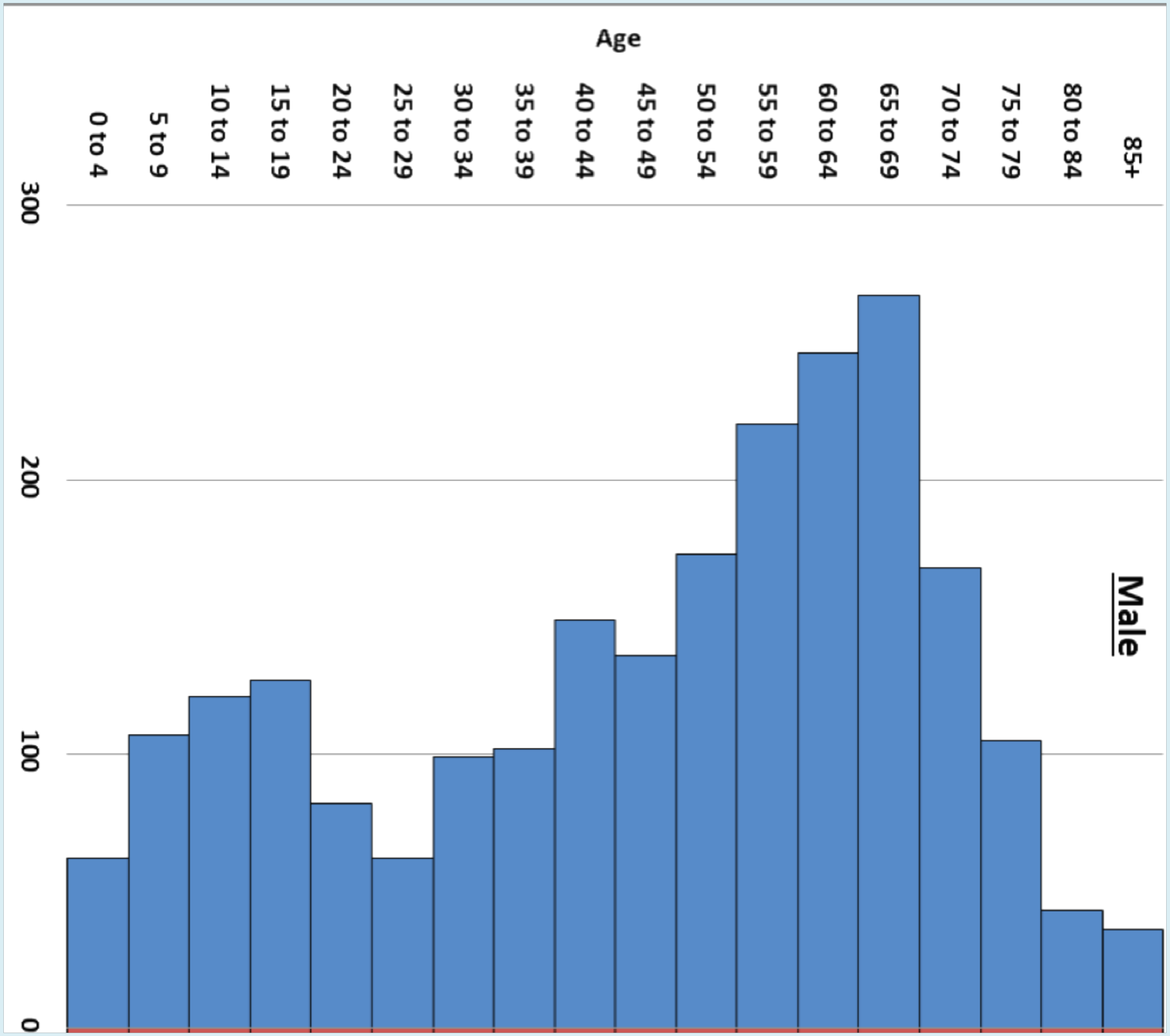
Negative Skew



Source: U.S. Census Bureau, Vintage 2015 Population Estimates.

Identifying Distribution Shape

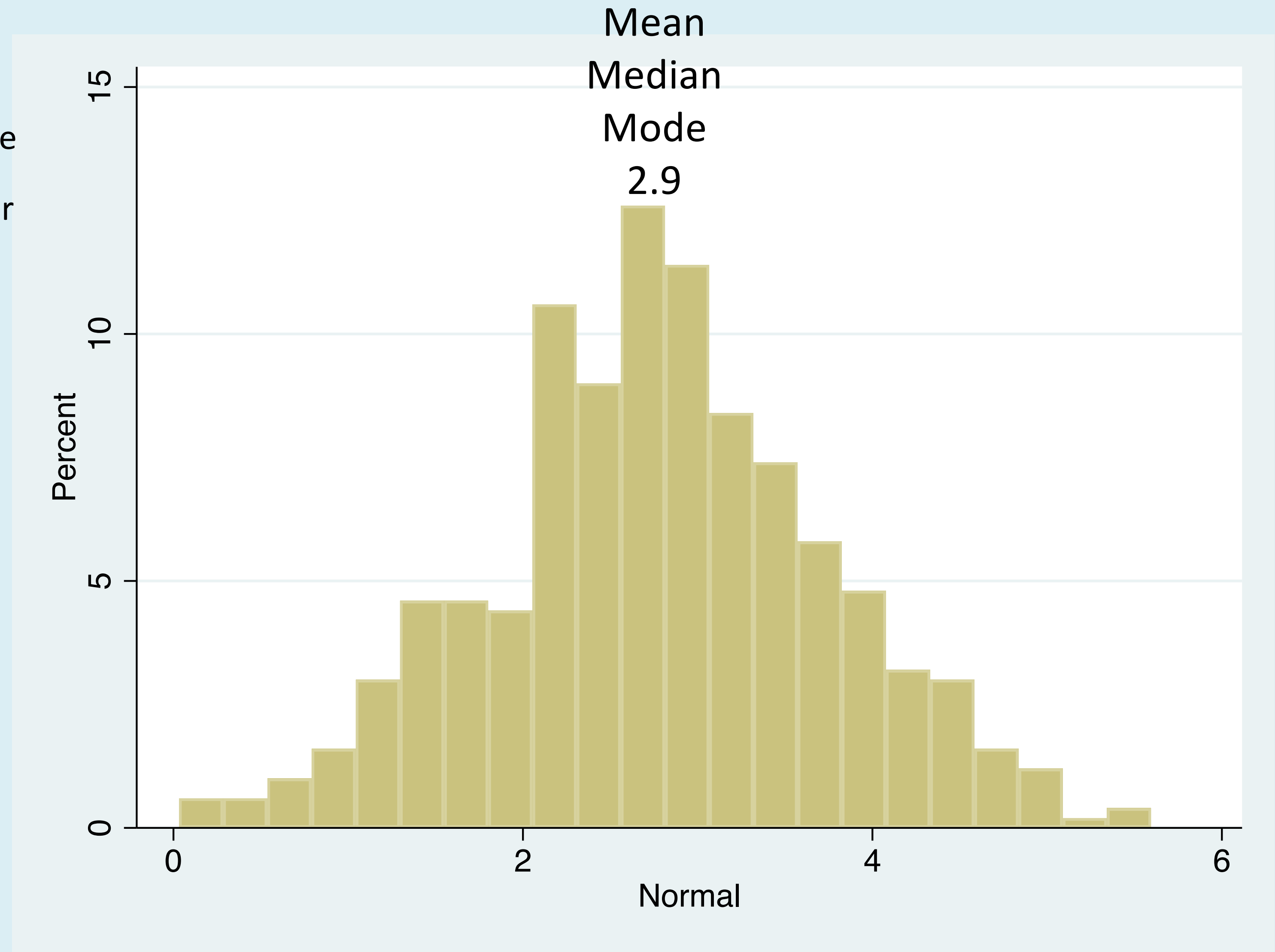
Negative Skew



Identifying Distribution Shape

Normal Distribution

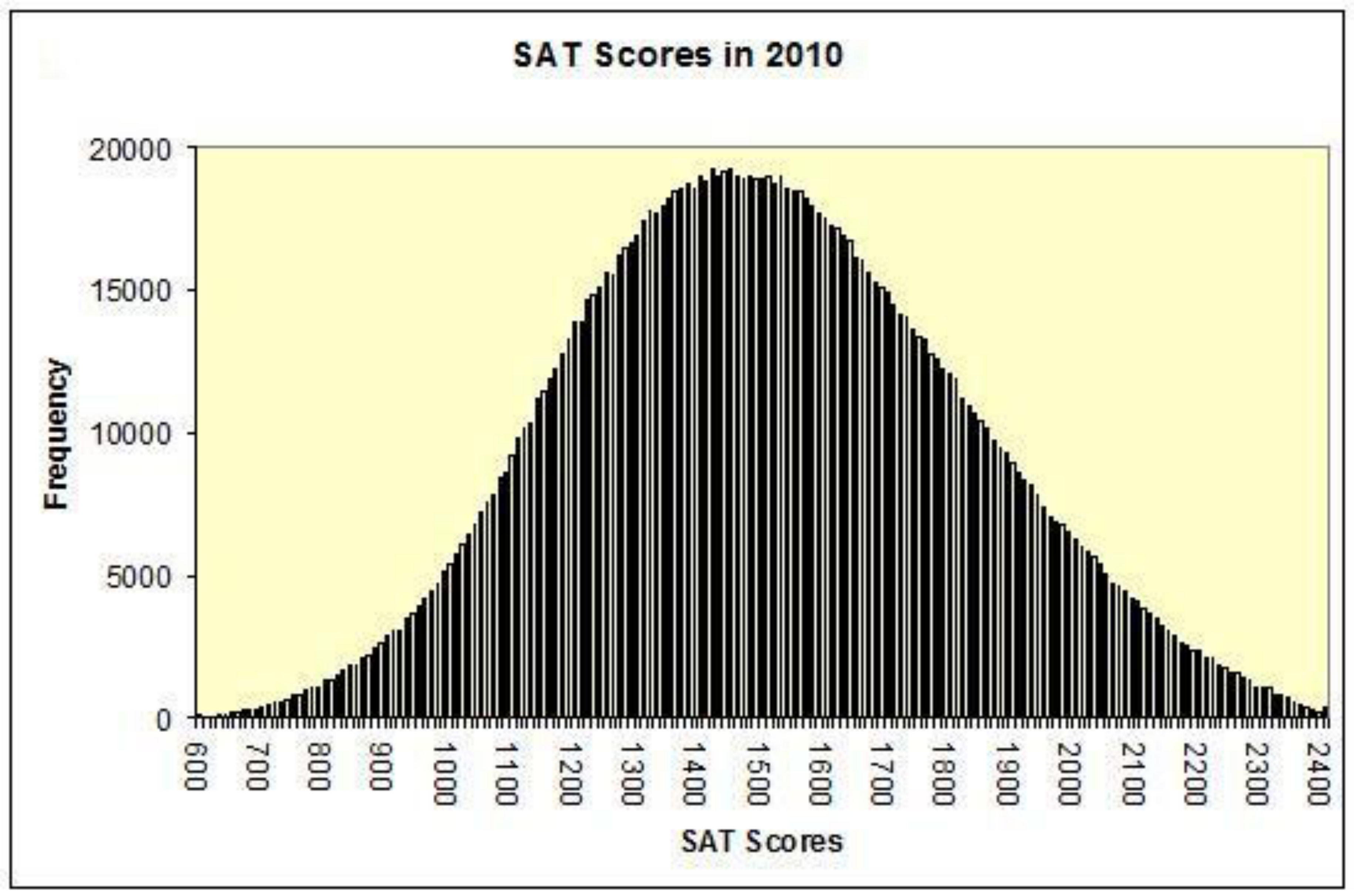
The mean and the median are approximately equal. The distribution is symmetrical, or bell-shaped.



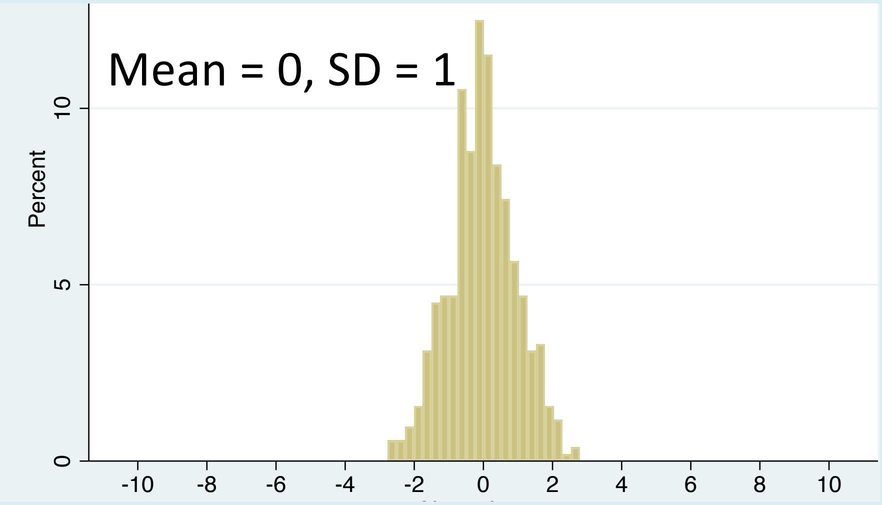
Mean
Median
Mode
2.9

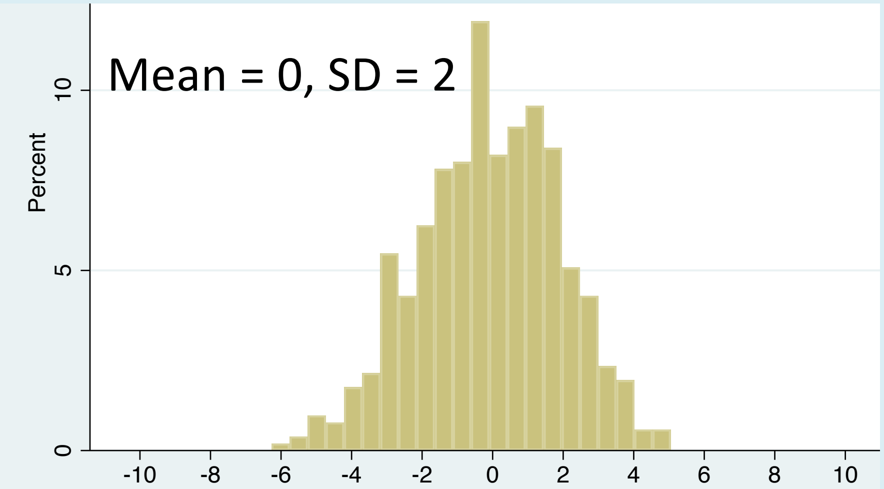
Identifying Distribution Shape

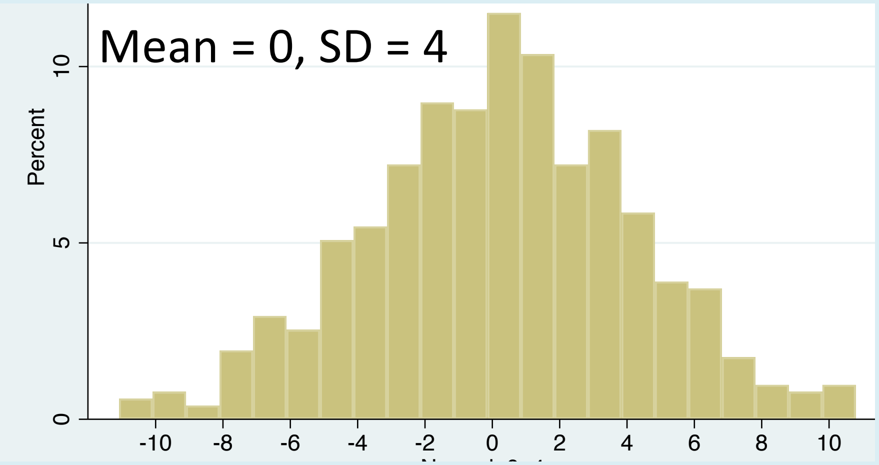
Normal Distribution



Comparing Variability







Each of these normal distributions of 500 cases has the same mean, 0.

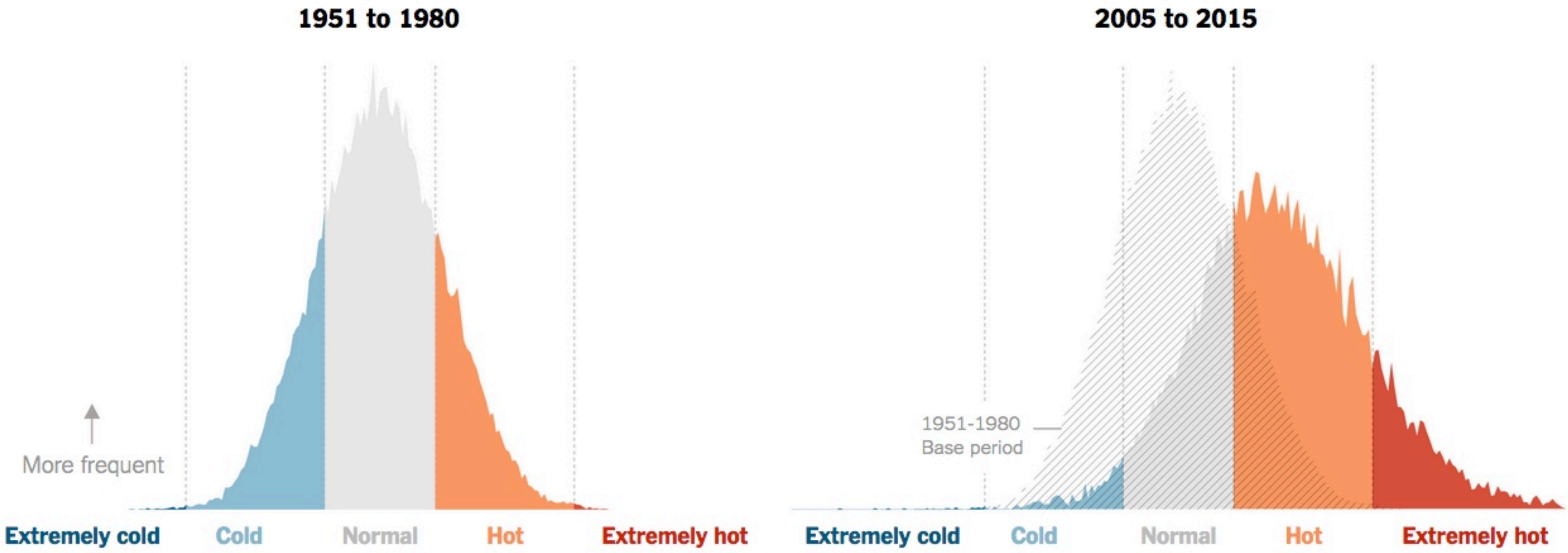
However, the amount of variance increases (top to bottom).

The standard deviations increase from 1, to 2, and 4.

There is no "correct" amount of variance as it is determined by the variable being measured.

As we'll see in the next unit, narrower distributions with smaller SDs can provide some advantages.

Comparing Variability



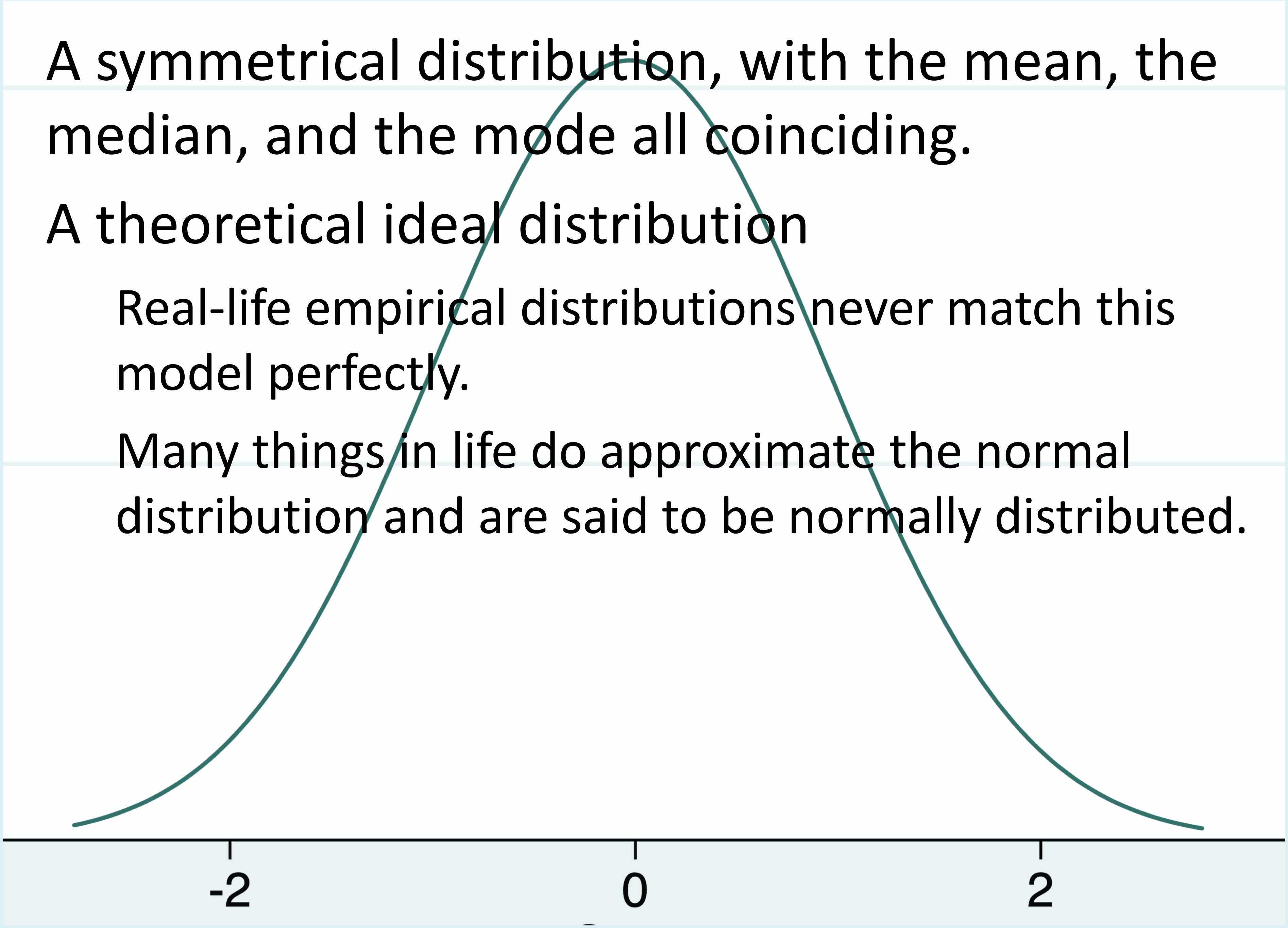
Mean increases
Variance increases

Normal Distribution

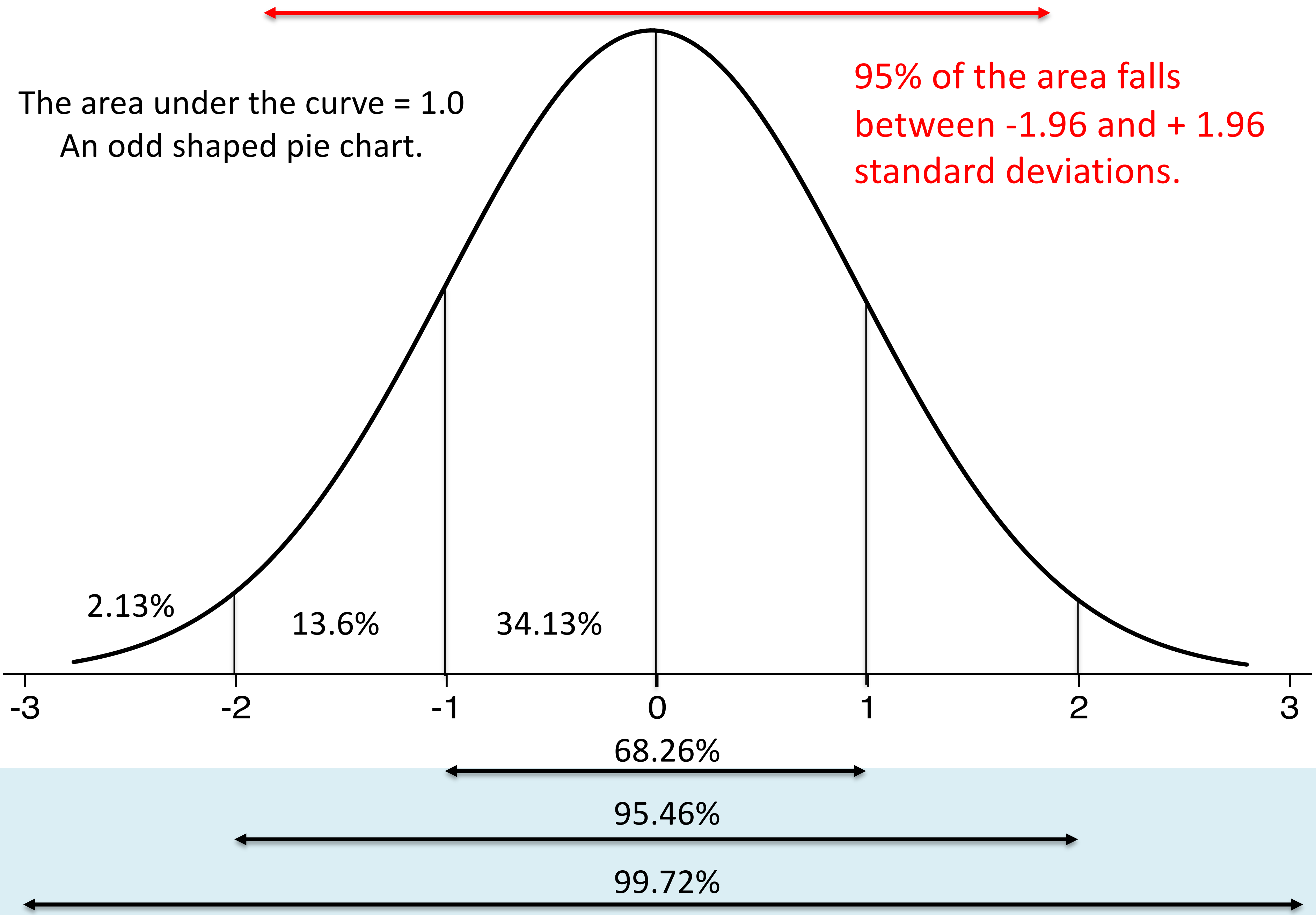
A symmetrical distribution, with the mean, the median, and the mode all coinciding.

A theoretical ideal distribution

- Real-life empirical distributions never match this model perfectly.
- Many things in life do approximate the normal distribution and are said to be normally distributed.



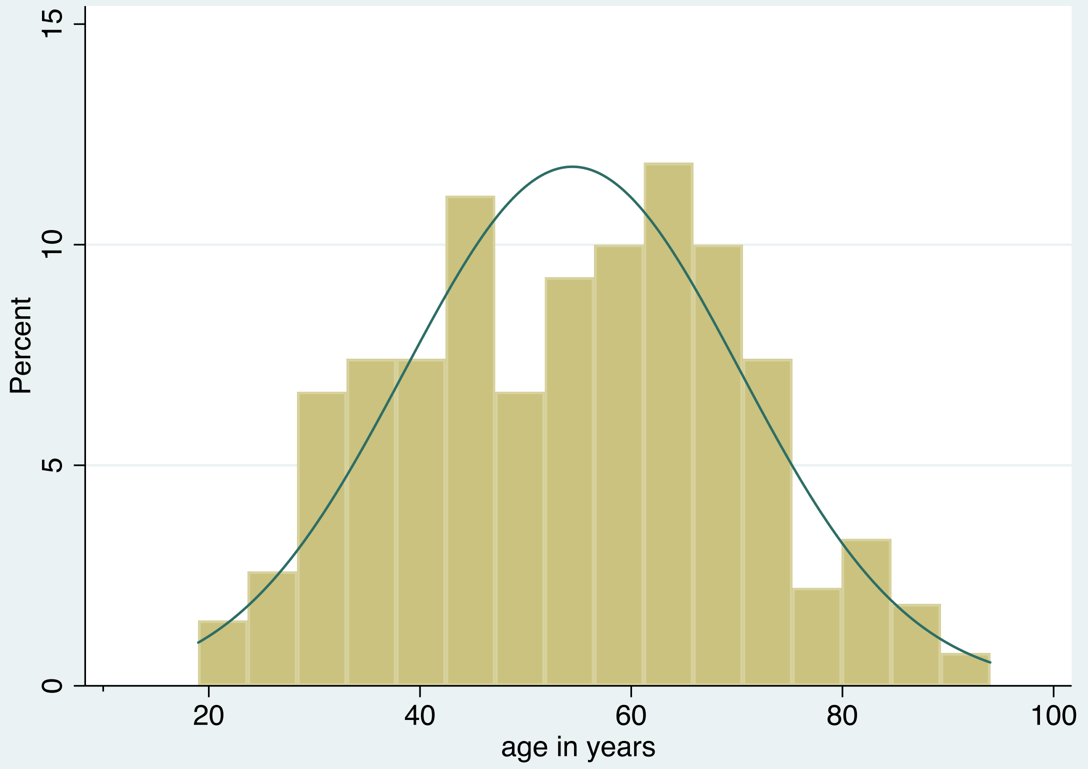
Normal Distribution



The area under the curve = 1.0
An odd shaped pie chart.

95% of the area falls between -1.96 and + 1.96 standard deviations.

Applying the normal distribution to actual data

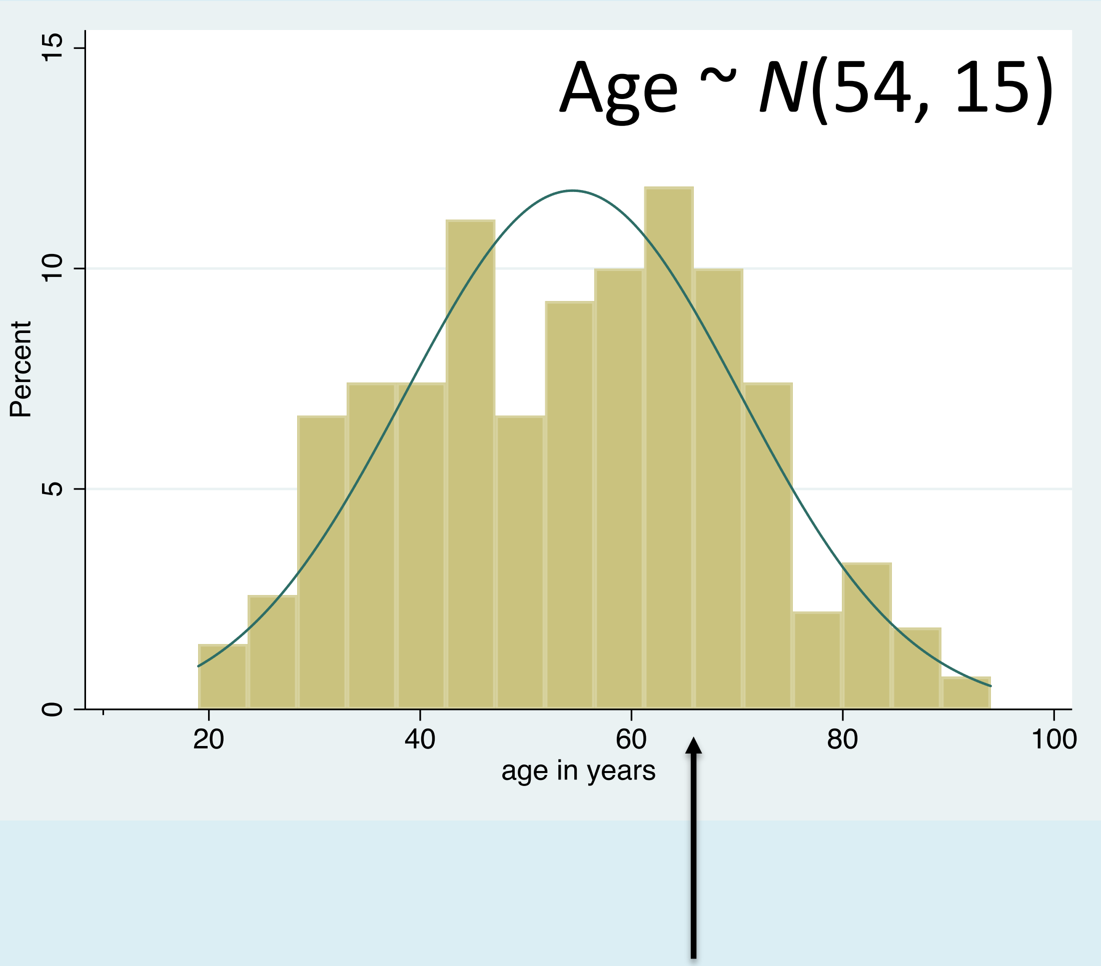


Taking the age measure we previously used, we can plot it and see that it's approximately normal, no extreme skew. So:

Age ~ *N*(54, 15)

Recalling the *empirical rule,* we would anticipate that 95% of the cases fall within 1.96 Std Dev (1.96 X 15 ~ 29 years) of the mean, thus: range is about (54-29 = 25) to (54+29 = 83)

Applying the normal distribution to actual data

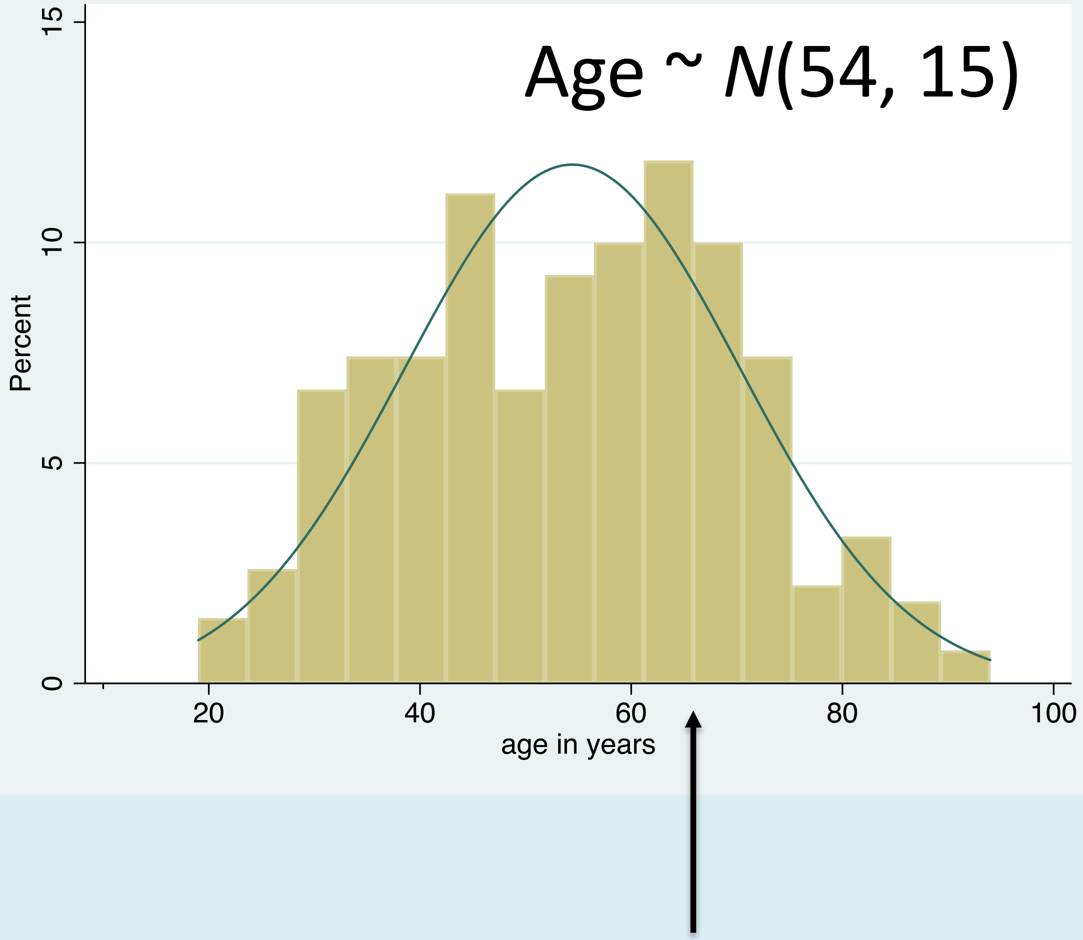


? –>

So, for planning purposes we want to know what proportion of the sample is 65 years old or older.

Using a table (or software) we can look at the normal distribution and find the proportion of the curve that is at 65 or greater.

Applying the normal distribution to actual data



Age ~ $N(54, 15)$

? –>

But the table, as we'll see, does not list years, only standard deviations and areas.

When standard deviations are listed in this manner they are called Z-scores.

So we need to convert our data that is in years to Z-scores.

Applying the normal distribution to actual data

$$Z = \frac{Y_i - \bar{Y}}{s_y}$$

The formula allows us to find the Z-score that matches any given value in the data. Here, we need the Z-score for 65.

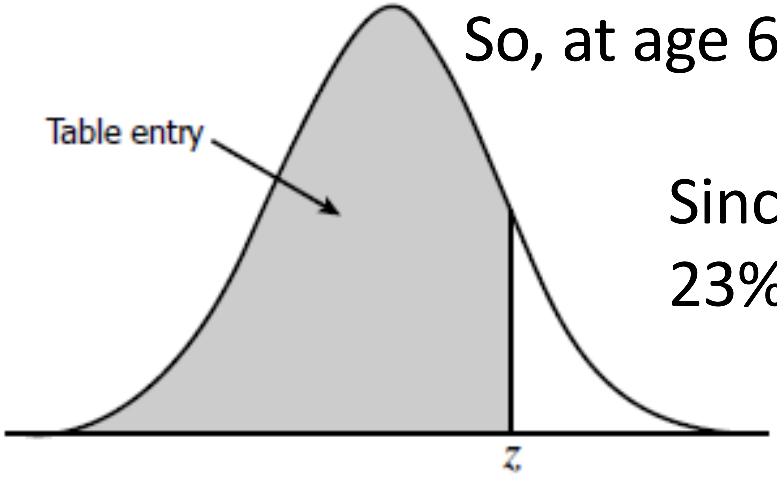
$$0.73 = \frac{65 - 54}{15}$$

Age ~ *N*(54, 15)

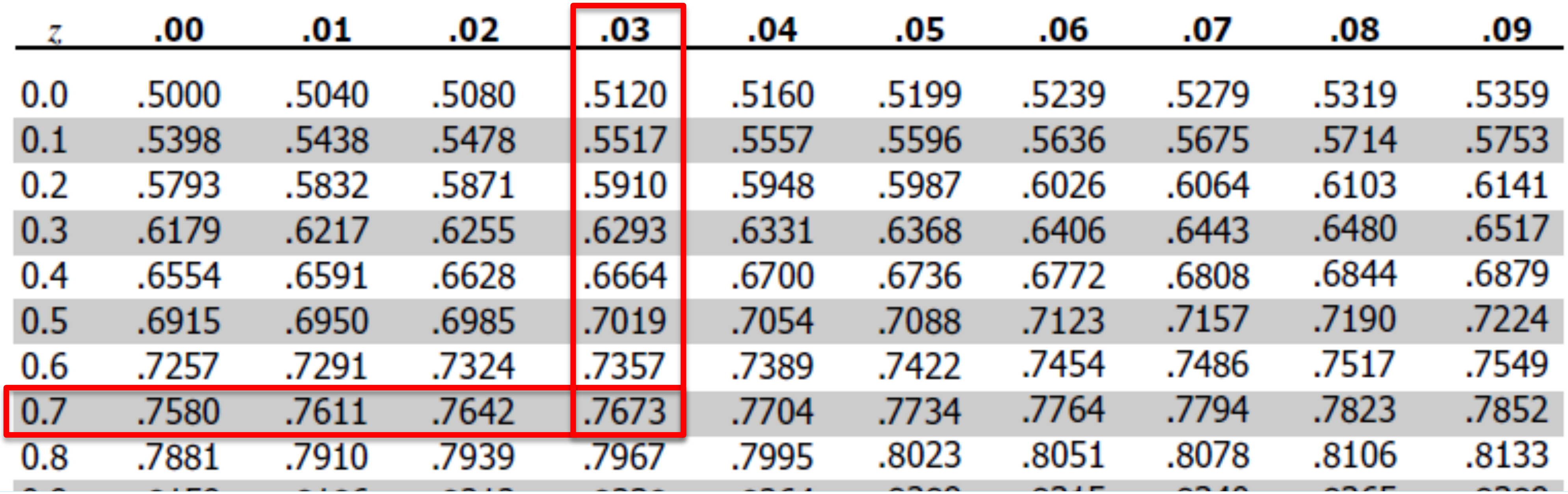
Now that we have the Z-score, we can find the area.
Z tables come in different formats, this one called the cumulative area, *which is the same as the percentile score.*

First, look up the area for our Z-score of 0.73 → .7673

So, at age 65, about 77% of the cases are *younger*.



Since the area = 1 we can then see that about 23% of the cases are age *65 <u>or older</u>.*



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133

Applying the normal distribution to actual data

$$Z = \frac{Y_i - \bar{Y}}{s_y}$$

Now suppose we want to find the proportion of individuals with ages between 30 and 65.

Age ~ *N*(54, 15)

$$-1.6 = \frac{30 - 54}{15}$$

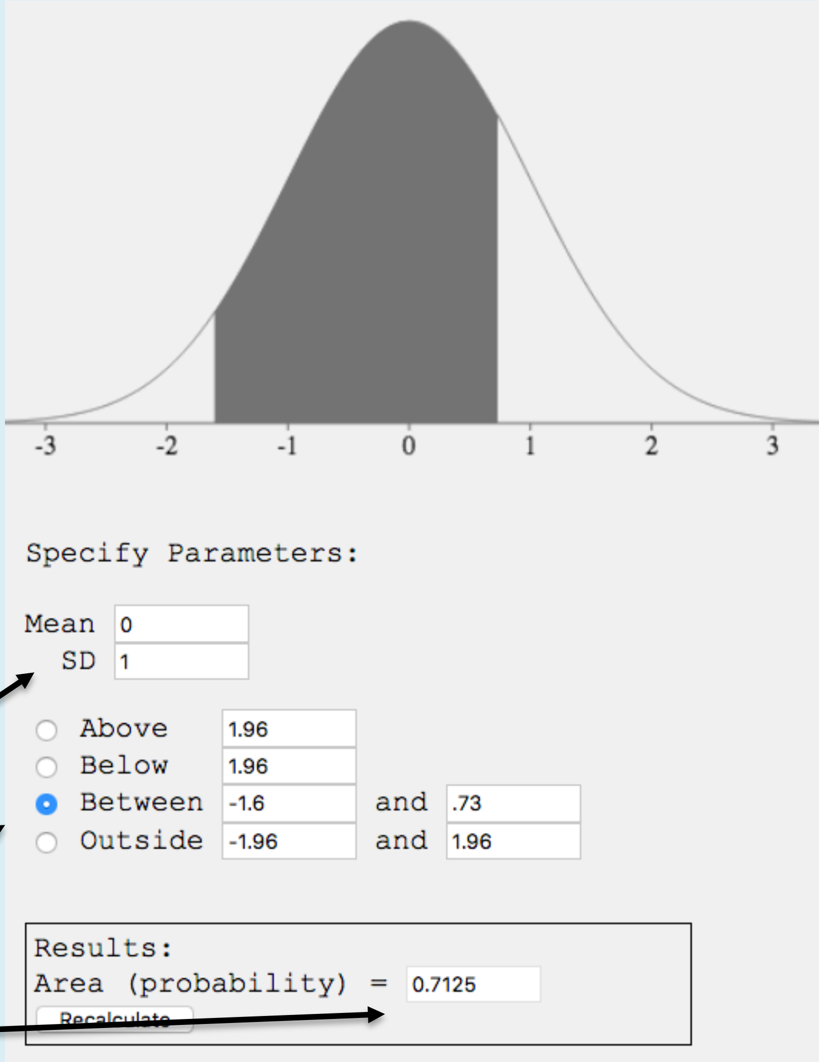
$$0.73 = \frac{65 - 54}{15}$$

Let’s use software for this …

http://onlinestatbook.com/2/calculators/normal_dist.html

$$-1.6 = \frac{30 - 54}{15}$$

$$0.73 = \frac{65 - 54}{15}$$



Specify Parameters:

Mean 0
SD 1

○ Above 1.96
○ Below 1.96
● Between -1.6 and .73
○ Outside -1.96 and 1.96

Results:
Area (probability) = 0.7125

1. Set the distribution to normal (0, 1)
2. Enter the low and high Z-scores
3. Visualize area and get value

71% of the cases are between 30 and 65

Finding Raw Values Corresponding to Normal Probabilities

Just work the formula backwards to solve for Y_{i}

$$Y_i = \bar{Y} + (Z * s)$$

Age ~ $N(54, 15)$

What age is at the 95th percentile?

Or, at what age or above lies 5% of the cases?

From table or software, the Z-score for area .95 = 1.645

So …

$54 + (1.65 * 15) \cong 78$

Finding Areas Under the Curve

1. ASK $P(Z > 2)$?

2. SKETCH

3. TABLE

4. INTERPRET

$P(Z < 2) = .95$

$1 - .95 = .05$ $P(Z > 2)$

ASK $P(Z > 2)$?

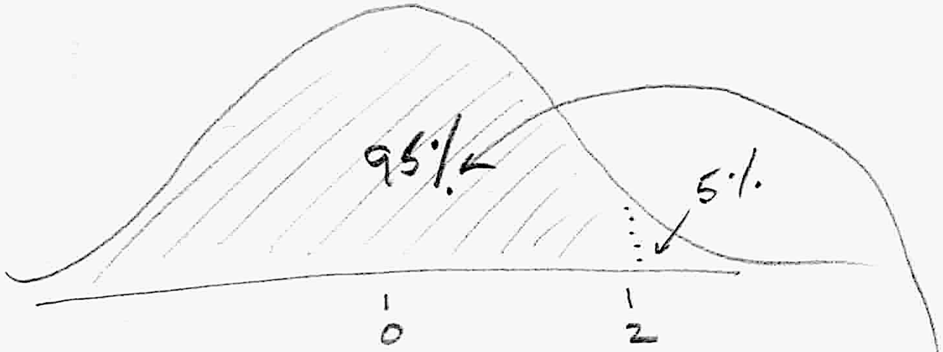


TABLE :

Z	.00	.01	. . .	.09
0.0	.50	.504	, , ,	.5359
0.1	.54			
:	:			
2.0	.95			
2.1				

Finding Probability for Given Value of X_i

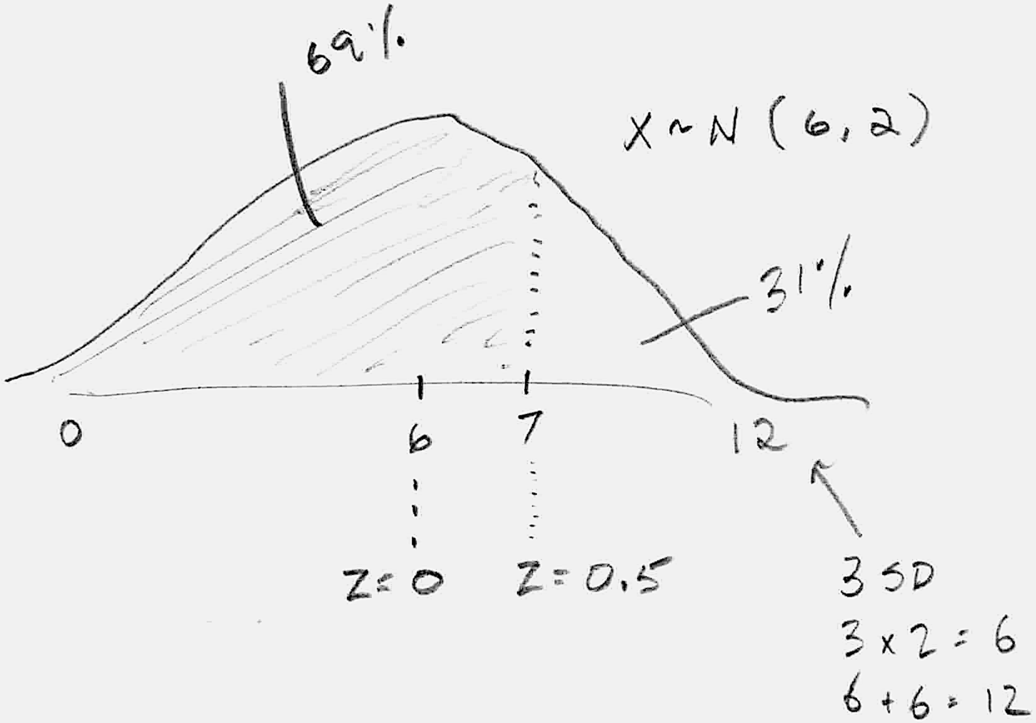
1. ASK $P(x_i > 7) = ?$
2. SKETCH
3. STANDARDIZE

$$Z = \frac{X_i - \bar{X}}{s} = \frac{7-6}{2} = 0.5$$

4. TABLE
5. INTERPRET

$P(x_i < 7) = .69$

$1 - .69 = .31 \quad P(x_i > 7)$



Z	.00	.01	...	.09
0.0	.5	.504		.536
0.1	.54			
0.2	.58			
0.3	.62			
0.4	.66			
0.5	.69			
0.6	.73			
⋮				

Key Terms

Standard Normal Distribution

Standard Normal Table

Standard (Z) Score

Area under the curve

Percentile

NOTE

There are two systems of symbols in statistics that work in parallel.

One uses the English alphabet.
For example, the mean is $\bar{X}$, and the standard deviation is s.

The other uses the Greek alphabet.
For example, mean is $\mu\ (\mathrm{mu})$ and the standard deviation is σ (sigma).

We’ll get into the details soon, but for now just to note that some of the videos use the Greek.

JTC270 **Lab 3** **Finding Areas** <JTC270 Lab Dataset.dta>

Formula: $Z = \frac{Y_i - \bar{Y}}{s_y}$ $Y_i = \bar{Y} + (Z * s)$

Links: Area widget http://onlinestatbook.com/2/calculators/normal_dist.html
Z Table http://www.z-table.com/

From our dataset we'll be working with one variable: rbratio. This is a measure of the way individuals in the study see the trade off between the potential risks of the uranium mine and the potential benefits.
rbratio $\approx N(2.063, 1.176)$ min = 0.22 max = 5 higher values indicate more risk than benefit

1) What is the probability that an individual has a score of 2.5 or less? $P(\text{rbratio} < 2.5)$
Sketch problem, calculate Z-score, get probability from Z table link.

2) What is the probability that an individual has an rbratio score between the mean and 2.5?
$P(\text{rbratio} > 2.063 < 2.5)$
Sketch problem, calculate Z-scores need two), get probability from Area widget link.

3). Find the 75th percentile on the distribution. This is also the value of rbratio at which $P(\text{rbratio} < .75)$.
Sketch problem, use Z table link to find Z for .75, calculate corresponding value for rbratio

In Stata. Open lab dataset. Here we will examine the distribution of rbratio, use software to create a new variable by calculating Z-scores for all 193 cases, examine that distribution, sort the dataset, find the Z-score that corresponds to an rbratio value of 4.5, and determine the probability of an individual in the study having an rbratio score less than 4.5.

rbratio $\approx N(2.063, 1.176)$ min = 0.22 max = 5 higher values indicate more risk than benefit

Examine distribution with histogram

Graphics --> Histogram
    Main: check data are continuous, select rbratio, Y axis = percent

4) What can we say about the shape of the distribution?

Calculate Z-scores. Note that since we have the summary statistics, it's simple to enter the Z-score equation.

Data --> Create or change data --> create new variable
    Main: Variable name: Zrbratio | Specify a value or an expression: (rbratio - 2.063) / 1.176
    OK

Examine the standardized distribution

Graphics --> Histogram
    Main: check data are continuous, select Zrbratio, Y axis = percent

5) How does this distribution differ from the distribution of rbratio?

Open the data browser (the Browse icon at top). scroll to the right to see the columns for rbratio and Zrbratio. Right-click on the top of the Zrbratio column, select Data, select Sort Data, OK

Scroll down to find where Zrbratio is closest to zero.

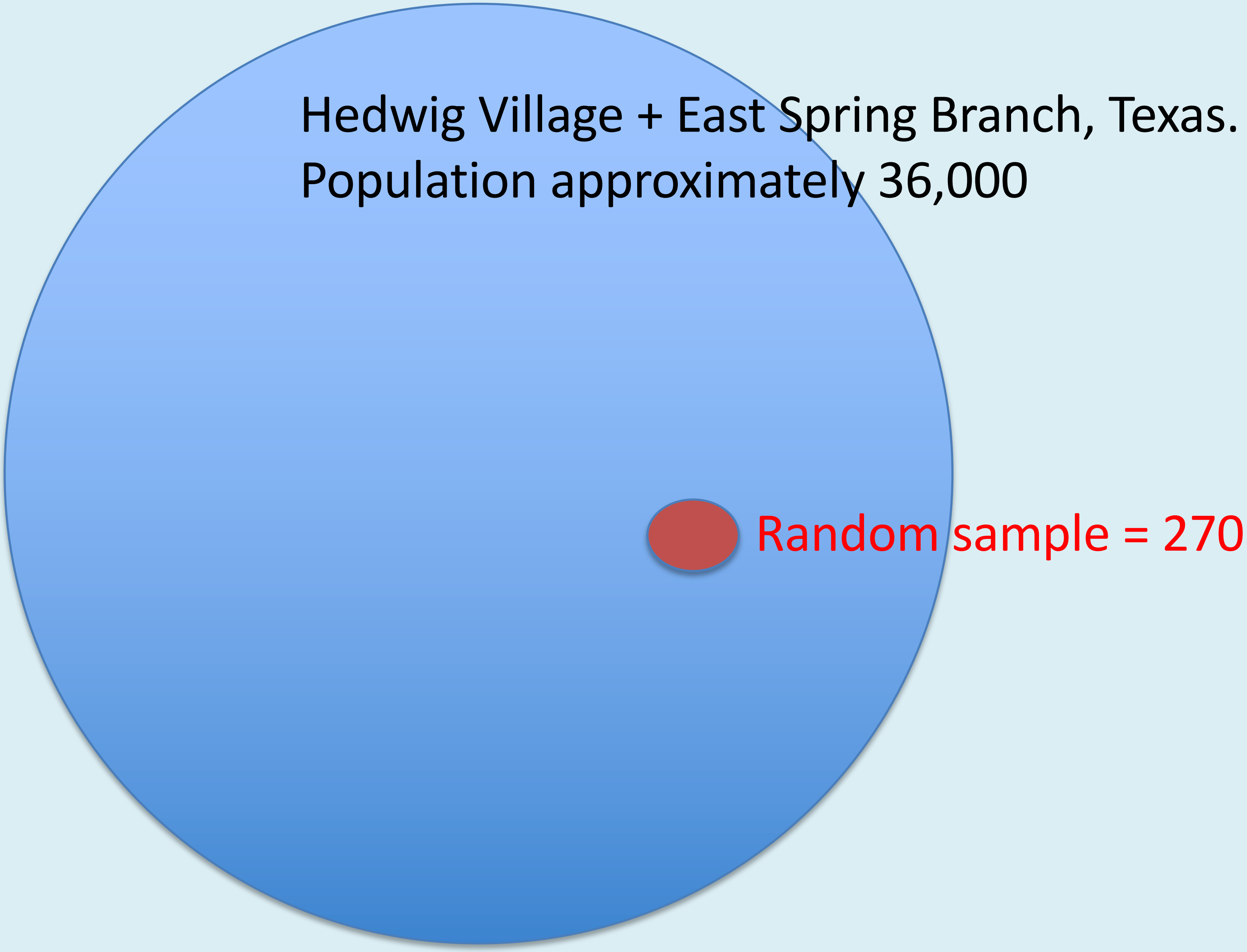
6) What can you say about the corresponding value of rbratio?

Scroll down to find the value 4.5 for rbratio.

7) What proportion of the distribution lies at 4.5 or less?
Use the area widget http://onlinestatbook.com/2/calculators/normal_dist.html

From Sample Data to Population Estimates

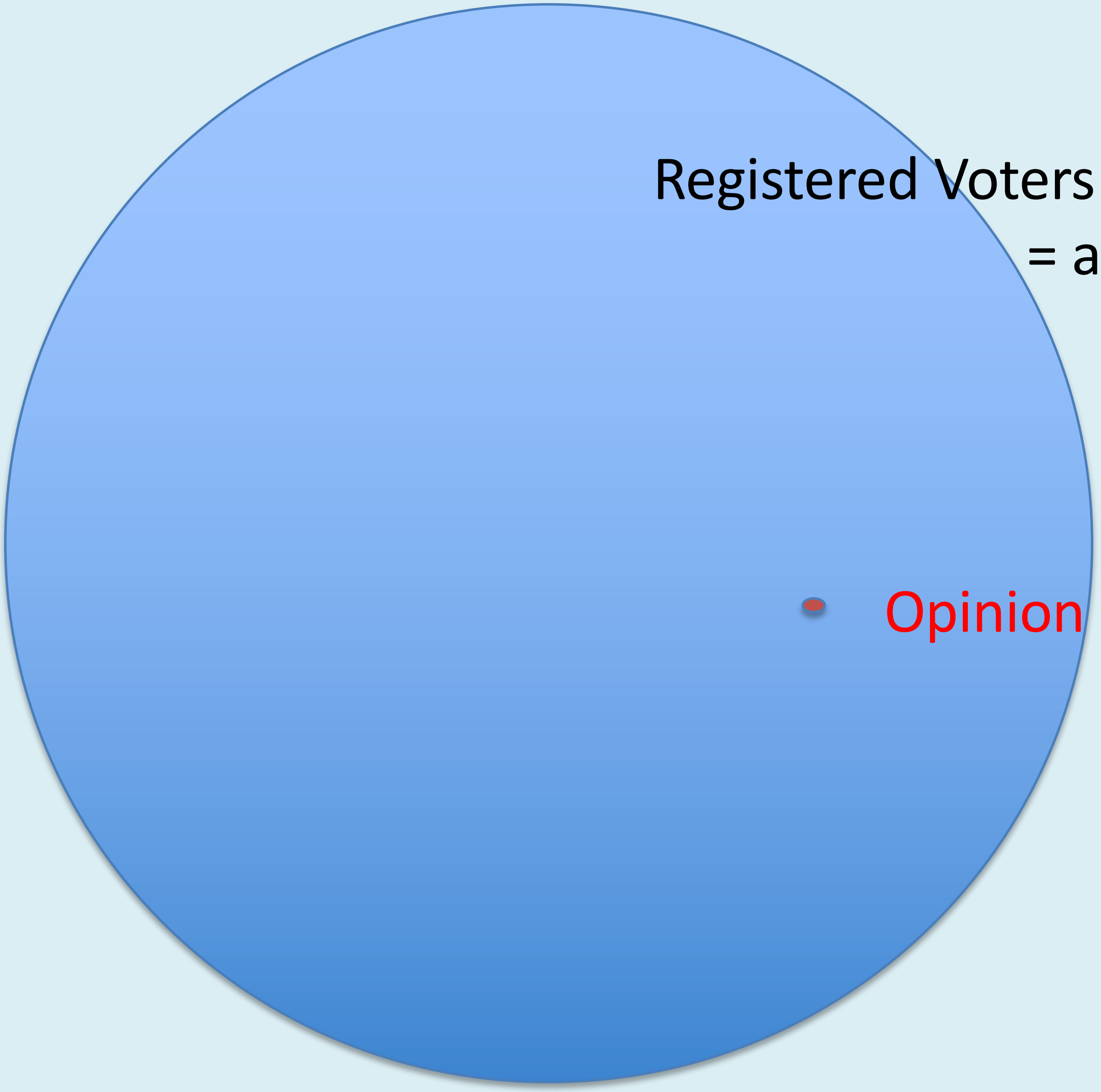
From Sample Data to Population Estimates



Hedwig Village + East Spring Branch, Texas.
Population approximately 36,000

Random sample = 270

From Sample Data to Population Estimates



Registered Voters in U.S. 2018
= approximately 153 Million

Opinion Poll sample = 500

Populations and Samples

Population:

All of the cases the researcher is interested in.

Parameter:

True central tendency and variance (unknowable).

Symbols:	μ	Mean
	σ	Standard Deviation
	π	Proportion

Populations and Samples

Sample:

A small selection from the population.

Statistic:

An estimate of the population characteristics.

Symbols:	$\bar{X}$	Mean	or M
	s	Standard Deviation	or SD
	p	Proportion	

Probability Sampling

In order for us to make the best estimate of the population using sample data, the sample must be done carefully.

Critically, samples must:

Be random and representative;

Each individual has and equal chance of selection;

Any combination (group) of individuals has and equal chance of selection.

Probability Sampling

Populations can be complicated!

So there are many approaches to drawing useful samples.

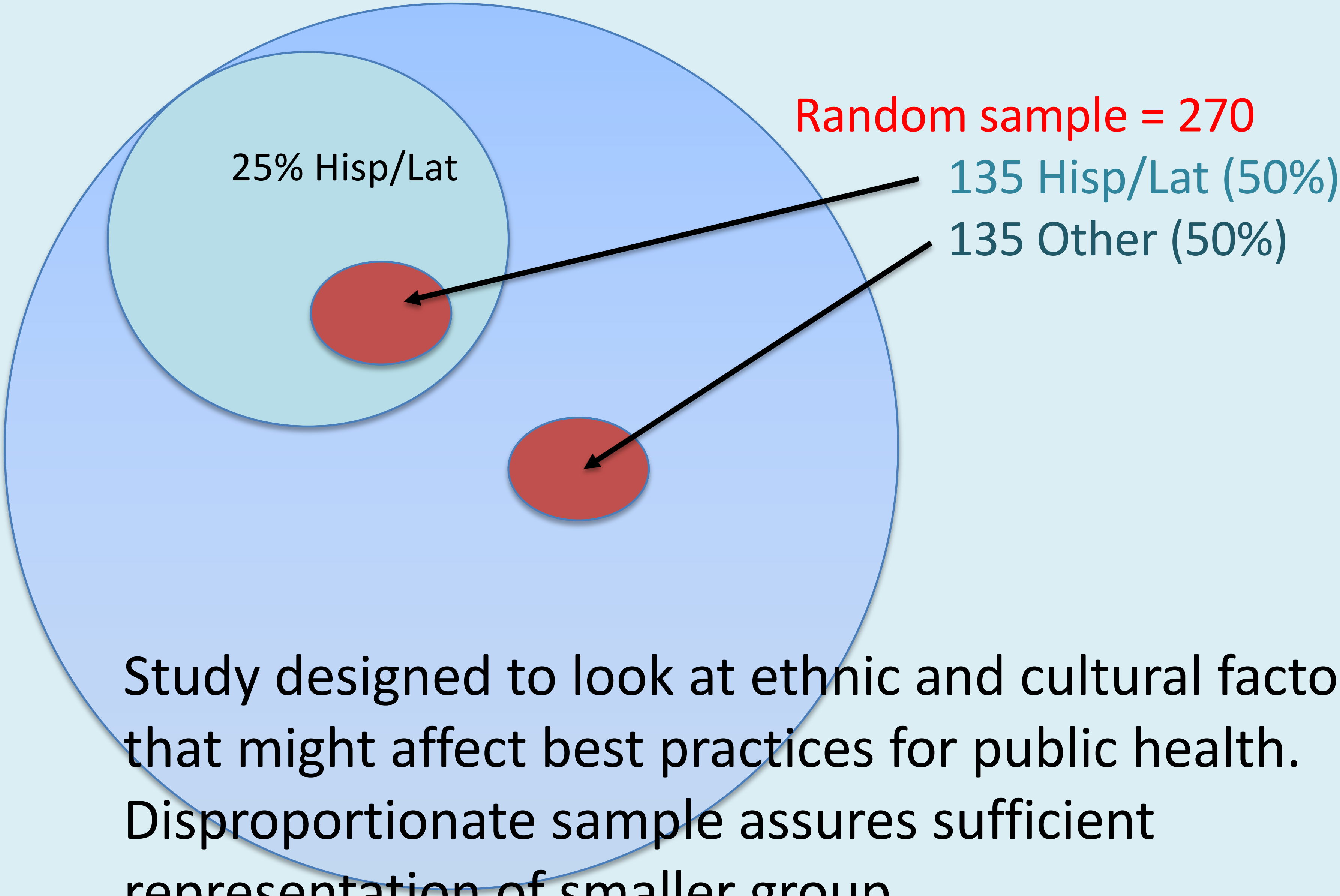
Simple: use random numbers (hat draw).

Systematic: Randomly ordered list of population, select every k^{th} individual.

Stratified: Randomly select *k* groups, then randomly select *k* individuals from each selected group.
Groups may be sampled to match their proportion in the population, or may be disproportionally sampled (often over-sampled).

Probability Sampling

Hedwig Village + Spring Branch, Texas. Population appx 36,000



25% Hisp/Lat

Random sample = 270

135 Hisp/Lat (50%)

135 Other (50%)

Study designed to look at ethnic and cultural factors that might affect best practices for public health. Disproportionate sample assures sufficient representation of smaller group.

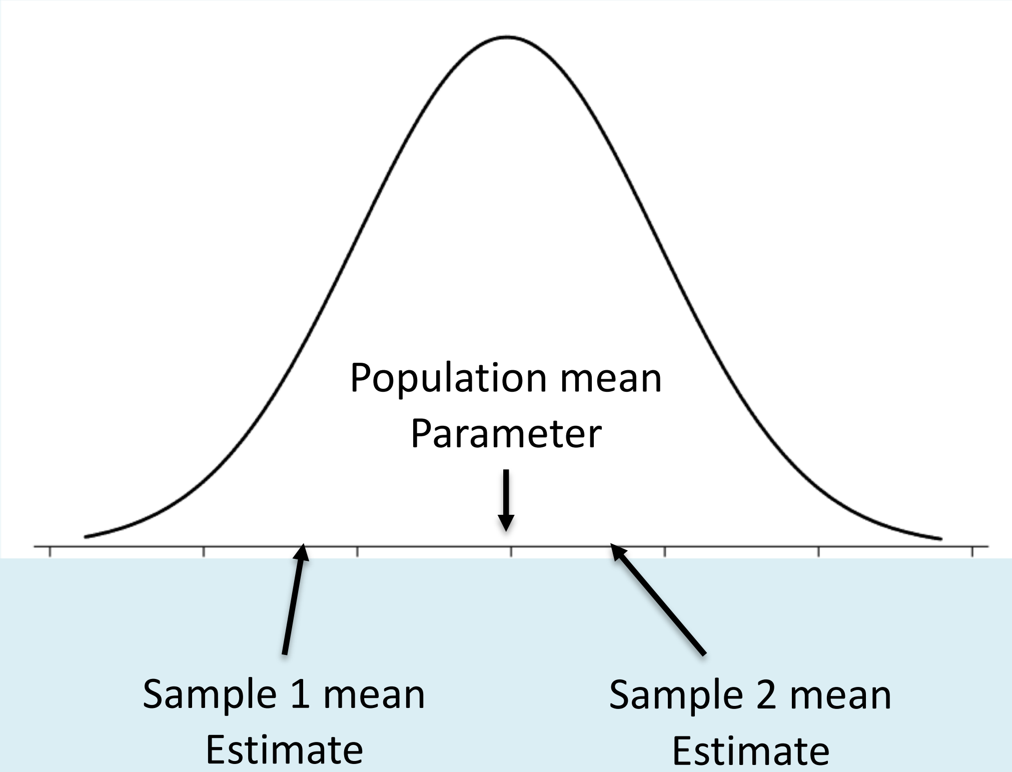
Sampling Distributions

An appropriate random sample from a population allows us to evoke a statistical principal called the Central Limit Theorem.

While mathematically complicated, the basic idea is approachable: If you draw a random sample of size N from a very large population you can find the mean for the sample.

We draw a second sample of size N, and get that mean.

Now we have two means.



Population mean
Parameter

Sample 1 mean
Estimate

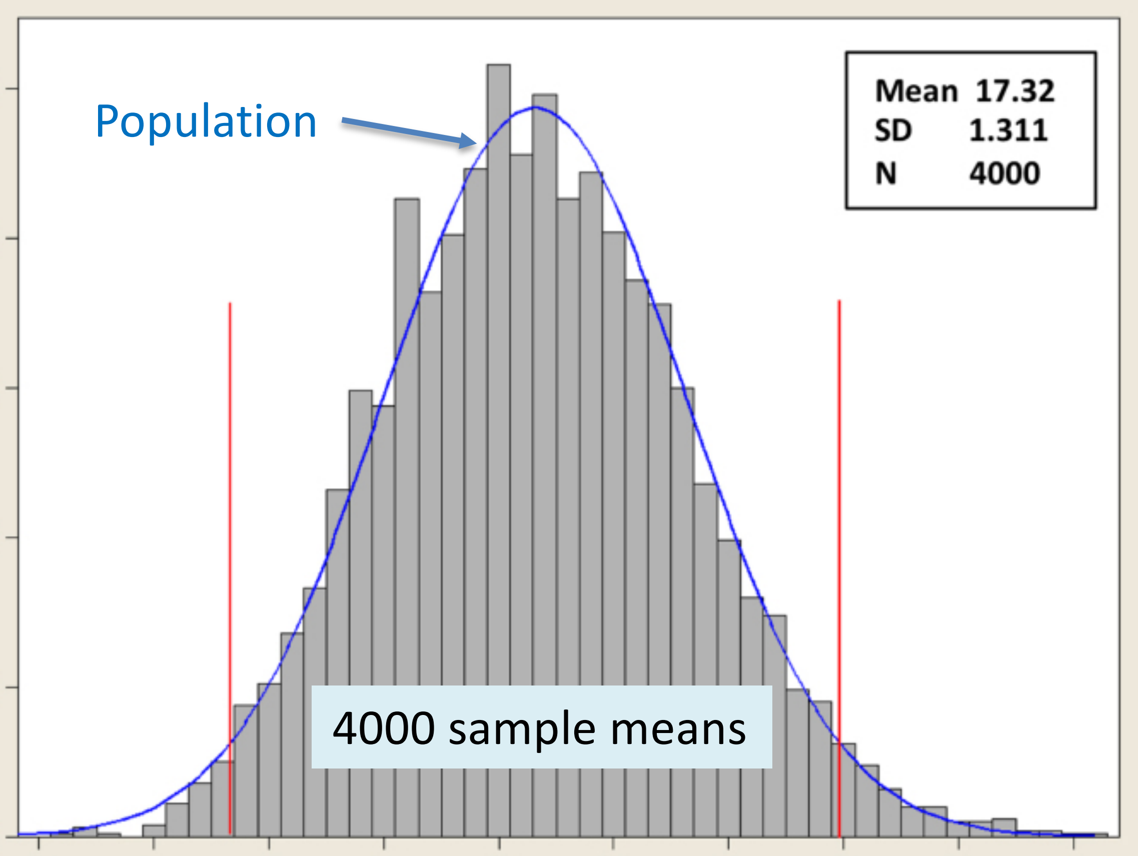
Sample 2 mean
Estimate

Sampling Distributions

Repeat that a large number of times.

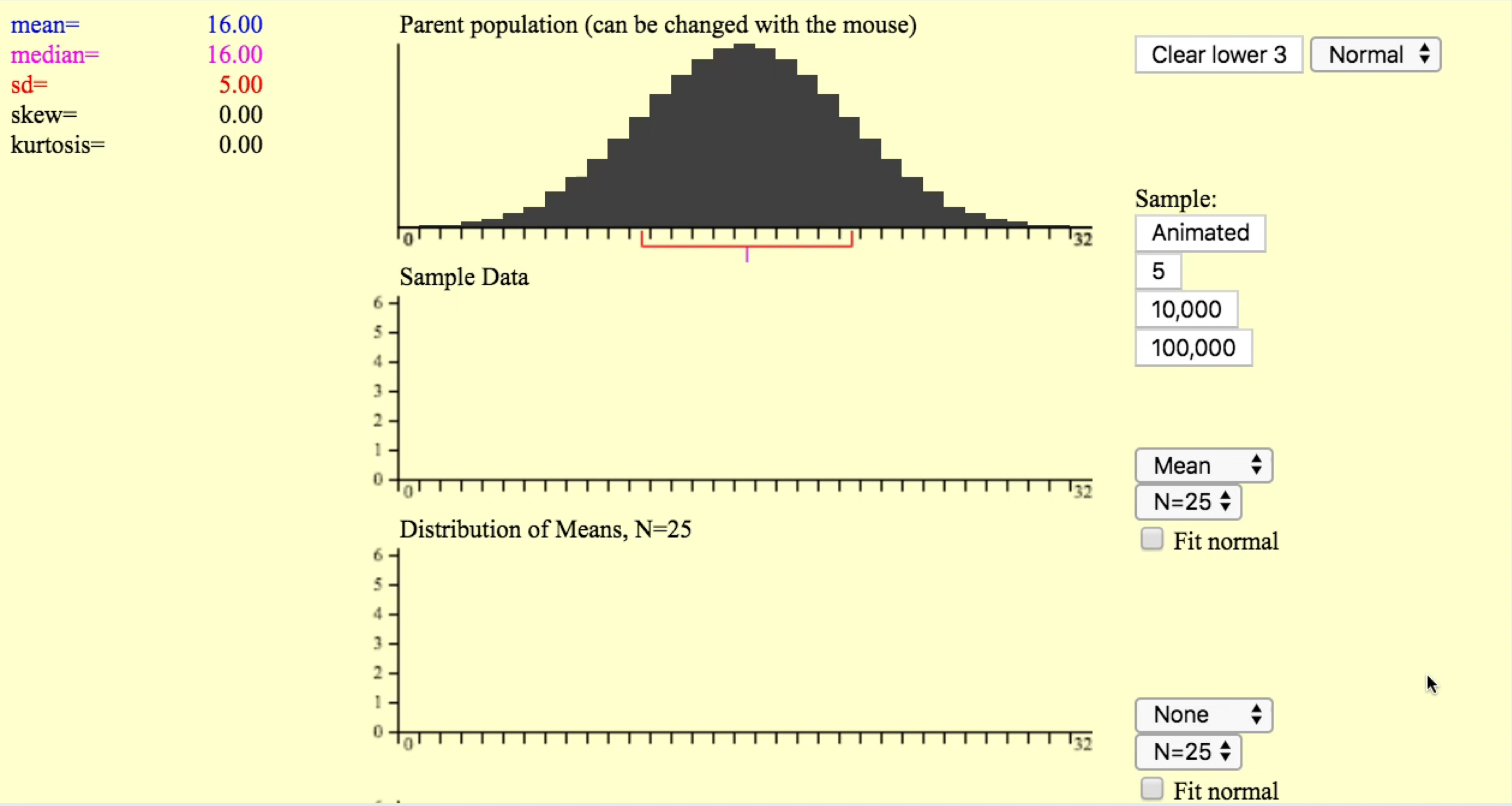
If we plot the distribution of those means, it will be normal (regardless of the actual shape of the population distribution!).

And the mean of the distribution of means will match the pop. mean.



This is called the sampling distribution of the mean.

Sampling Distributions



mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00

Let’s put this to work: Estimation

We select a random sample from a population and use a sample statistic to estimate a population parameter.

This is called a point estimate, our best guess at one number. We can make point estimates about any parameter (mean, standard deviation, the difference between two means, etc.)

BUT it is an *estimate!*

How good is the estimate?
We can calculate that from the sample data.
It’s called a confidence interval, or “margin of error”

Let's put this to work: Estimation

Must decide how confident we need to be about the estimate. How often would it be OK to be wrong? Typically 5%. We call that value alpha (α)

This is the 95% confidence **level**
(1 – α since we want to be right most of the time).

Because it's a probability, it has a Z-score.

Since we might be wrong in either direction, to get the Z score we divide alpha in half (.025), subtract from 1 (.975) and consult the normal table where P .975 → $Z = 1.96$

This is also called the critical value for Z.

Let’s put this to work: Estimation

How do we use this Z-score to get a confidence interval?

$$\mathrm{CI} = \bar{\mathrm{X}} \pm Z * SE_{\bar{x}}$$

The point estimate plus-or-minus the Z-score for our desired confidence level times the standard error.

Thanks to the Central Limit Theorem we know that the standard deviation in the population is smaller than the one we have in our data. So we must make an adjustment.

Sample standard deviation

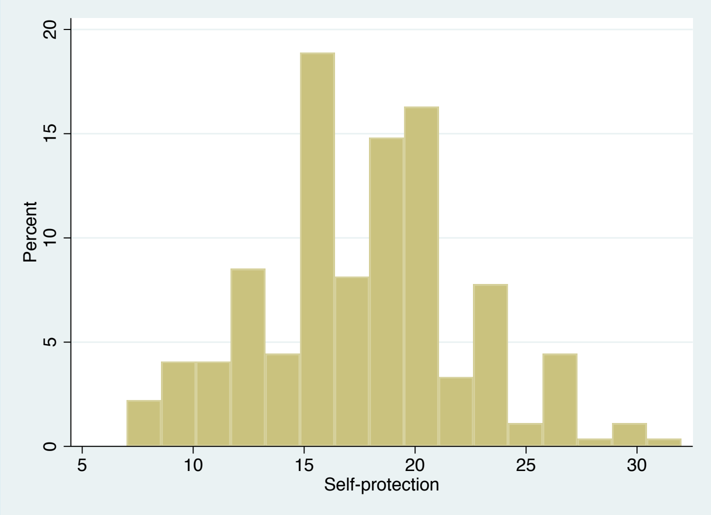
$$SE_{\bar{x}} = \frac{SD}{\sqrt{N}}$$

Let’s put this to work: Estimation

$$\text{CI} = \bar{\text{X}} \pm Z * SE_{\bar{x}} \quad \textit{where} \quad SE_{\bar{x}} = \frac{SD}{\sqrt{N}}$$

Since we’re working with a mean, we must have interval-level data. Our score for mosquito self-protection:

protect ~N (17.7, 4.7)



Z = 1.96

N = 270, Sqrt N = 16.4

SE = 4.7 ÷ 16.4 = 0.29

CI = 17.7 +/- 1.96 * 0.29 = 0.57

Let’s put this to work: Estimation

Z = 1.96

N = 270, Sqrt N = 16.4

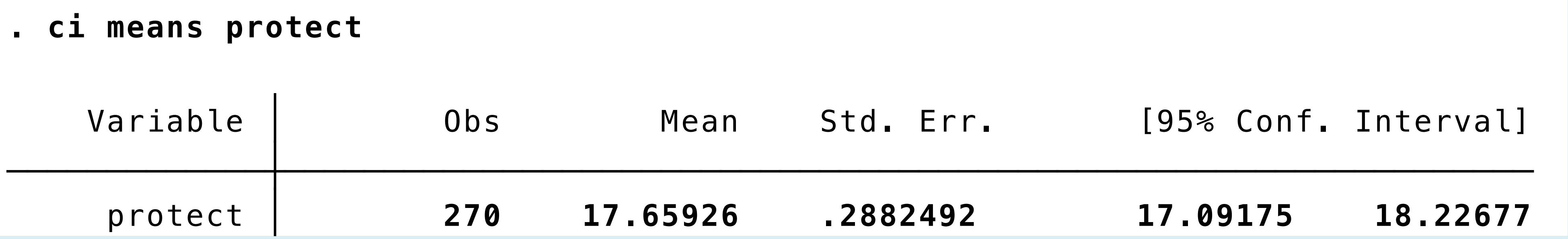
SE = 4.7 ÷ 16.4 = 0.29

$$\mathrm{CI} = \bar{\mathrm{X}} \pm Z * SE_{\bar{x}}$$

CI = 17.7 +/- 1.96 * 0.29 = 0.57

Thus, we can state that based on our data we are 95% confident that the true population value for protect is 17.7 plus-or-minus 0.57.

Or, we can say that at alpha = .05 the estimated mean is 17.7 (CI 17.3, 18.27)



```
. ci means protect

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     protect |        270    17.65926    .2882492        17.09175    18.22677
```

We can also use this for proportions, nominal data

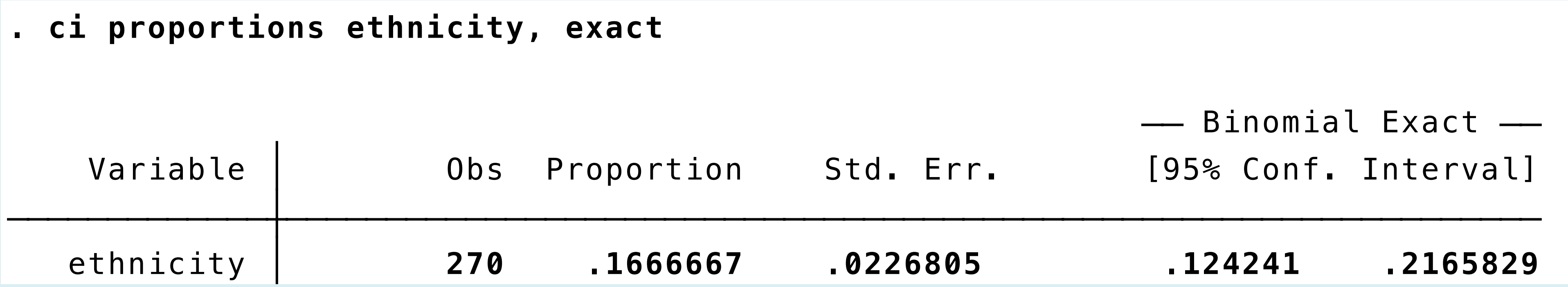
$$\mathrm{CI} = p \pm Z * SE_p \qquad \textit{where} \qquad SE_p = \sqrt{\frac{p(1-p)}{N}}$$

N = 270 17% report to be Hispanic/Latino

$$Z = 1.96 \qquad SE_p = \sqrt{\frac{.17\,(1\ -\ .17)}{270}} = .023$$

$$\mathrm{CI} = .17 +/- 1.96 * 0.023 = 0.045$$

We are 95% confident that the true proportion for ethnicity is 17% plus-or-minus 4.5%. Or, we can say that at alpha = .05 the estimated proportion is 17% (CI 12.5%, 21.5%).

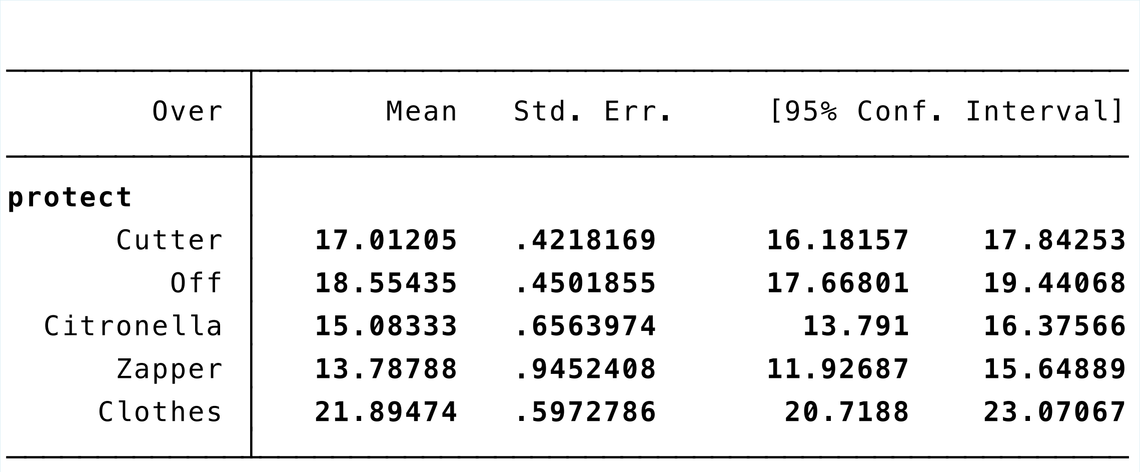


```
. ci proportions ethnicity, exact


                                                         -- Binomial Exact --
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ethnicity |        270    .1666667    .0226805         .124241    .2165829
```

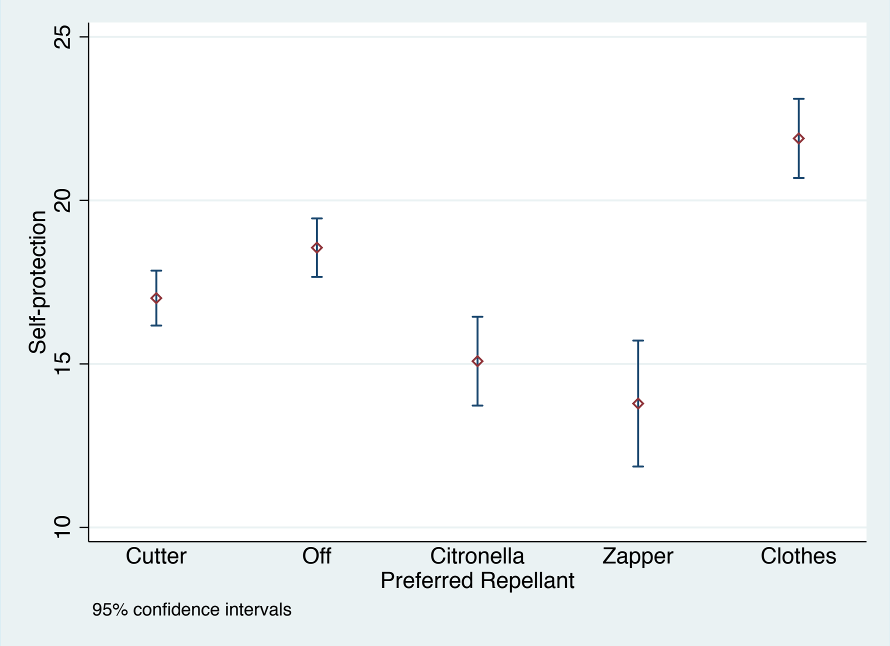
Interpreting Confidence Intervals

We can compare confidence intervals between two variables measured in the same way.

Or much more common: one variable across groups:



```
--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
protect      |
      Cutter |   17.01205   .4218169      16.18157    17.84253
         Off |   18.55435   .4501855      17.66801    19.44068
  Citronella |   15.08333   .6563974        13.791    16.37566
      Zapper |   13.78788   .9452408      11.92687    15.64889
     Clothes |   21.89474   .5972786       20.7188    23.07067
--------------------------------------------------------------
```



We'll do this more formally, but you can see that it's very clear that the population mean for the Clothes group is higher than the other four, and the others have enough overlap that they may not be different.

Confidence Interval Width Factors:

N and SE:

Larger samples make smaller standard errors, narrower CIs

Confidence Level:

If we select a higher level of confidence, say 99%,
the the Z-score that we get is larger, making the CI wider.

For our mean on the protect score, if:

95% confidence	N = 270	CI +/- .57
	N = 50	CI +/- .67
	N = 500	CI +/- .21

N = 270, SE = .29	CL = 95%	Z = 1.96	CI +/- .57
	CL = 90%	Z = 1.64	CI +/- .48
	CL = 99%	Z = 2.58	CI +/- .74

Key Terms

Parameter
Population
Sample
Statistic
Estimation
Point estimate
Sampling Dist. of the Mean
Standard Error of the Mean
Probability Sampling
Simple Random Sample
Systematic Random Sampling
Stratified Random Sample
Proportionate Stratified Sample
Disproportionate Stratified Sample
Central Limit Theorem
Confidence level
Confidence interval
Margin of error

JTC270 Lab 4 Confidence Intervals <JTC270 Lab Dataset.dta>

$$\mathrm{CI} = \bar{\mathrm{X}} \pm Z * SE_x \qquad SE_x = \frac{SD}{\sqrt{N}} \qquad \mathrm{CI} = p \pm Z * SE_p \qquad SE_p = \sqrt{\frac{p(1-p)}{N}}$$

For this portion of the Lab we'll work with a three "generic" variables:
X is an interval measure, Y (0/1) and Group (A/B) are a nominal variables Use the 95% CL where Z = 1.96

For X ~ N (30, 2) N = 200

Within Group A X ~ N (20, 2) N = 100 In Canvas enter Q1) The standard error Q2) The confidence interval

Within Group B X ~ N (40, 3) N = 100 In Canvas enter Q3) The standard error Q4) The confidence interval

Q5) What can you observe about the two Confidence Intervals above?
a) both are normal b) they do not overlap c) they are equal d) they overlap

Y is a nominal variable with values 0/1, it's p is .73 (73% of the cases are 1s) N = 200

Within Group A p of Y = .56 N = 50 In Canvas enter Q6) The standard error Q7) The confidence interval

Within Group A p of Y = .40 N = 150 In Canvas enter Q8) The standard error Q9) The confidence interval

Q10) What can you observe about the two Confidence Intervals above?
a) both are normal b) they do not overlap c) they are equal d) they overlap

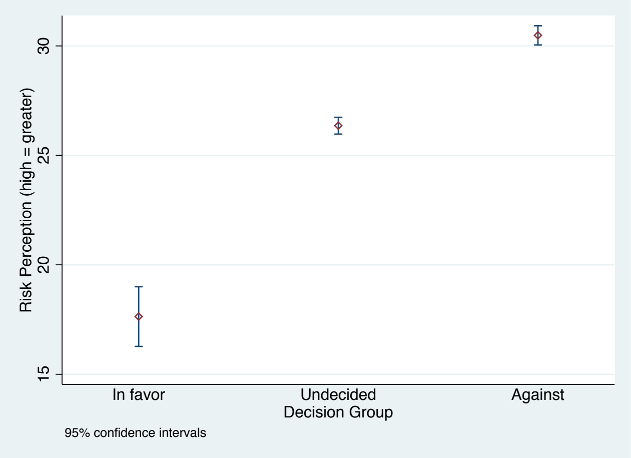
In Stata. We'll briefly look at the variable <riskper> that described perception of risk for the mining operation (high values = riskier). riskper ~N (26.4, 1.4). Here we simply want to look at the confidence intervals for riskper over three groups using the variable <decide> (1 = in favor, 2 = neutral, 3 = against).

Statistics --> Summaries tables tests --> Confidence Intervals
Main tab: click on button for means, enter riskper for Variables
by/if/in tab: click repeat commands by groups, enter variable decide *Submit*

This is a plot of the results:

Q11) The CI for the "in favor" group is widest. From what you can observe in your output which of these is likely the cause?
a) in favor has smallest SD
b) in favor has the lowest mean value
c) in favor has the smallest N
d) in favor is more normal

Q12) Since the CIs do not overlap, which of these might be true?
a) on average everybody is undecided
b) the variance in each group is about equal
c) no conclusions can be drawn
d) the probability of the group means being equal is small



Testing Hypotheses

with confidence intervals

Statistical Hypothesis Testing

Given that we now have a tool (probability) to evaluate how confident we are about an estimate of a population parameter made with sample data, we can make formal questions about the population.

Such formal questions are hypotheses.

Statistical Hypothesis Testing

A single hypothesis exists in two forms, which mirror one another:

The null hypothesis, if supported, tells us that we cannot accept the results (whatever they are) at our selected confidence level.

The alternative hypothesis (also called the research hypothesis) may be accepted if we have evidence that the null hypothesis is wrong, that it can be rejected.

Statistical Hypothesis Testing

In practice we usually focus on the research hypothesis.

The research hypothesis is a clear, simple, direct statement of what we have reason to think is true about the population, based on sample data.

Statistical Hypothesis Testing

Often expressed in terms and symbols representing the population, so here we have Greek letters showing up.

Our first, most basic hypothesis test will be to use the confidence interval to evaluate our population point estimate against a known value.

We'll do for for means and for proportions.

Statistical Hypothesis Testing

Consider our sample of 270 in two Houston neighborhoods. Is this population representative of the entire Houston area with respect to age?

We might estimate our neighborhood population as being older, or younger, that the whole city.

This is called a two-tailed test.

One-tailed tests would specify our neighborhood being older, or younger (exclusive).

Statistical Hypothesis Testing

The two-tailed or non-directional hypothesis will be our default approach. The research hypothesis:

The age in our neighborhoods is equal to the average age in Houston.

$$H_a: \mu_1 = \mu_2$$

where $\mu_1 =$ mean age neighborhood

$\mu_2 =$ mean age Houston

Statistical Hypothesis Testing

We have census data, which is not an estimate but a known value, that tells us the average age in Houston is approximately 35 years. We can substitute this into the hypothesis:

H_a: $\mu_1 = 35$

where $\mu_1 =$ mean age neighborhood

We'll also state our confidence level with the hypothesis. In this case, we stay at the 95% level.

Statistical Hypothesis Testing

H_a: $\mu_1 = 35$ Let's look at the data:

Neighborhood age:

M = 54,   SD = 16,   SE = .97,   95% CI = 52, 56

Note that we could have easily found the 95% confidence interval:

for Z = 1.96

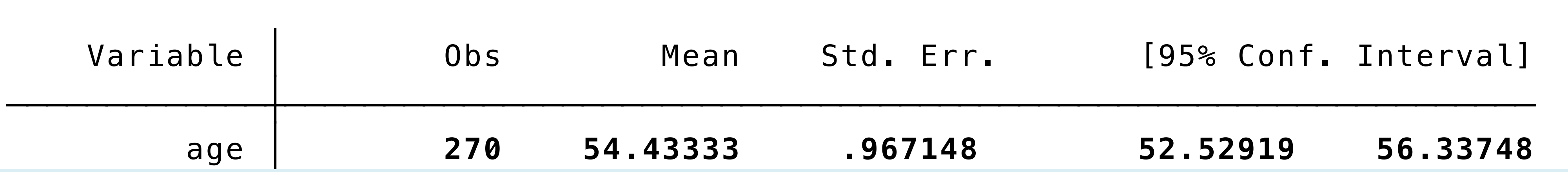
M +/- 1.96 * .97   round that to 2 X 1 = 2

54 - 2 = 52

54 + 2 = 56

Statistical Hypothesis Testing

H_a: $\mu_1 = 35$ Let’s look at the output:



```
    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
         age |       270    54.43333     .967148        52.52919     56.33748
```

Our point estimate is 54 years, plus/minus 2.

Or, we’re 95% confident that the true population value for the neighborhood lies somewhere between 52 and 56. 35 does not fall in this range.

SO the **null** hypothesis $\mu_1 \neq 35$ is not rejected.

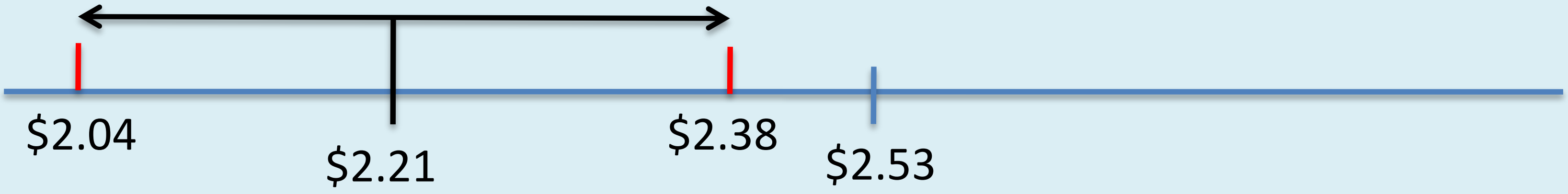
The **alternative** hypothesis is not accepted.

Statistical Hypothesis Testing

M national price of gas/gal: $2.53 (value known)

M price of gas/gal in CO from random sample of 500 stations = $2.21 using the estimated standard error and selecting alpha = .05 we find that the confidence interval is $0.17 stated $2.21 (2.04, 2.38).

Does the test value (national average) fall within the CI?



M = 2.21, SE = .087 CI = $2.21 +/- 1.96(.087) = .17

Statistical Hypothesis Testing

Let's look at proportions.

The approach for hypotheses is the same.

H_a: $\pi_1 = \pi_2$

where $\pi_1 =$ % respondents report Hispanic/Latino

$\pi_2 =$ % Hispanic/Latino in Houston (44%)

Statistical Hypothesis Testing

H_a: $\pi_1 = .44$ Let's look at the data:

Neighborhood percentage:

p = .17,   SD = .37,   SE = .02,   95% CI = .13, .21

Note that we could have easily found the 95% confidence interval:

for Z = 1.96

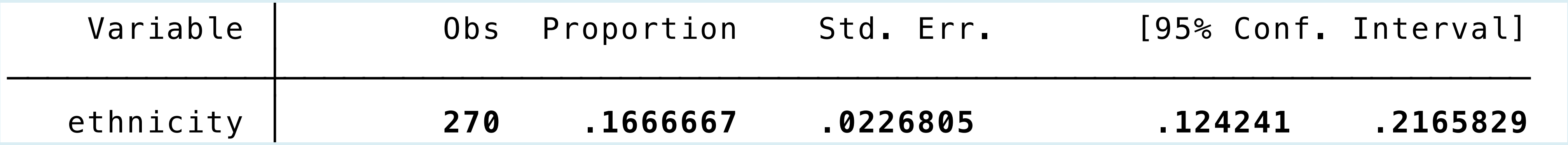
p +/- 1.96 * .02   round that to 2 X .02 = .04

.17 - .04 = .13

.17 + .04 = .21

Statistical Hypothesis Testing

H_a: $\pi_1 = .44$ Let’s look at the output:



```
    Variable |        Obs  Proportion    Std. Err.      [95% Conf. Interval]
-------------+--------------------------------------------------------------
   ethnicity |        270    .1666667    .0226805       .124241     .2165829
```

Our point estimate is .17 (or 17%), plus/minus 4%.

We’re 95% confident that the true neighborhood population value lies between 12.4% and 21.6%.

44% does not fall in this range.

SO the alternative hypothesis is rejected.

Statistical Hypothesis Testing

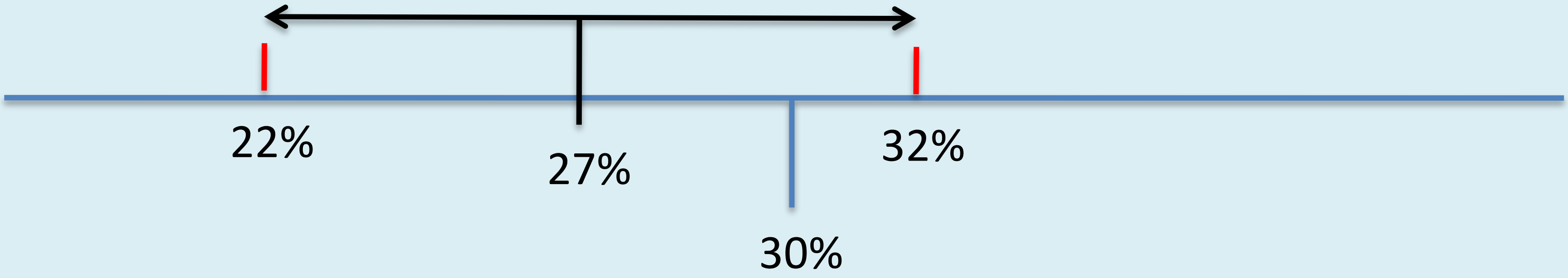
gas stations in US with grocery points: 30% *known value*

Ho: $\mu_1 \neq \mu_2$ *Ha:* $\mu_1 = \mu_2$

% in CO from random sample of 250 stations = 27%

p = .27 p(1-p)/n = .0008 SE = .028

95% CI = .27 +/- 1.96 * .028 = .054 or 5%



Does the test value (national proportion) fall within the CI?

Yes! We can reject the null hypothesis, and accept the alternative.

Review: still must find the CI … means

Z = 1.96

N = 270, Sqrt N = 16.4

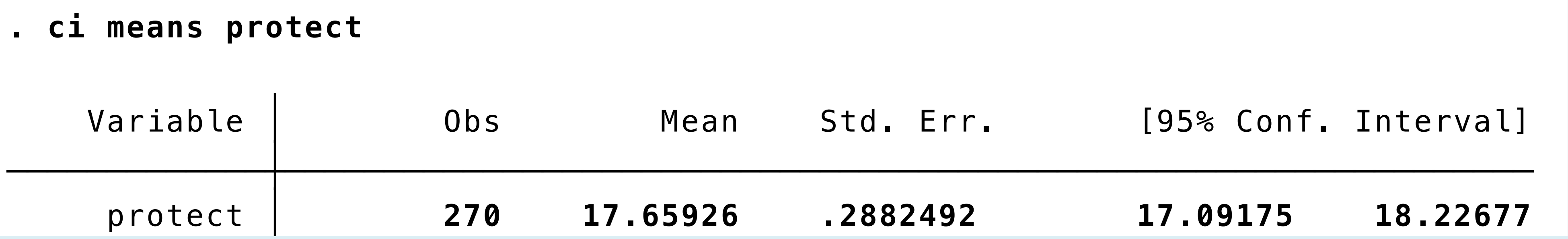
SE = 4.7 ÷ 16.4 = 0.29

$$CI = \bar{X} \pm Z * SE_{\bar{x}}$$

CI = 17.7 +/- 1.96 * 0.29 = 0.57

Thus, we can state that based on our data we are 95% confident that the true population value for protect is 17.7 plus-or-minus 0.57.

Or, we can say that at alpha = .05 the estimated mean is
17.7 (CI 17.3, 18.27)



```
. ci means protect

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     protect |        270    17.65926    .2882492        17.09175    18.22677
```

Review: still must find the CI … proportions

$$\mathrm{CI} = p \pm Z * SE_p$$

$$SE_p = \sqrt{\frac{p(1-p)}{N}}$$

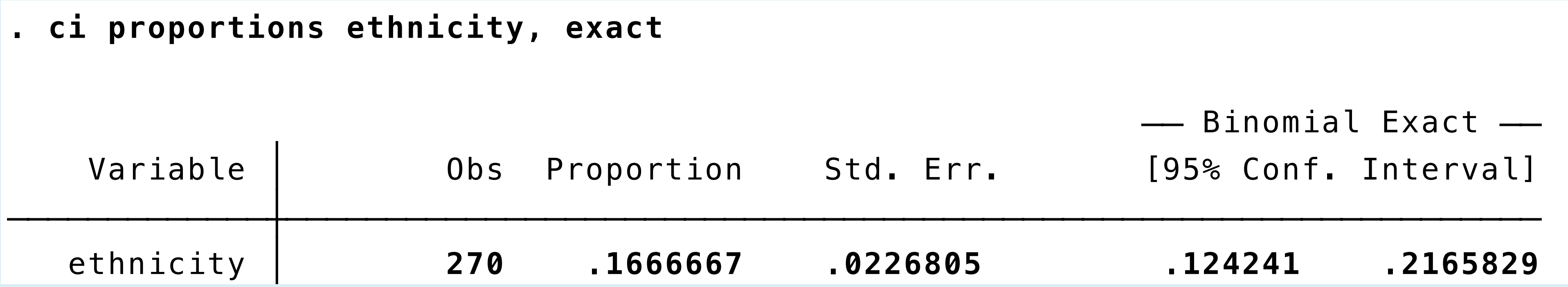
N = 270 17% report to be Hispanic/Latino

Z = 1.96

$$\mathrm{SE_p} = \sqrt{\frac{.17\,(1 - .17)}{270}} = .023$$

CI = .17 +/- 1.96 * 0.023 = 0.045

We are 95% confident that the true proportion for ethnicity is 17% plus-or-minus 4.5%. Or, we can say that at alpha = .05 the estimated proportion is 17% (CI 12.5%, 21.5%).



```
. ci proportions ethnicity, exact


                                                         — Binomial Exact —
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   ethnicity |        270    .1666667     .0226805         .124241    .2165829
```

Hypothesis Testing Notes …

One-Tailed Tests

A “directional” hypothesis test where the alternative is stated in such a way that the probability is entirely in one tail of a sampling distribution. $H_a: \mu_1 > \mu_2$

Z values are smaller:

95% Z_{stat}	2-tail = 1.96
	1-tail = 1.64

so CIs are narrower

Not used unless direction can be strongly justified.

Hypothesis Testing Notes …

Recall, these are estimates, and there's always a chance we might be wrong!

Type I Error

The probability associated with rejecting a null hypothesis when its true (false positive)

Type II Error

The probability associated with failing to reject a null hypothesis when it is false (false negative)

Hypothesis Testing Notes …

1. Making assumptions

2. Stating the hypotheses and selecting alpha/CL

3. Specifying the test statistic

4. Computing the test statistic

5. Making a decision and interpreting the results

Key Terms

- Null hypothesis
- Research hypothesis
- One-tailed test
- Two-tailed test
- Type I Error
- Type II Error

JTC270 Lab 5 Testing Hypothses with Confidence Intervals <JTC270 Lab Dataset.dta>

$$\mathrm{CI} = \overline{\mathrm{X}} \pm Z * SE_x \qquad SE_x = \frac{SD}{\sqrt{N}} \qquad \mathrm{CI} = p \pm Z * SE_p \qquad SE_p = \sqrt{\frac{p(1-p)}{N}}$$

This lab does not introduce any new formulae, but uses those from the previous lab to test hypotheses about population estimates from sample data, as compared to known population parameters.

A. Testing a sample mean against a parameter

In survey reseach demographic variables such as age are typically included, which allows the sample estimate to be compared against a known parameter from, often, the Census. Let's consider a random sample of adult individuals in Larimer County, N = 200. The mean for age in the sample is 45 years, with a standard deviation of 10. We also have the most recent value for age in Larimer County from the US Census. It is not an estimate, but a parameter. The mean is 35 years. We'll use the two-tailed 95% CL where Z = 1.96, and the convention where the sample mean us μ_1 and the census mean is μ_2.

Q1) Given this scenario, what would we want to test for a non-directional alternative hypothesis?

a) $\mu_1 \neq \mu_2$ b) $\mu_1 = \mu_2$ c) $\mu_1 > \mu_2$ d) $\mu_1 < \mu_2$

Q2) What is the confidence interval around our point estimate for age from the sample? CI = +/- _______.

Q3) What can we say about the value of age in our sample versus the census?

a) The true population value is within the 95% CI of our sample, sample looks good
b) The true population value is not within the 95% CI of our sample, sample looks not so good

B. Testing a sample proportion against a parameter

Here we have a study about consumption of red meat (which according to today's news is good for you). It's a small study based in the city of Big Hat, Montana (no deragatory intent!). 50 households were randomly sampled and asked what percent of all their meals included beef: 76%. It turns out that the Big Hat Grocers' Association recently checked with all 500 households and found the figure to be lower, 70%.

We'll use the two-tailed 95% CL where Z = 1.96, and the convention where the sample proportion us π_1 and the parameter proportion is π_2.

Q4) Given this scenario, what would we want to test for a non-directional alternative hypothesis?

a) $\pi_1 \neq \pi_2$ b) $\pi_1 = \pi_2$ c) $\pi_1 > \pi_2$ d) $\pi_1 < \pi_2$

Q5) What is the confidence interval around our point estimate for beef consumption? CI = +/- _______.

Q6) What can we say about the value of beef consumption in our sample versus the city census?

a) The true population value is within the 95% CI of our sample, sample looks good
b) The true population value is not within the 95% CI of our sample, sample looks not so good

Open our semester dataset in Stata.

C. Testing a sample mean against a parameter with Stata
From Census data we know that the mean age in Larimer County is 35 years.

Statistics --> Summaries Tables and Tests --> Summary and Decriptive Statistics --> Confidence Intervals
check: means
variables: age *submit*

Q7) What can we say about the value of age in our sample?
a) The true population value is within the 95% CI of our sample, sample looks good
b) The true population value is not within the 95% CI of our sample, sample looks not so good

D. Testing a sample proportion against a parameter with Stata
Note: In software when you are using a binary nominal variable the % is reported as the mean. For example, if the mean is .35 then the percentage of yes (value = 1) versus no (value = 0) is 35%. It's for this reason we often score such variables as 0/1.

From a recent study we know that approximately 20% of household water in Larimer County comes from wells as opposed by surface/treated sources.

Statistics --> Summaries Tables and Tests --> Summary and Decriptive Statistics --> Confidence Intervals
check: proportions
variables: water *submit*

Q8) What can we say about the proportion of well water use in our sample compared to the county?
a) The true population value is within the 95% CI of our sample, sample looks good
b) The true population value is not within the 95% CI of our sample, sample looks not so good

E. Visualizing multiple confidence intervals with Stata.
Stata does not provide a GUI menu for confidence intervals. Like many such programs some functions require coding, with abundant online support. First, obtain the needed table output:

Statistics --> Summaries Tables and Tests --> Summary and Decriptive Statistics --> Confidence Intervals
main check: means Variables = enviro by/if/in check repear command, enter variable = actions

Then, in the main window of Stata, using the Command pane, enter the code:

```
ciplot enviro, by(actions)
```

return

This calls a plot of the variable enviro (environmentalism score) across the 5 levels of actions taken in the mining issue. Review the codebook for details on these variables.

Q9) What might you suggest **<u>is not true</u>** about this plot?
a) There is a possibility that the mean for enviro is the same across all levels of action.
b) If the true (but unknown) mean for enviro is 21 this is, overall, a representative sample.
c) The enviro score for those taking no actions is different from those who took 4+ actions.
d) The most acurate estimate is that most sample respondent took no actions.

Testing differences between two sample means or two sample proportions

by

Looking at confidence intervals

and

Finding achieved significance

Testing means and proportions

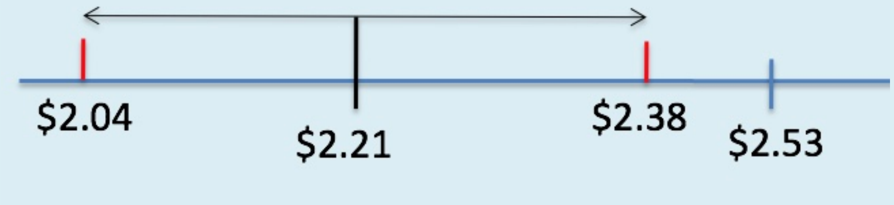
Returning to our use of confidence intervals to test a sample mean/proportion against a parameter . . .

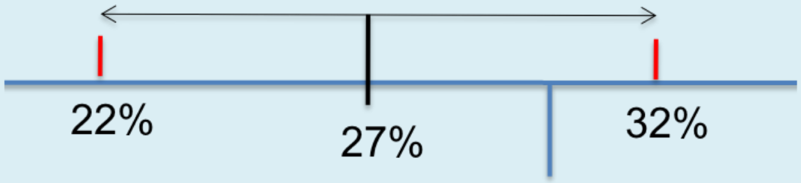
$$SE_x = \frac{SD}{\sqrt{N}}$$

$$SE_p = \sqrt{\frac{p(1-p)}{N}}$$

$$\mathrm{CI} = \bar{\mathrm{X}} \pm Z * SE_x$$

$$\mathrm{CI} = p \pm Z * SE_p$$





Testing means and proportions

Let’s extend this conceptually to look at two sample means, each with it’s own confidence interval.

The variables must be related . . . e.g.:

same measure, population at two points in time

same measure, population that has two groups, etc.

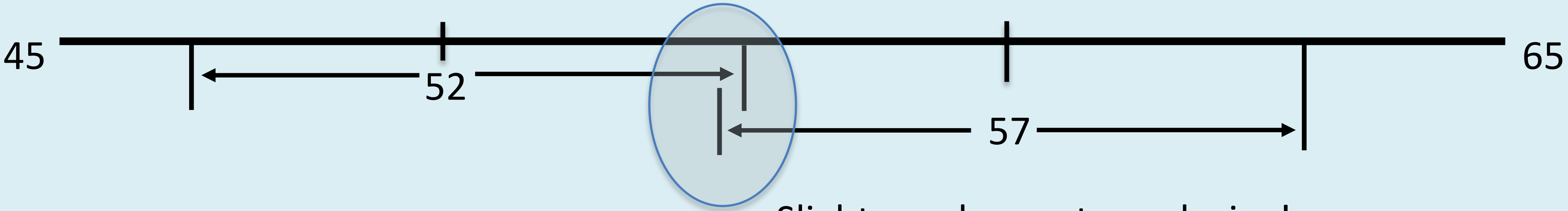
Construct confidence interval around both means and examine for overlap.

Testing means and proportions

Consider the variable age, and the two neighborhood areas that are in our sample:

$$SE_x = \frac{SD}{\sqrt{N}} \qquad \mathrm{CI} = \bar{\mathrm{X}} \pm Z * SE_x$$

Spring Branch	Hedwig Village
M = 52 SD = 16	M = 57 SD = 15
SE = 1.34 N = 149	SE = 1.36 N = 121
CI +/- 1.96 * 1.34 = 2.6	CI +/- 1.96 * 1.36 = 2.7
52 (49.4, 54.6)	57 (54.3, 59.7)



Slight overlap, not conclusive!

Testing means and proportions

Consider the variable ethnicity, and the two neighborhood areas that are in our sample:

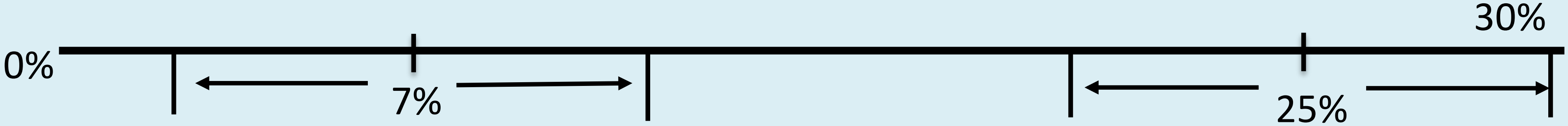
$$SE_p = \sqrt{\frac{p(1-p)}{N}} \qquad \text{CI} = \ p \pm Z * SE_p$$

Hedwig Village N = 121
p = .07 SE = .02

CI +/- 1.96 * .02 = .04
7% (3%, 11%)

Spring Branch N = 149
p = .25 SE = .04

CI +/- 1.96 * .04 = .08
25% (18%, 32%)



No overlap, conclusive!

Testing means and proportions

This approach illustrates the concept of a means or proportion comparison test, but it is not the appropriate way to conduct such a test.

In a close call, how much overlap matters?

We need a more precise way to test these differences.

Testing means and proportions

Consider: in the population there are two groups, each has a true parameter for some variable.

So, there is also a true parameter value for the difference between the two means: $\mu_1 - \mu_2$

We can treat that as a point estimate, find a confidence interval for it, and calculate a precise value for the probability that the difference is zero.

Thus, we test H_a: $(\mu_1 - \mu_2) \neq 0$

Testing means and proportions

This is the t-test.

A point estimate of $\mu_1 - \mu_2$ allows us to imagine a very large number of such estimates, a standard deviation from those estimates, and an estimated standard error. *Just as done with a single mean.*

By testing H_a: $(\mu_1 - \mu_2) \neq 0$ we can arrive at a probability that the difference is far enough from zero that it would not occur by chance, at out given confidence level.

Testing means and proportions

To do so we calculate a test statistic that is associated with a probability function, a curve.

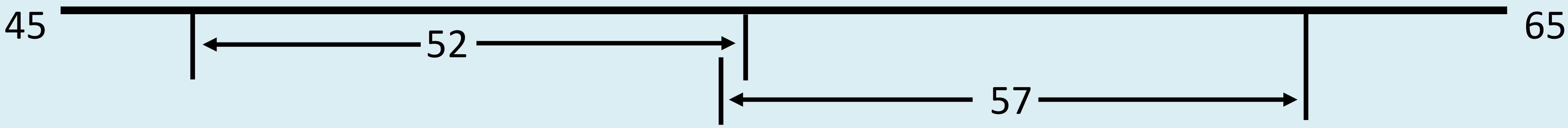
We've worked with Z.
Here we use its close counterpart t.

The value for t depends on its degrees of freedom, which is an adjustment to the sample size.

When the degrees of freedom (*df*) is small, t gets adjusted. When large (N = 50+) t = Z.

Testing means and proportions

Let's return to this …



Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

The difference in means is 57 – 52 = 5 our point estimate

$$\bar{Y}_1 - \bar{Y}_2$$

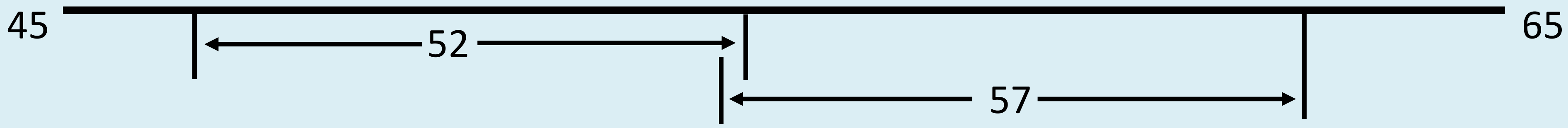
We'll use the same formula as for finding confidence intervals, so we also need to have a standard error to go with this estimate.

$$\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

! Must square the SD

Testing means and proportions

Let’s return to this …



Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

Recall our formula for an area under the curve, it had the form:

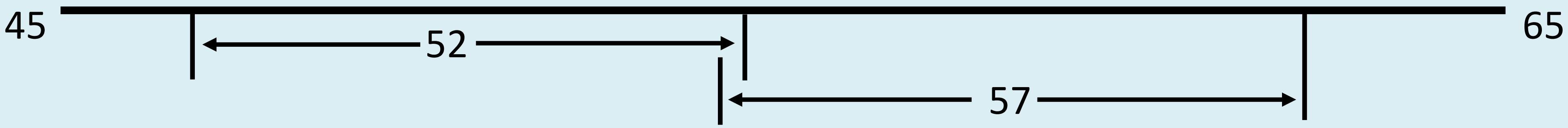
$$Z = \frac{Y_i - \bar{Y}}{s_y}$$

The test statistic Z was found by dividing the point estimate by its standard deviation. Same approach here, different characters:

$$t_{(N_1 + N_2 - 2)} = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{\overline{Y}_1 - \overline{Y}_2}}$$

Testing means and proportions

Let’s return to this …



Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

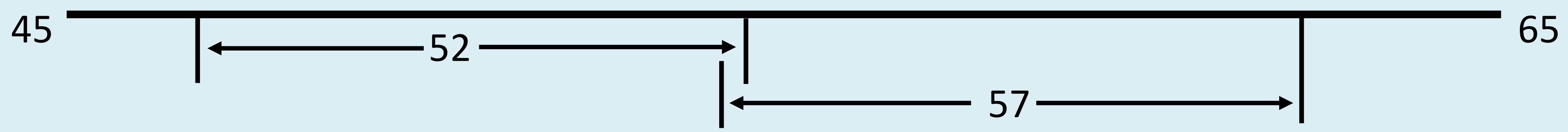
SO: $\bar{Y}_1 - \bar{Y}_2 = 5$

$$\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{16^2}{149} + \frac{15^2}{121}} = \sqrt{\frac{256}{149} + \frac{225}{121}} = 1.89$$

$$t = \frac{5}{1.89} = 2.65 \qquad df = 149 + 121 - 2 = 268$$

Testing means and proportions

Let’s return to this …

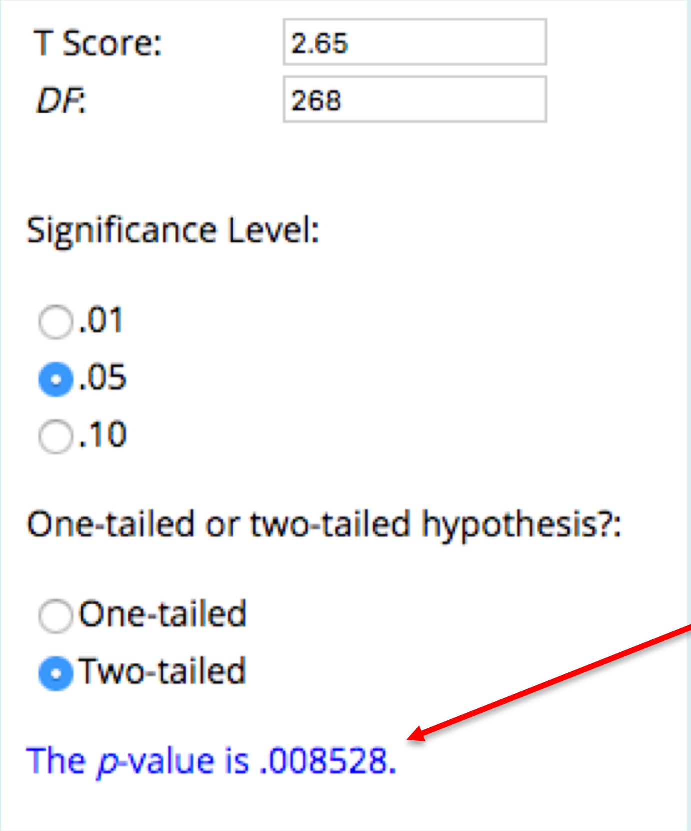


Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

$$t_{(\mathrm{df} = 268)} = 2.65$$

Treat this just like a Z value, and get the probability that you would see a t-value of 2.65



T Score: 2.65

DF: 268

Significance Level:

○ .01
● .05
○ .10

One-tailed or two-tailed hypothesis?:

○ One-tailed
● Two-tailed

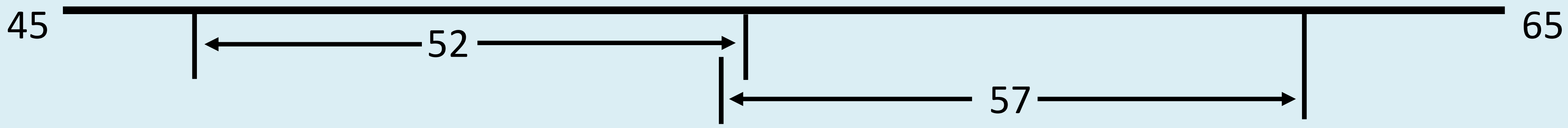
The *p*-value is .008528.

Achieved p-value

www.socscistatistics.com/pvalues/tdistribution.aspx

Testing means and proportions

Let's return to this …

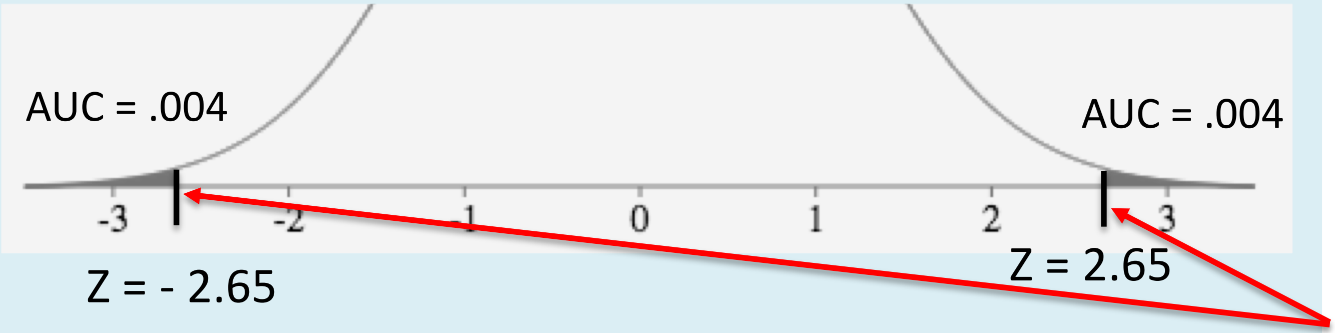


Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

The probability of seeing a difference of 5 years, given the data, is very small (8/1000). That's the chance that the null hypothesis is true, so we reject it and accept the alternative hypothesis.

The cut-off value at the 95% CL is p ≤ .05

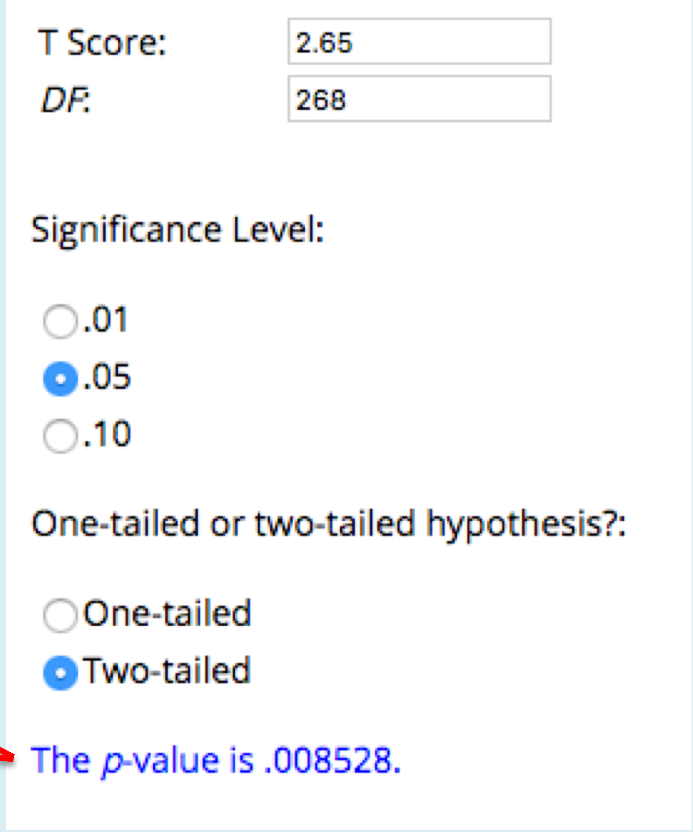


AUC = .004

AUC = .004

Z = - 2.65

Z = 2.65



T Score: 2.65

DF: 268

Significance Level:

○.01
●.05
○.10

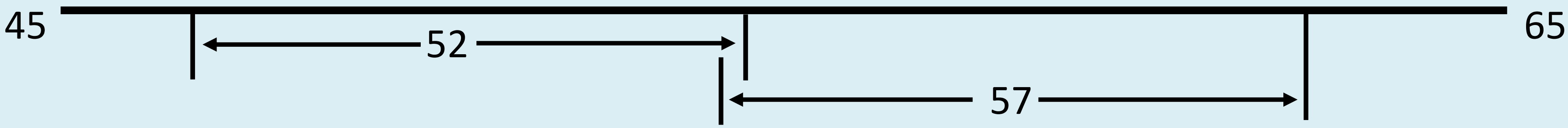
One-tailed or two-tailed hypothesis?:

○One-tailed
●Two-tailed

The *p*-value is .008528.

Testing means and proportions

Let’s return to this …



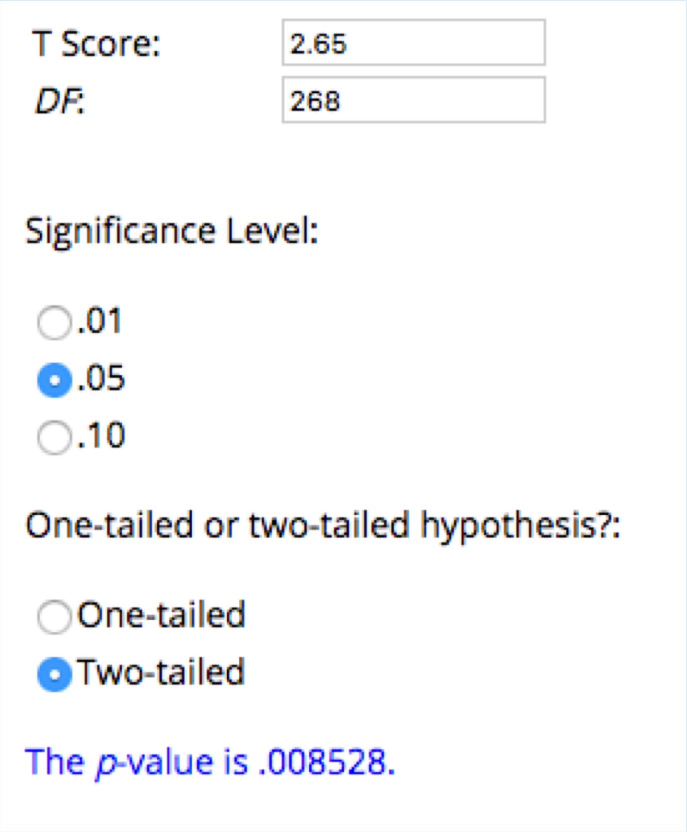
Spring Branch Age ~N(52, 16) N = 149

Hedwig Village Age ~N(57, 15) N = 121

The cut-off value at the 95% CL is p ≤ .05

Since the achieved p-value is .008 (much smaller than .05) we state that the difference is significant.

Notice how this precision differs from the ambiguity of the confidence interval.



T Score: 2.65

DF: 268

Significance Level:

○ .01
● .05
○ .10

One-tailed or two-tailed hypothesis?:

○ One-tailed
● Two-tailed

The *p*-value is .008528.

Testing means and proportions

That was fun . . . Let’s do the same with proportions!

$$\mathrm{SE}_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$$

Hedwig Village N = 121	Spring Branch N = 149
p = .07 SE = .02	p = .25 SE = .04

$$.04 \quad = \sqrt{\frac{.07\,(1-.07)}{121} + \frac{.25\,(1-.25)}{149}}$$

Testing means and proportions

That was fun . . . Let's do the same with proportions!

Hedwig Village N = 121	Spring Branch N = 149
p = .07 SE = .02	p = .25 SE = .04

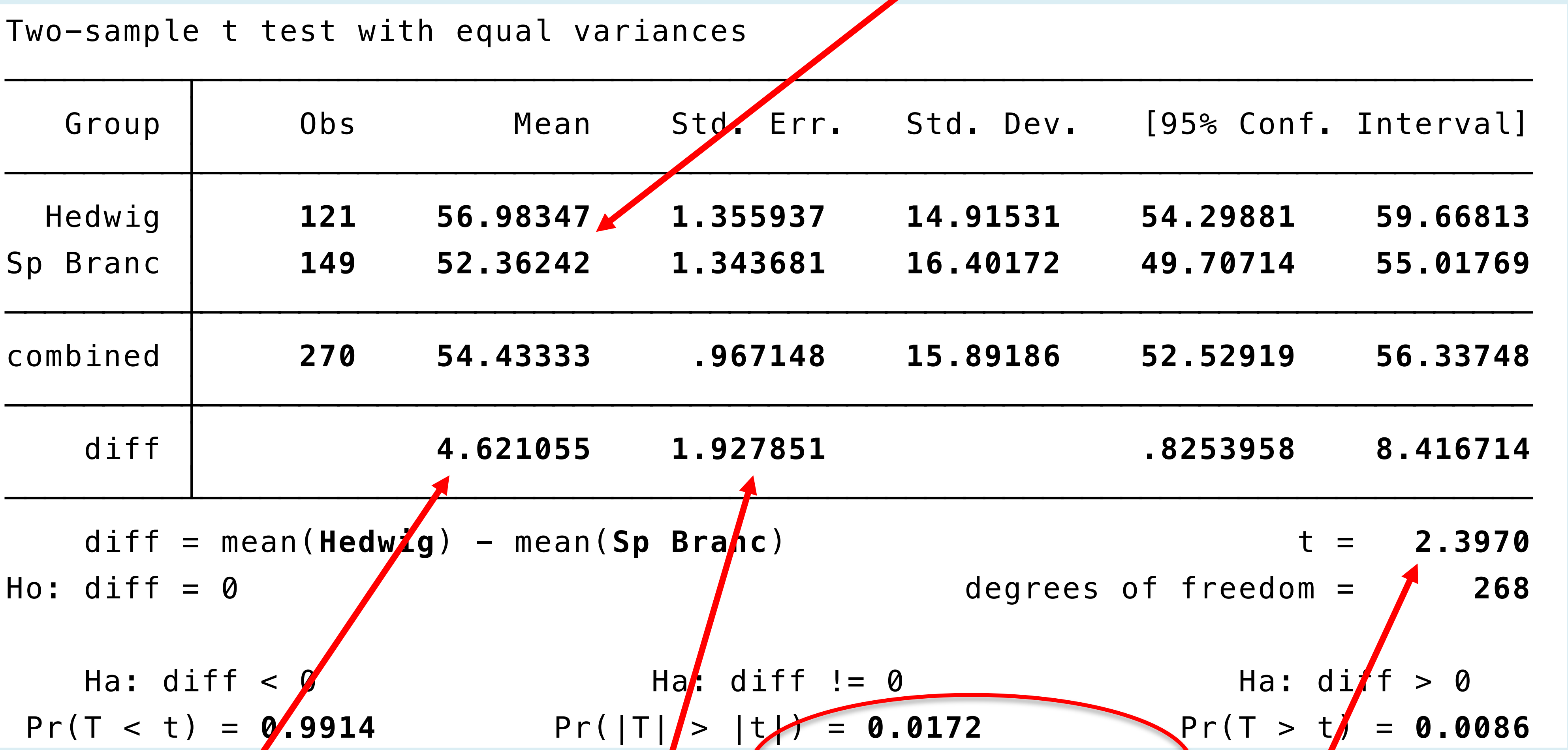
$$Z = \frac{p_1 - p_2}{SE_p} = \frac{.25 - .07}{.04} = 4.5$$

This is a *very large* Z-statistic, given we know that about 99% of the distribution falls +/- 3. We conclude that p < .0001

Testing means and proportions

With Stata . . .

Means for each group



```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Hedwig |     121    56.98347    1.355937    14.91531    54.29881    59.66813
Sp Branc |     149    52.36242    1.343681    16.40172    49.70714    55.01769
---------+--------------------------------------------------------------------
combined |     270    54.43333     .967148    15.89186    52.52919    56.33748
---------+--------------------------------------------------------------------
    diff |            4.621055    1.927851                .8253958    8.416714
------------------------------------------------------------------------------
    diff = mean(Hedwig) - mean(Sp Branc)                          t =   2.3970
Ho: diff = 0                                     degrees of freedom =      268

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9914         Pr(|T| > |t|) = 0.0172          Pr(T > t) = 0.0086
```

Two-tailed p-value

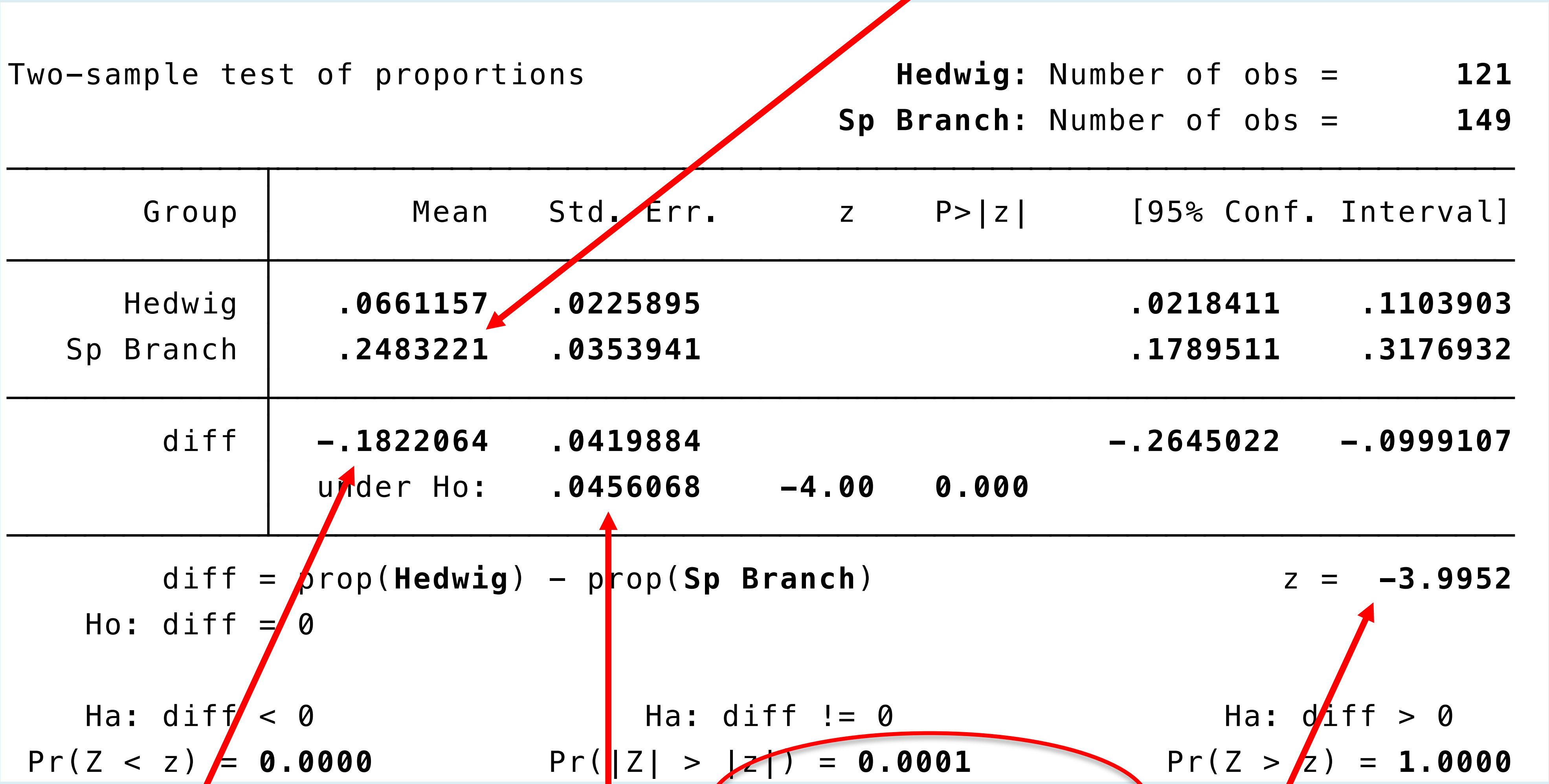
t-statistic (diff ÷ SE)

Difference in means, and its SE

Testing means and proportions

With Stata . . .

Proportions for each group



```
Two-sample test of proportions                   Hedwig: Number of obs =       121
                                              Sp Branch: Number of obs =       149
-------------------------------------------------------------------------------
       Group |       Mean    Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
      Hedwig |   .0661157    .0225895                       .0218411    .1103903
   Sp Branch |   .2483221    .0353941                       .1789511    .3176932
-------------+-----------------------------------------------------------------
        diff |  -.1822064    .0419884                      -.2645022   -.0999107
             |  under Ho:    .0456068    -4.00   0.000
-------------------------------------------------------------------------------
        diff = prop(Hedwig) - prop(Sp Branch)                     z =  -3.9952
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0000         Pr(|Z| > |z|) = 0.0001          Pr(Z > z) = 1.0000
```

Two-tailed p-value

Z-statistic (diff ÷ SE)

Difference in proportions, and its SE

Testing means and proportions

Keywords

Standard (Z) Score
Parameter
Population
Sampling Distribution
Sampling Distribution of the Mean
Sampling Error
Standard Error of the Mean
Two-tailed test (non-directional)

Alternative hypothesis
Null hypothesis
SE for difference btwn means
Statistical hypothesis testing
Alpha
Confidence level
P-value
Test statistic

JTC270 Lab 6 Testing Hypotheses about Means and Proportions <JTC270 Lab Dataset.dta>

$$t_{(\text{df N1 + N2 - 2})} = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}} \quad \textit{where} \quad \text{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

$$Z = \frac{p_1 - p_2}{SE_p} \quad \textit{where} \quad \text{SE}_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{N_1} + \frac{p_2(1 - p_2)}{N_2}}$$

A. Testing two sample means

Here we'll look at the uranium mine data. One of the first thigs analysts do when looking into a dataset is examine demographics for any interesting differences. We'd like to know if the mean age is significantly different in our two areas, Wellington and the surrounding Rural community. Here are the descriptive stats:

Wellington Age ~N(54, 15) N = 117 Rural Age ~N(46, 13) N = 76

Q1) Given this scenario, what would we want to test for a non-directional null hypothesis?

a) $\mu_1 \neq \mu_2$ b) $\mu_1 = \mu_2$ c) $\mu_1 > \mu_2$ d) $\mu_1 < \mu_2$

Q2) What is the value for the SE? _______.

Q3) What is the value for t? _______.

https://www.socscistatistics.com/pvalues/tdistribution.aspx

Q4) What is the achieved p-value?

a) p = .05
b) p > .01
c) p < .001

B. Testing two sample proportions

Continuing with out demographic analysis, we'd like to know if the proportion of men-to-women is significantly different in our two areas, Wellington and the surrounding Rural community.
Here are the descriptive stats: Wellington 74% male N = 76 Rural 63% make N = 117

Q5) Given this scenario, what would we want to test for a non-directional null hypothesis?

a) $\pi_1 \neq \pi_2$ b) $\pi_1 = \pi_2$ c) $\pi_1 > \pi_2$ d) $\pi_1 < \pi_2$

Q6) What is the value for the SE? _______.

Q7) What is the value for Z? _______.

Q8) What is the achieved p-value? _______.
https://www.socscistatistics.com/pvalues/normaldistribution.aspx

Open our semester dataset in Stata.

C. Testing two sample means with Stata

Here we'll continue looking at differences between Wellington and the Rural area. We'll look at our variable that measure risk perception, with high values indicating high risk for the mine. **Ha: $\mu_1 \neq \mu_2$**

Statistics --> Summaries Tables and Tests --> Classical Tests --> T-Test (mean comparison)
check: Two-sample using groups
variable: riskper   Group: town   *submit*

Q9) What is the difference in risk perception scores between the areas? ________.

Q10) What is the result of the test?
a) reject the null, accept the alternative
b) accept the null, reject the alternative

D. Testing two sample proportions with Stata

Note: In software when you are using a binary nominal variable the % is reported as the mean. For example, if the mean is .35 then the percentage of yes (value = 1) versus no (value = 0) is 35%. It's for this reason we often score such variables as 0/1.

Continuing with our comparison, we'd like to know if the proportion of well-to-treated water is the same.
Note that water is scored 0 = well 1 = treated   **Ha: $\pi_1 \neq \pi_2$**

Statistics --> Summaries Tables and Tests --> Classical Tests --> Proportion test
check: Two-sample using groups
variable: water   Group: town   *submit*

Q11) What is the difference in the proportion of household with treated water between the areas? ________.

Q12) What is the result of the test?
a) reject the null, accept the alternative
b) accept the null, reject the alternative

Evaluating Independence:

Cross-tabulation and the chi-square test.

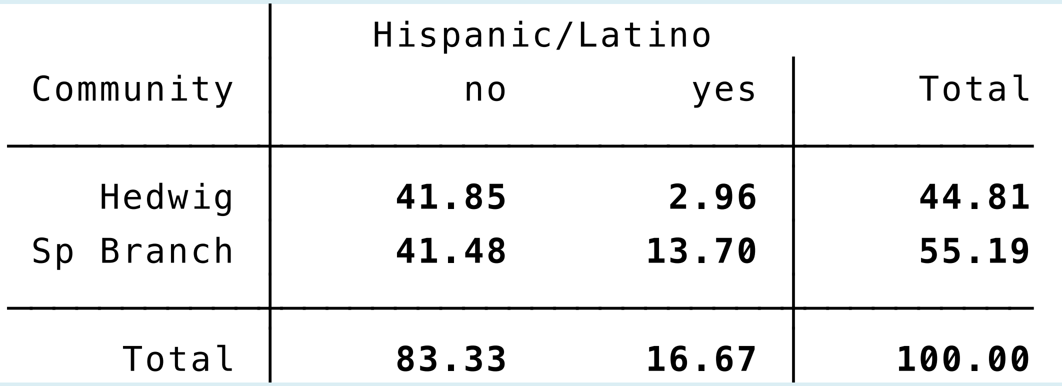
Expanding the number of levels

because, it's a complicated world

Z-test for two proportions $$Z = \frac{p_1 - p_2}{SE_p}$$

This is a useful test, but is limited to dichotomous nominal-level variables that are expressed as a ratio (a proportion) of one condition (level) versus another.

e.g., the West-Nile ethnicity variable can be viewed as such:



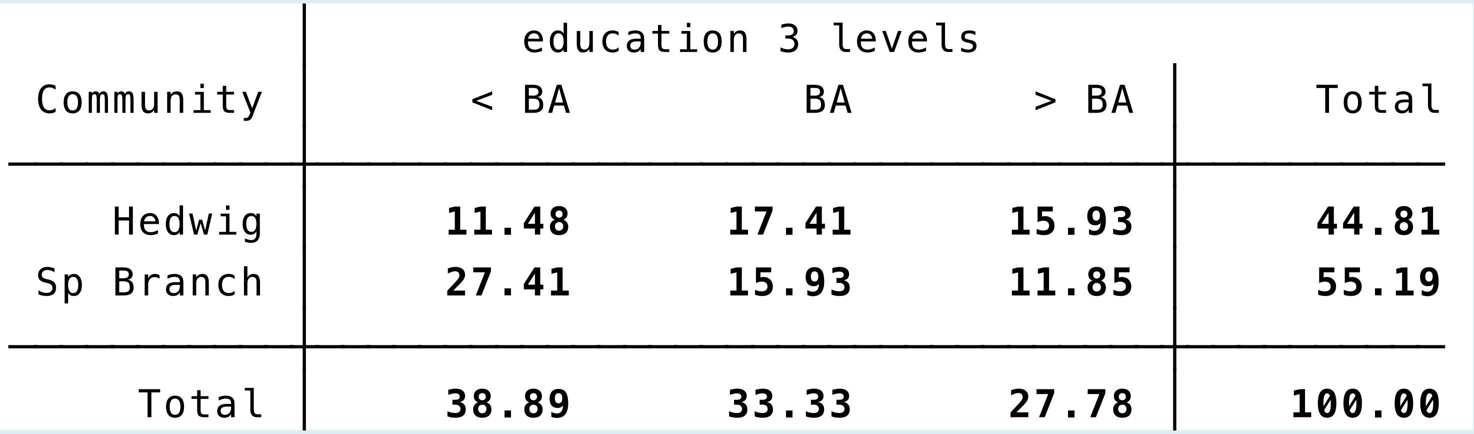
```
           |   Hispanic/Latino
 Community |        no        yes |     Total
-----------+----------------------+----------
    Hedwig |     41.85       2.96 |     44.81
 Sp Branch |     41.48      13.70 |     55.19
-----------+----------------------+----------
     Total |     83.33      16.67 |    100.00
```

Expanding the number of levels

A level is one of the categories of a categorical or ordinal variable. For the variable “area” one level is Spring Branch and the other level is Hedwig Village.

But what if there are more than two levels?

Consider, educational attainment as a three-level variable that can be examined across the two levels of area:



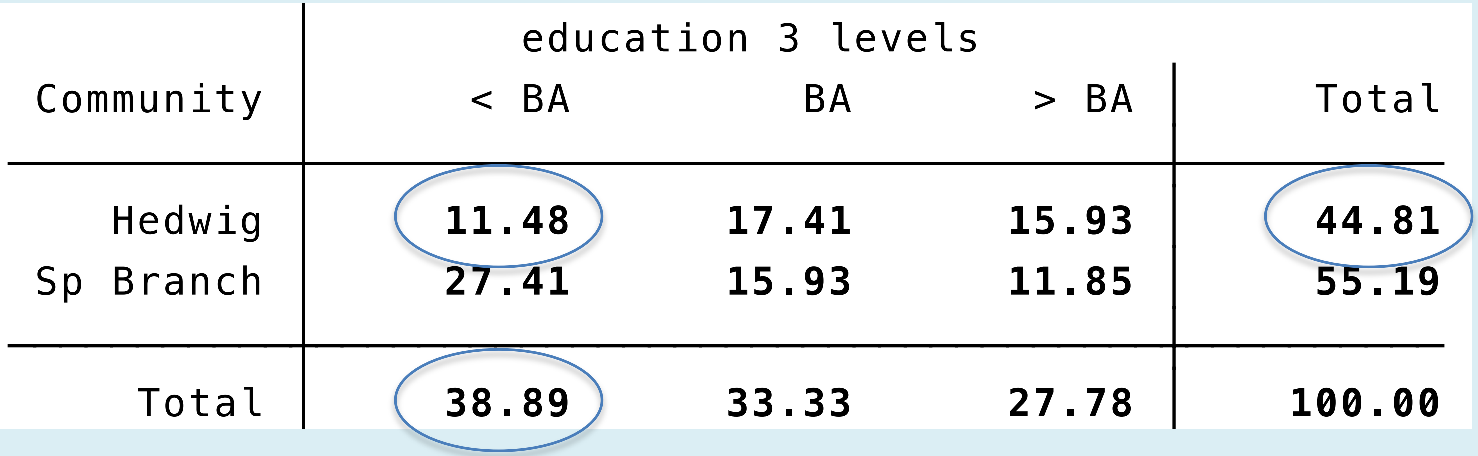
```
           |      education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |     11.48      17.41      15.93 |     44.81
 Sp Branch |     27.41      15.93      11.85 |     55.19
-----------+---------------------------------+----------
     Total |     38.89      33.33      27.78 |    100.00
```

Expanding the number of levels

Reading the table:

Each of the 6 cells in the table is the percentage of the total number of cases (i.e., including both areas) that fall into that condition. Thus, we can make a statement such as:

Of the total number of participants 11.5% have less than a BA *and* live in Hedwig Village, of all participants 38.9% have less than a BA, and of all participants 44.8% reside in Hedwig.

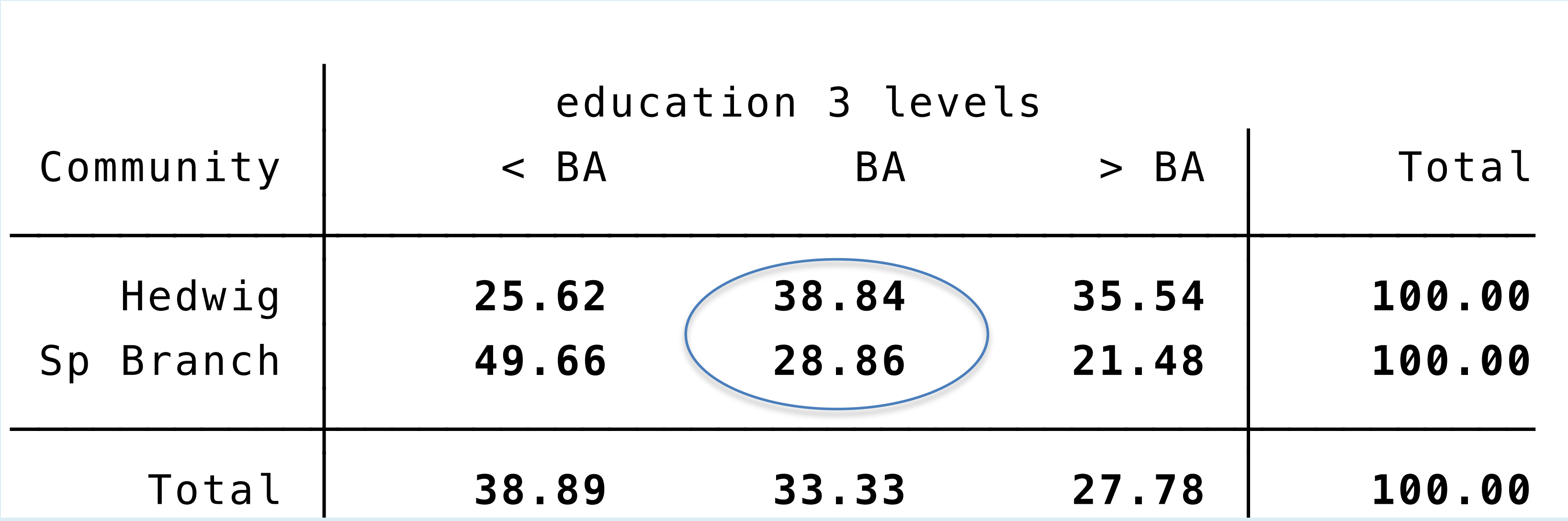


```
           |      education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |     11.48      17.41      15.93 |     44.81
 Sp Branch |     27.41      15.93      11.85 |     55.19
-----------+---------------------------------+----------
     Total |     38.89      33.33      27.78 |    100.00
```

Expanding the number of levels

Enriching the table by rows:

While a “raw” table can be useful, it’s much more useful to percentage the table across levels, as such:



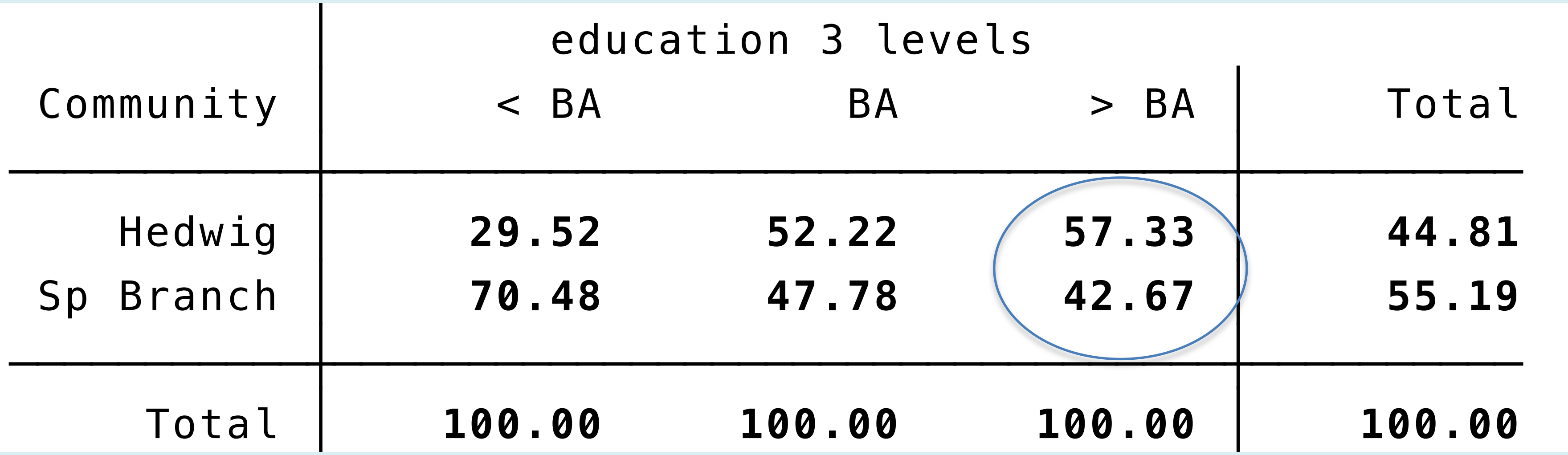
```
           |       education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |     25.62      38.84      35.54 |    100.00
 Sp Branch |     49.66      28.86      21.48 |    100.00
-----------+---------------------------------+----------
     Total |     38.89      33.33      27.78 |    100.00
```

Now we can make comparisons: In Hedwig Village 38.8% of participants have a BA while in Spring Branch 28.9% do.

Expanding the number of levels

Enriching the table by columns:

We can also switch the table to look at the column variable:

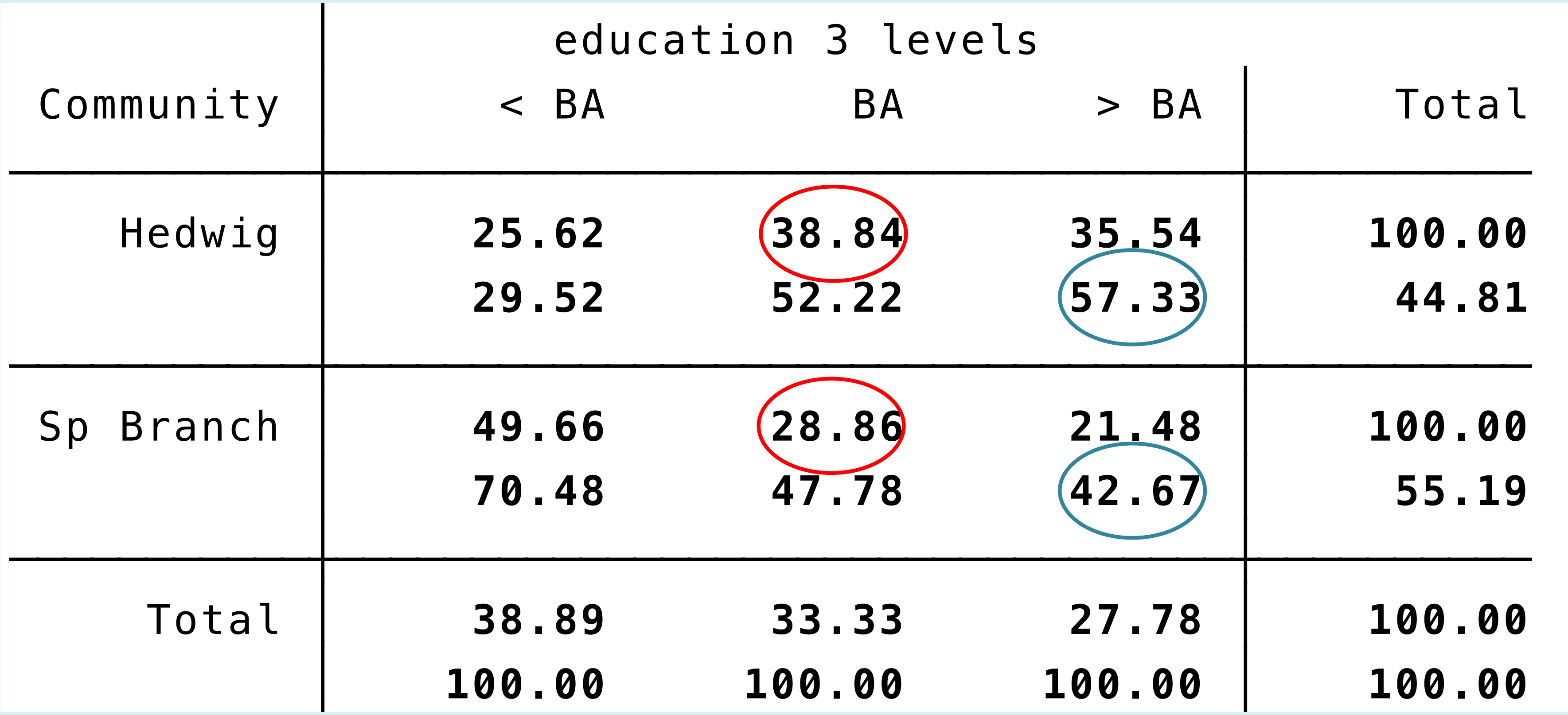


```
           |       education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |     29.52      52.22      57.33 |     44.81
 Sp Branch |     70.48      47.78      42.67 |     55.19
-----------+---------------------------------+----------
     Total |    100.00     100.00     100.00 |    100.00
```

Now we can make other comparisons: Of the participant who have a degree greater than a BA 57.3% are in Hedwig Village versus 42.7% who are in Spring Creek.

Expanding the number of levels

Enriching the table by rows and columns:

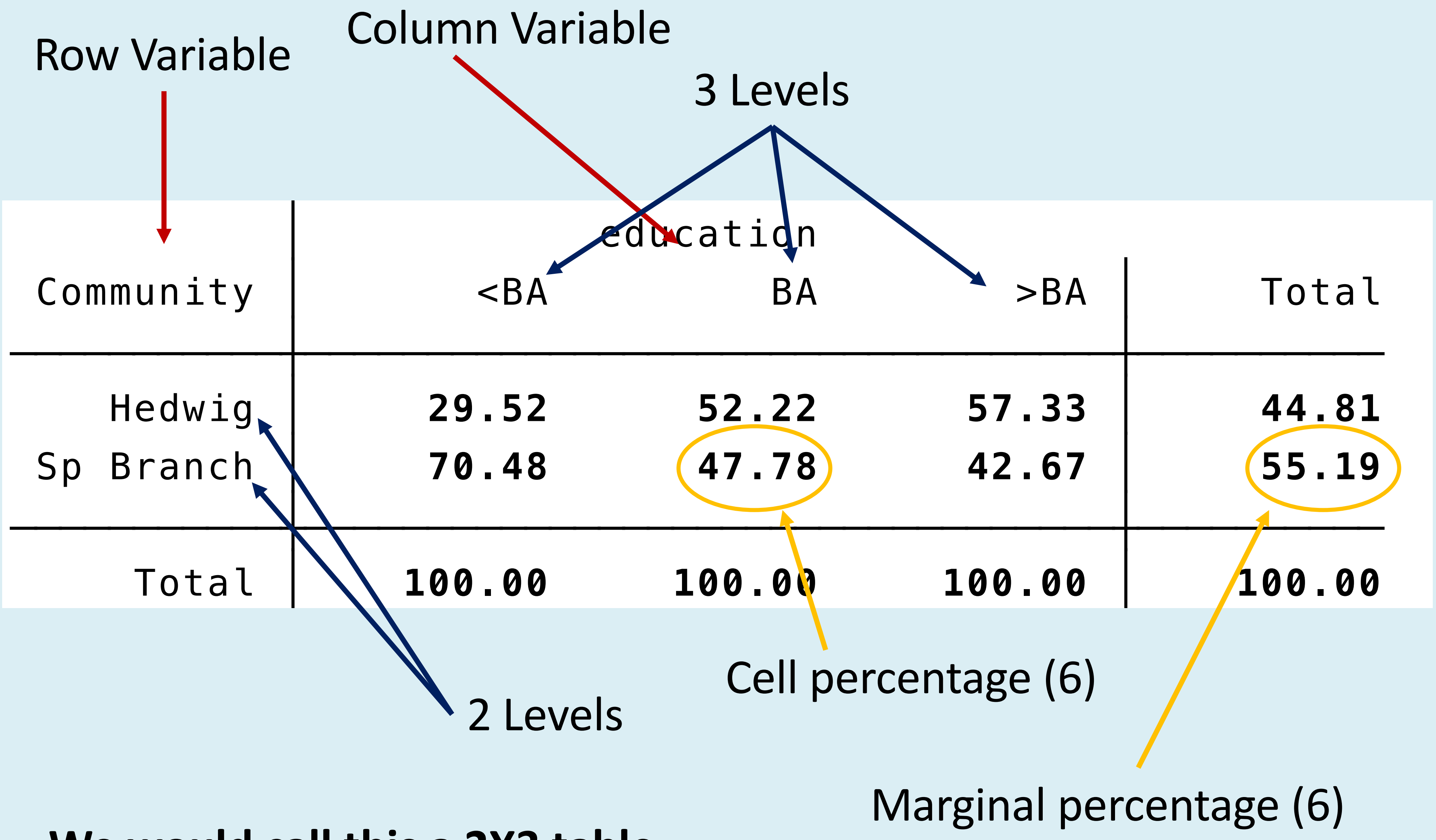


	education 3 levels			
Community	< BA	BA	> BA	Total
Hedwig	25.62 29.52	38.84 52.22	35.54 57.33	100.00 44.81
Sp Branch	49.66 70.48	28.86 47.78	21.48 42.67	100.00 55.19
Total	38.89 100.00	33.33 100.00	27.78 100.00	100.00 100.00

We can find our same conclusions in the fuller table.

Usually, we will choose either rows or columns, not both.

Table Components



Row Variable

Column Variable

3 Levels

```
           |          education
 Community |       <BA         BA        >BA |     Total
-----------+---------------------------------+----------
    Hedwig |     29.52      52.22      57.33 |     44.81
 Sp Branch |     70.48      47.78      42.67 |     55.19
-----------+---------------------------------+----------
     Total |    100.00     100.00     100.00 |    100.00
```

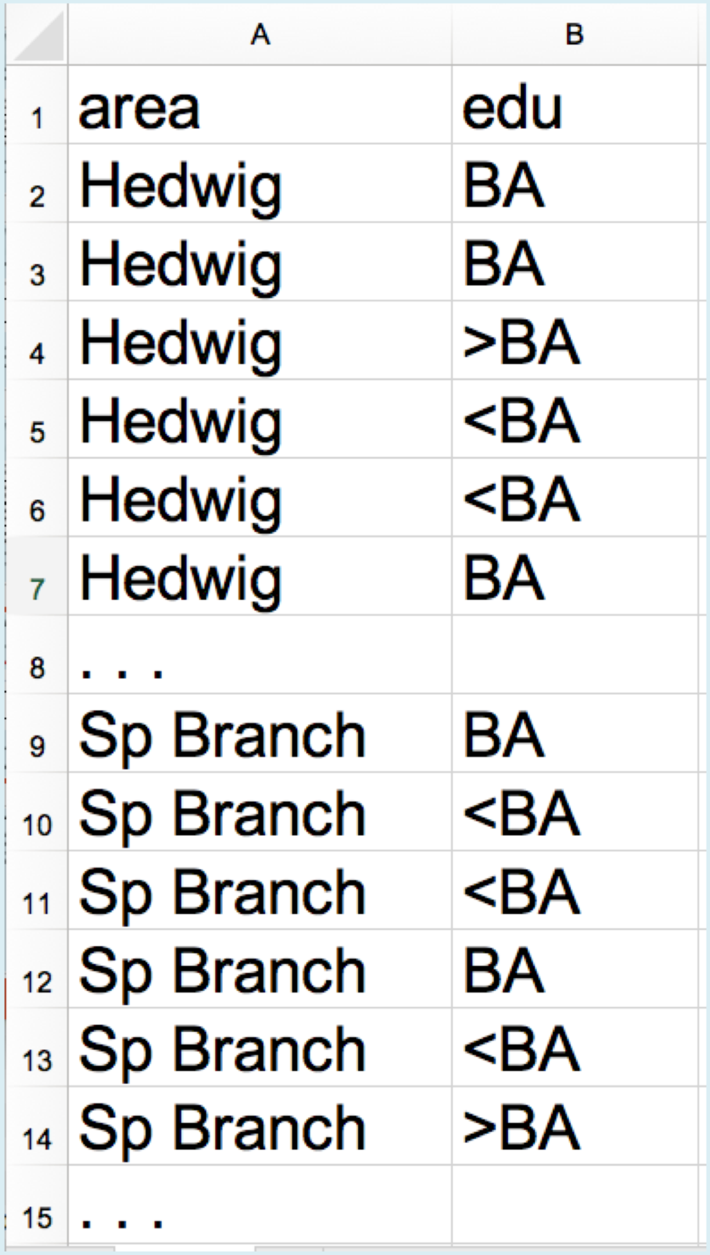
2 Levels

Cell percentage (6)

Marginal percentage (6)

We would call this a 2X3 table

Table Components



	A	B
1	area	edu
2	Hedwig	BA
3	Hedwig	BA
4	Hedwig	>BA
5	Hedwig	<BA
6	Hedwig	<BA
7	Hedwig	BA
8	. . .	
9	Sp Branch	BA
10	Sp Branch	<BA
11	Sp Branch	<BA
12	Sp Branch	BA
13	Sp Branch	<BA
14	Sp Branch	>BA
15	. . .	

It's often helpful to think about how the data are arranged in the spreadsheet.

For data entry:

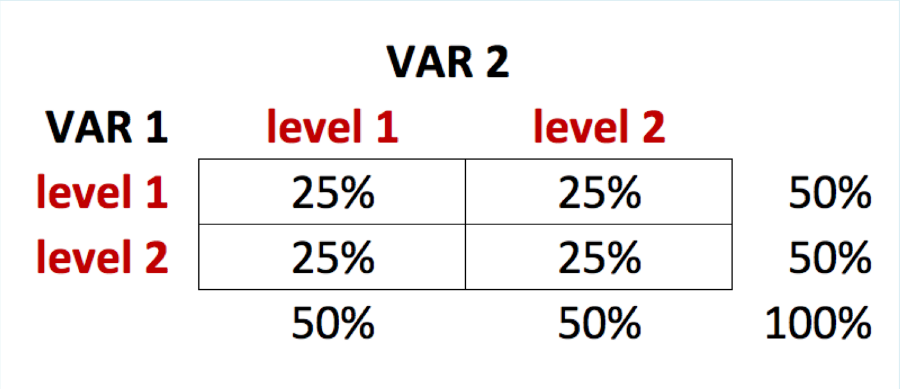
Hedwig = 1
Sp Branch = 2

<BA = 1
BA = 2
>BA = 3

Are the variables related?

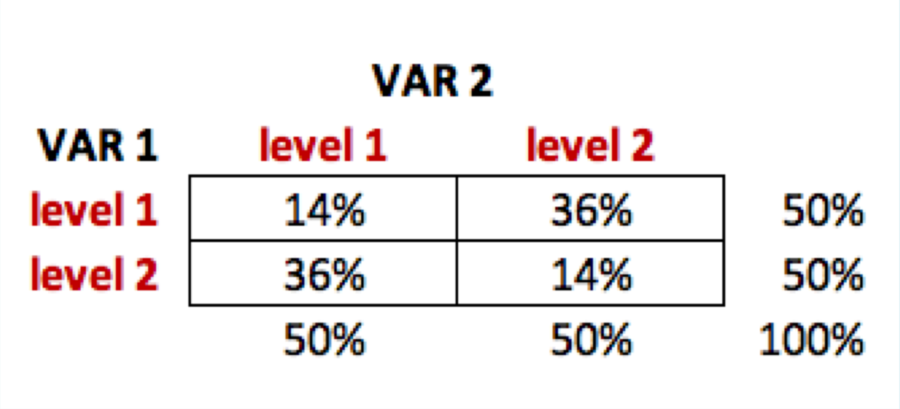
Specifically, are the percentages presented in the full table sufficiently different from each other that we can conclude that such differences can't be attributed to random chance?

A random 2X2 table would be:



	VAR 2		
VAR 1	level 1	level 2	
level 1	25%	25%	50%
level 2	25%	25%	50%
	50%	50%	100%

A non-random 2X2 would be:



	VAR 2		
VAR 1	level 1	level 2	
level 1	14%	36%	50%
level 2	36%	14%	50%
	50%	50%	100%

Are the variables related?

Because our data comes from an appropriate random sample we can use the same statistical techniques for probability and inference.

We declare the variable roles typically based on some theory or logic:

Independent variable = Area
Dependent variable = Education

We might expect the area to have an effect on education levels, but perhaps not vice-versa.

Hypothesis Testing

We are testing statistical independence.

This is the absence of an association.

Do we have evidence to conclude that *in the population* the percentages of the dependent variable within each category of the independent variable are statistically identical?

Thus: H_o: VAR1 and VAR2 are independent.

H_a: VAR1 and VAR2 are dependent.

The Chi-Square Test

Assumes that data come from a random sample.

Requires that data be at the nominal or ordinal level.

The test statistic: $$x^2 = \sum_{cells}^{all} \frac{(f_e - f_o)^2}{f_e}$$

where: $$f_e = \frac{(Col.\ marginal)(Row\ marginal)}{N}$$

N is the total number of observations

The Chi-Square Test

Expected frequencies (f_e)

The cell frequencies (counts) that would be expected in a bivariate table if the two variables were statistically independent.

Observed frequencies (f_o)

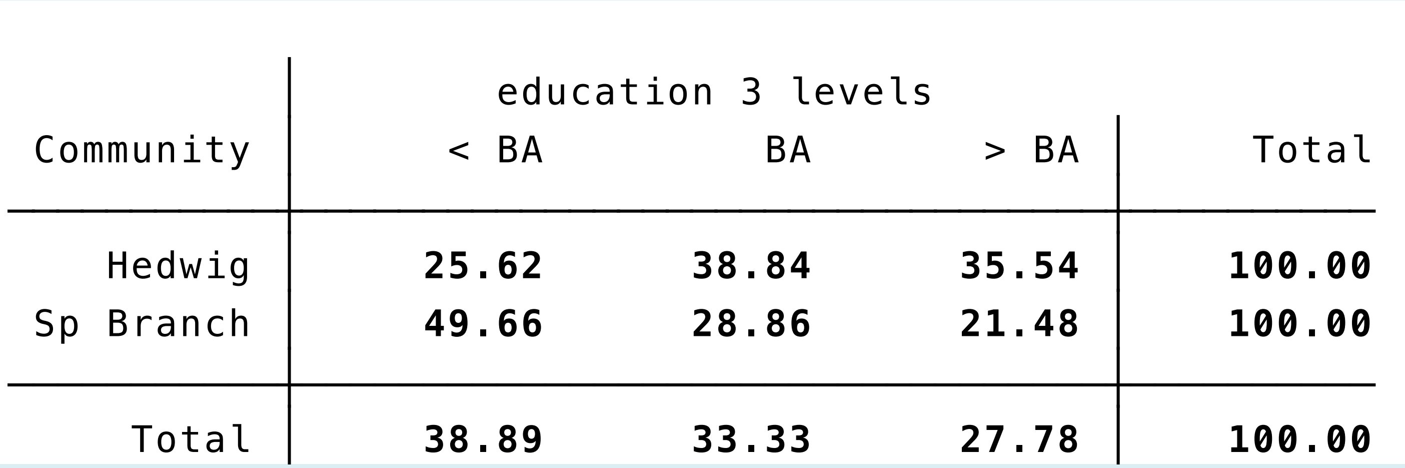
The cell frequencies (counts) that are actually observed in a bivariate table.

The test asks if the two are sufficiently different as to reject the null hypothesis of independence.

The Chi-Square Test

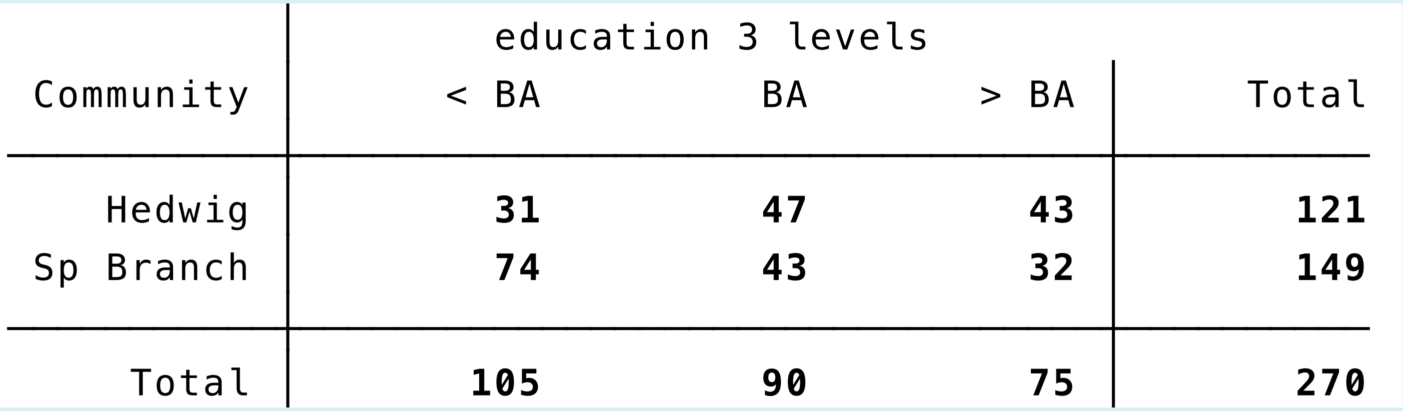
Expected frequencies (f_e)

Row proportions →



```
           |       education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |     25.62      38.84      35.54 |    100.00
 Sp Branch |     49.66      28.86      21.48 |    100.00
-----------+---------------------------------+----------
     Total |     38.89      33.33      27.78 |    100.00
```

Frequencies →

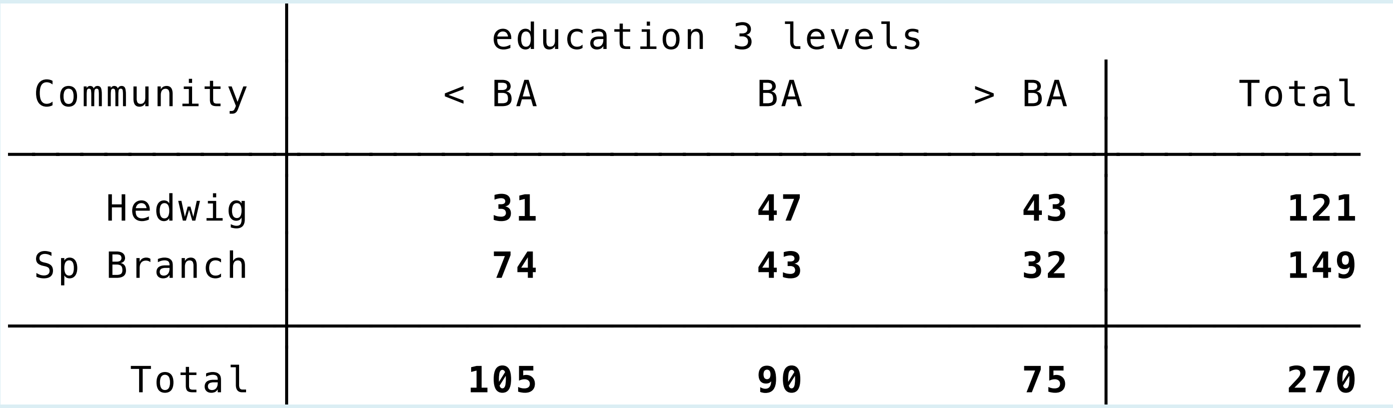


```
           |       education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |        31         47         43 |       121
 Sp Branch |        74         43         32 |       149
-----------+---------------------------------+----------
     Total |       105         90         75 |       270
```

The Chi-Square Test

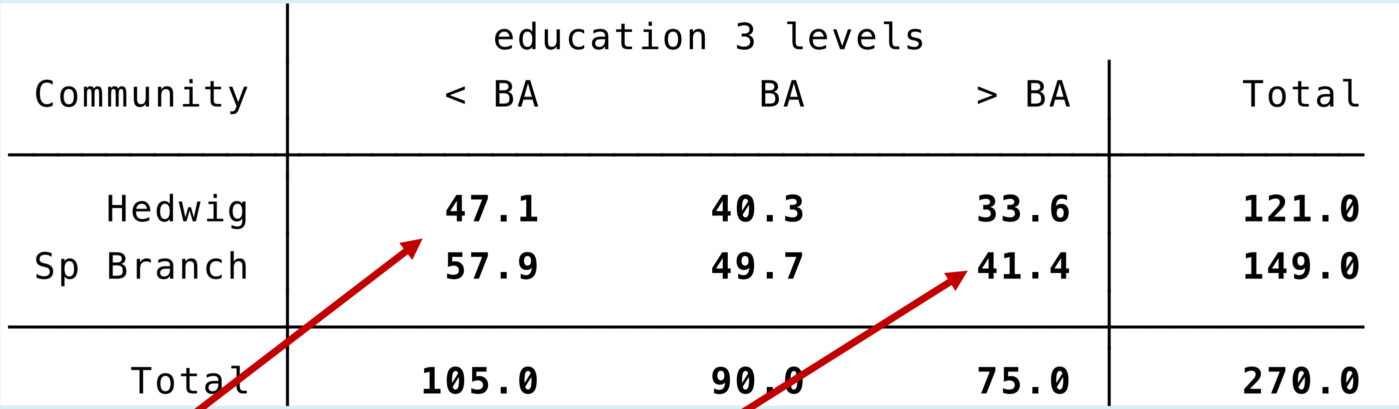
Expected frequencies (f_e)

Frequencies →



```
           |      education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |        31         47         43 |       121
 Sp Branch |        74         43         32 |       149
-----------+---------------------------------+----------
     Total |       105         90         75 |       270
```

Expected Frequencies →



```
           |      education 3 levels
 Community |      < BA         BA       > BA |     Total
-----------+---------------------------------+----------
    Hedwig |      47.1       40.3       33.6 |     121.0
 Sp Branch |      57.9       49.7       41.4 |     149.0
-----------+---------------------------------+----------
     Total |     105.0       90.0       75.0 |     270.0
```

121 * 105 ÷ 270 = 47.1

149 * 75 ÷ 270 = 41.4

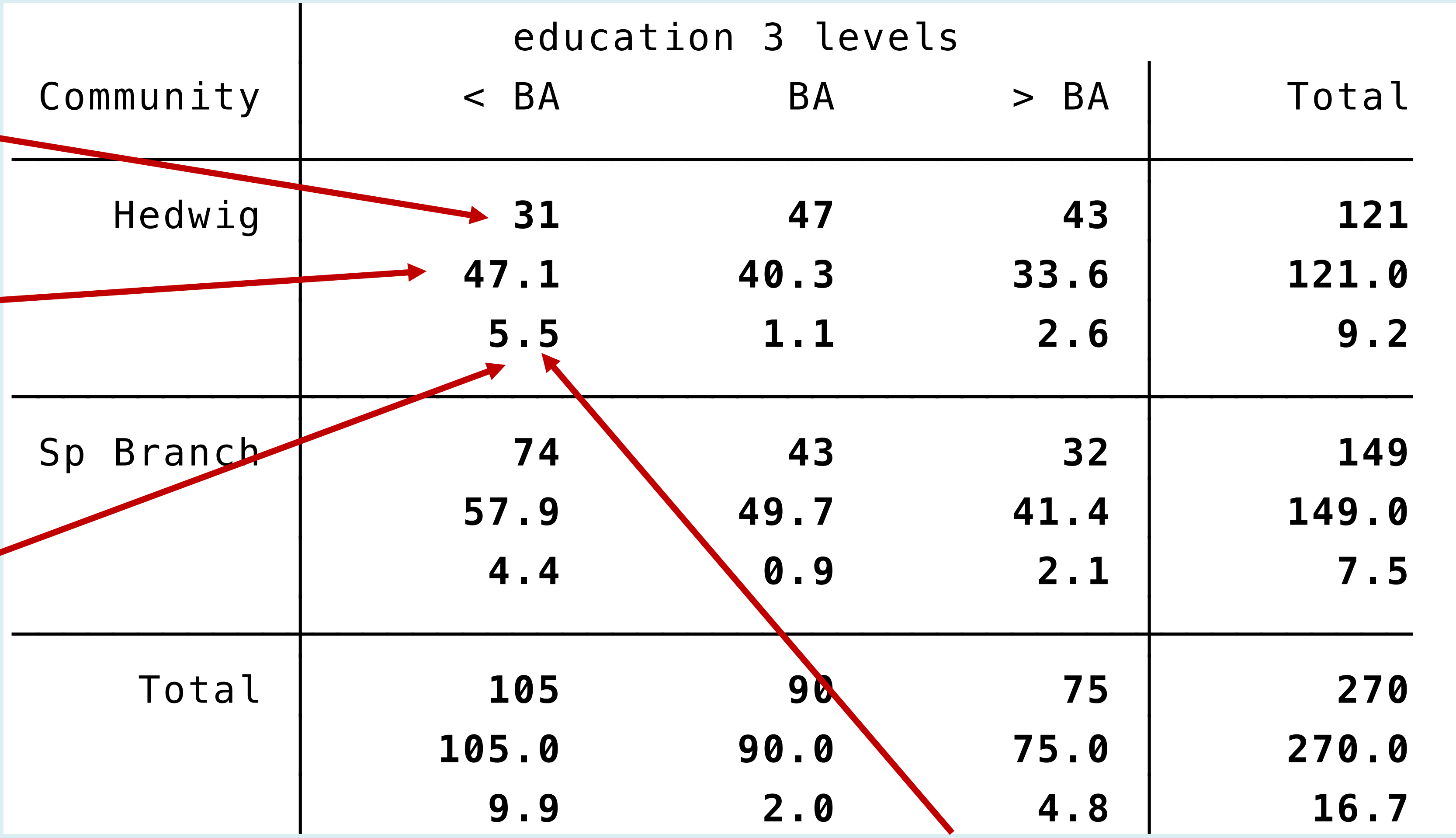
Repeat for each cell

The Chi-Square Test

f_o

f_e

$$\frac{(f_e - f_o)^2}{f_e}$$



```
                    education 3 levels
  Community    < BA        BA      > BA      Total
--------------------------------------------------
     Hedwig      31        47        43        121
               47.1      40.3      33.6      121.0
                5.5       1.1       2.6        9.2
--------------------------------------------------
  Sp Branch      74        43        32        149
               57.9      49.7      41.4      149.0
                4.4       0.9       2.1        7.5
--------------------------------------------------
      Total     105        90        75        270
              105.0      90.0      75.0      270.0
                9.9       2.0       4.8       16.7
```

$$(47.1 - 31)^2 \div 47.1 = 5.5$$

$$x^2 = \sum_{cells}^{all} \frac{(f_e - f_o)^2}{f_e}$$

$$\chi^2 = 5.5 + 1.1 + 2.6 + 4.4 + 0.9 + 2.1 = 16.7$$

The Chi-Square Test

$$x^2 = \sum_{cells}^{all} \frac{(f_e - f_o)^2}{f_e} \quad = 5.5 + 1.1 + 2.6 + 4.4 + 0.9 + 2.1 = \mathbf{16.7}$$

This is a test statistic, just as we had with *Z* and *t*

It has a probability distribution, so we can look up how likely it is that we found that test statistic, and see if it might happen by chance more than, say, 5 times in 100 (95% CL, α = .05).

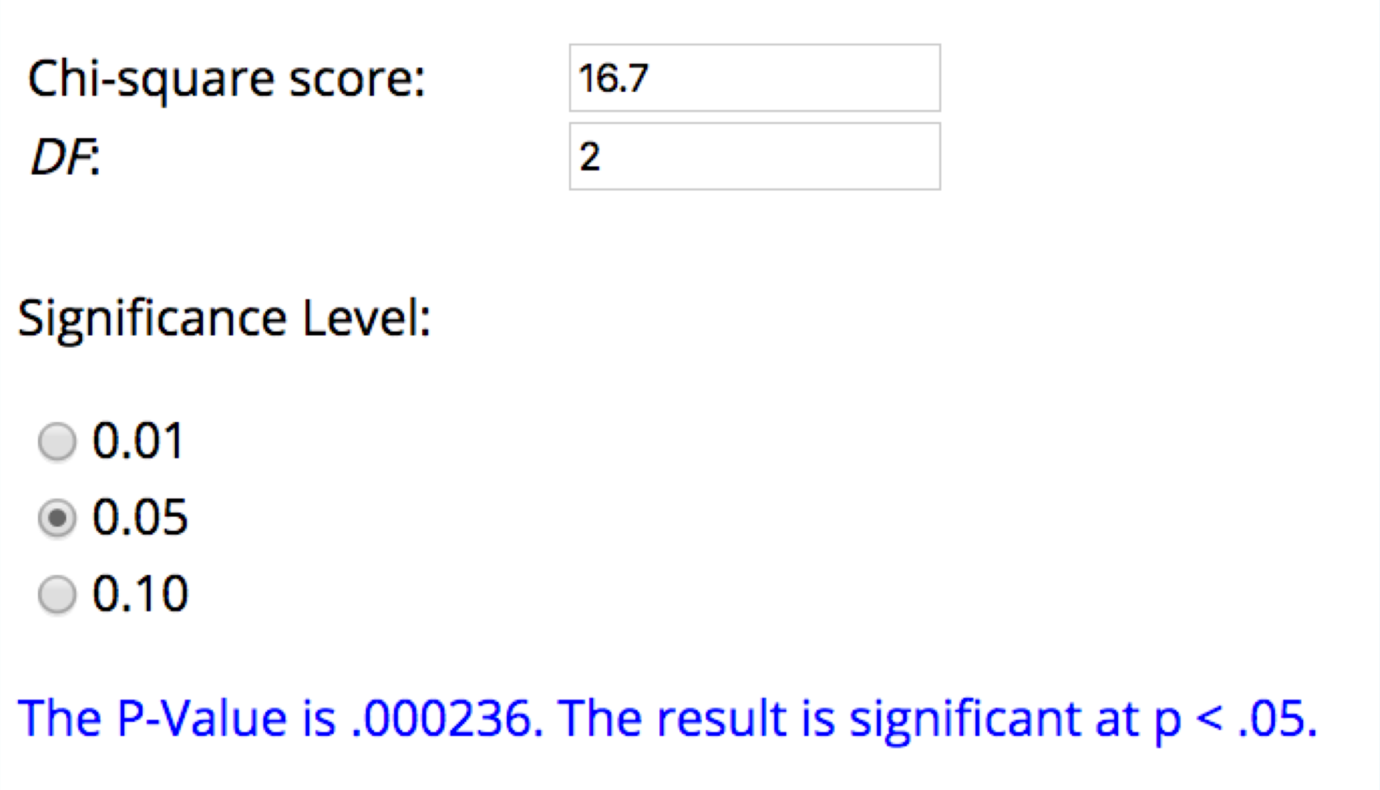
It also requires the use of degrees of freedom:

df **= (# rows – 1)(# columns – 1)**

The Chi-Square Test

$$x^2 = \sum_{cells}^{all} \frac{(f_e - f_o)^2}{f_e} = 5.5 + 1.1 + 2.6 + 4.4 + 0.9 + 2.1 = \mathbf{16.7}$$

df = (2-1)(3-1) = 2



Chi-square score:	16.7
DF:	2

Significance Level:

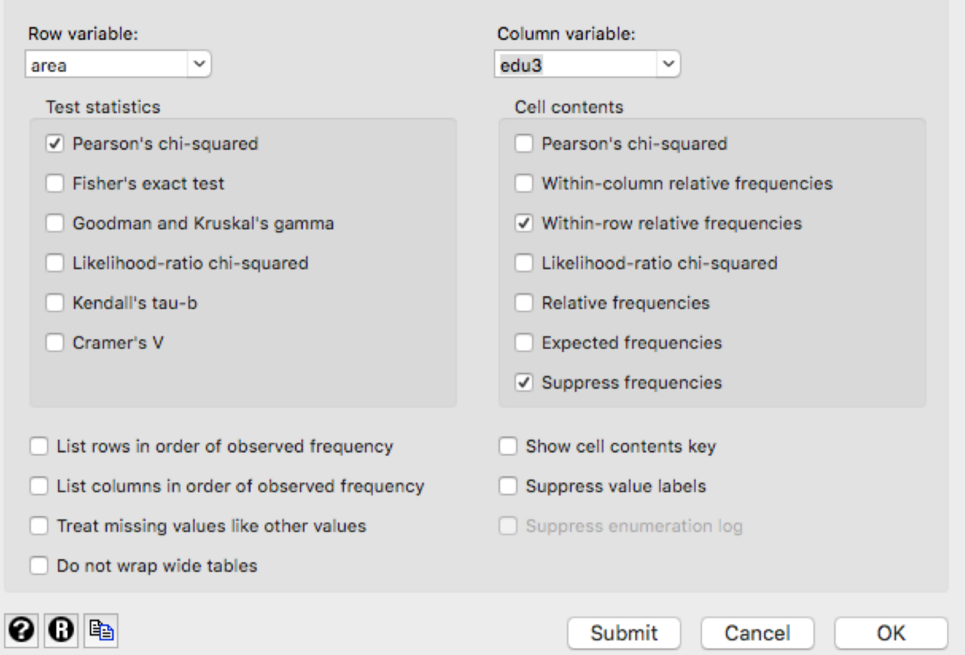
○ 0.01
● 0.05
○ 0.10

The P-Value is .000236. The result is significant at p < .05.

Accept the alternative hypothesis: the two variables are dependent on each other, there is a relationship.

https://www.socscistatistics.com/pvalues/chidistribution.aspx

The Chi-Square Test



Row variable: area

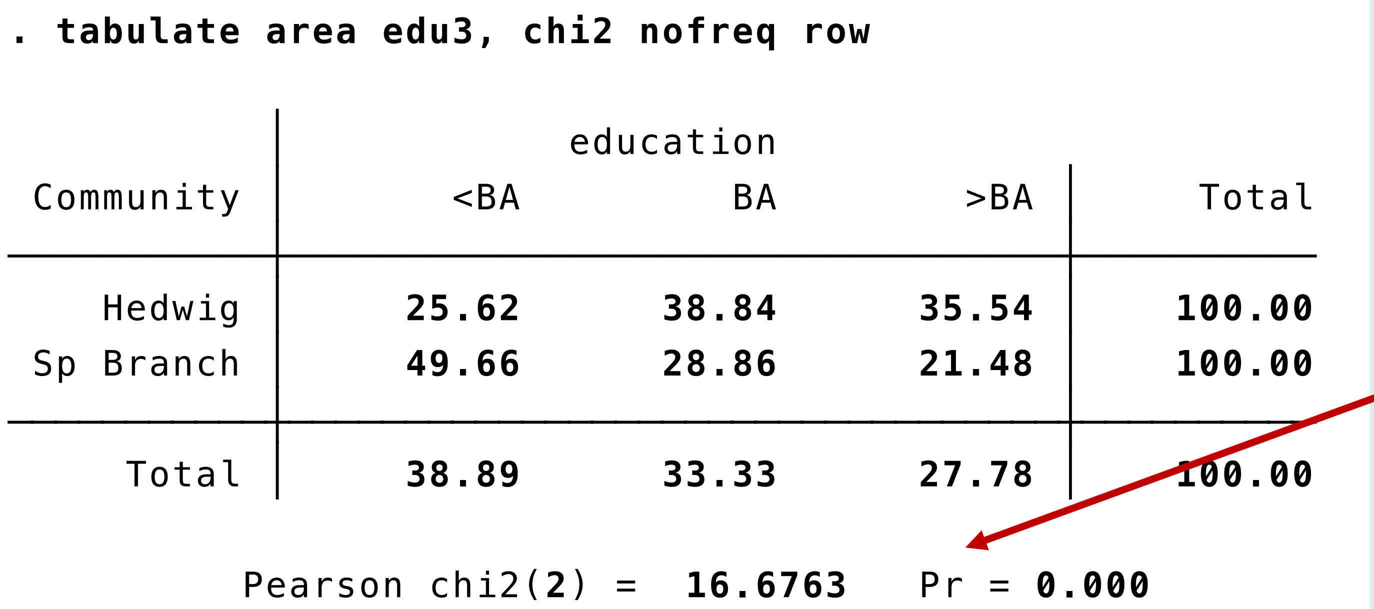
Column variable: edu3

Test statistics
- ☑ Pearson's chi-squared
- ☐ Fisher's exact test
- ☐ Goodman and Kruskal's gamma
- ☐ Likelihood-ratio chi-squared
- ☐ Kendall's tau-b
- ☐ Cramer's V

Cell contents
- ☐ Pearson's chi-squared
- ☐ Within-column relative frequencies
- ☑ Within-row relative frequencies
- ☐ Likelihood-ratio chi-squared
- ☐ Relative frequencies
- ☐ Expected frequencies
- ☑ Suppress frequencies

- ☐ List rows in order of observed frequency
- ☐ List columns in order of observed frequency
- ☐ Treat missing values like other values
- ☐ Do not wrap wide tables
- ☐ Show cell contents key
- ☐ Suppress value labels
- ☐ Suppress enumeration log

In dialog box select row and column variables, select Pearson's chi-squared, either within-row or within-column relative frequencies, and suppress frequencies



```
. tabulate area edu3, chi2 nofreq row

           |            education
 Community |       <BA         BA        >BA |     Total
-----------+---------------------------------+----------
    Hedwig |     25.62      38.84      35.54 |    100.00
 Sp Branch |     49.66      28.86      21.48 |    100.00
-----------+---------------------------------+----------
     Total |     38.89      33.33      27.78 |    100.00

          Pearson chi2(2) =  16.6763   Pr = 0.000
```

In output observe and report the percentages, the chi-square statistic, and it's p-value (here, $p < .001$)

Chi-Square Test Limitations

- Does not offer much information about the strength of the relationship or its substantive significance in the population

- Sensitive to sample size

- Sensitive to small expected frequencies in one or more of the cells in the table

Advanced Application

Crosstabulation can be used to check for the problem of confounding, also called spuriousness.

An analysis may indicate dependence between two variables that is not real, but is due to the action of another, third variable.

To examine this possibility we use Elaboration.

Let’s demonstrate.

Advanced Application

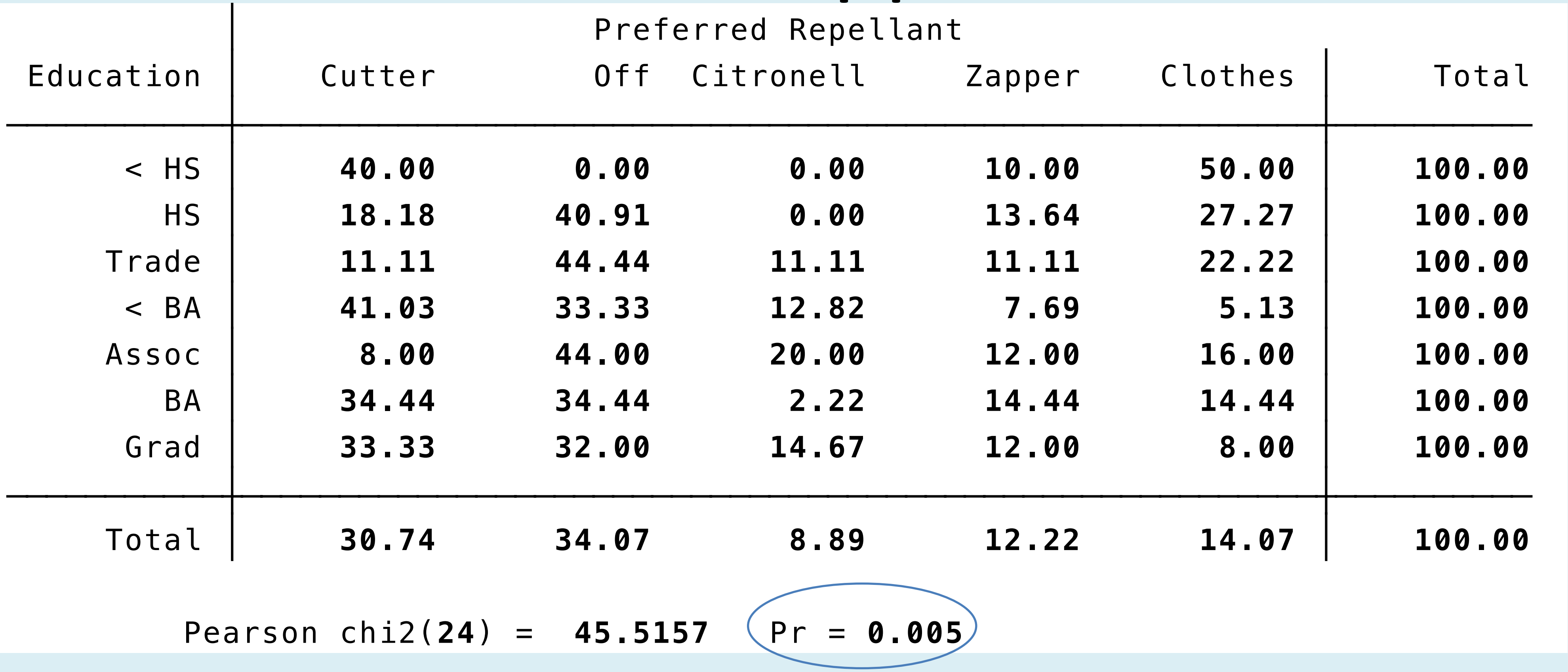
Elaboration Process

1. Divide the observations into subgroups on the basis of the control variable … as many subgroups as there are categories in the control variable.

2. Re-examine the relationship between the original variables separately for the control variable subgroups.

3. Compare the partial relationships with the original bivariate relationship for the total group

Advanced Application



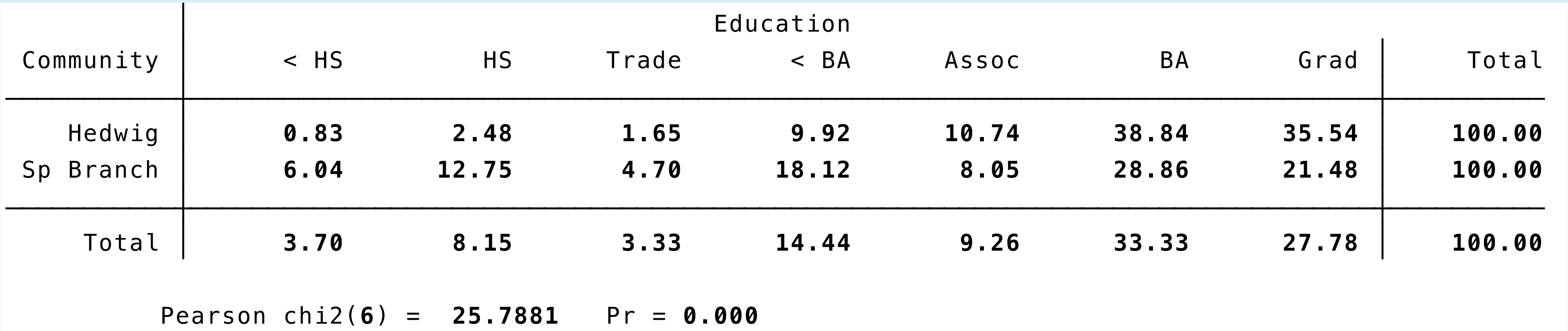
```
           |                Preferred Repellant
 Education |    Cutter        Off  Citronell     Zapper    Clothes |     Total
-----------+-------------------------------------------------------+----------
      < HS |     40.00       0.00       0.00      10.00      50.00 |    100.00
        HS |     18.18      40.91       0.00      13.64      27.27 |    100.00
     Trade |     11.11      44.44      11.11      11.11      22.22 |    100.00
      < BA |     41.03      33.33      12.82       7.69       5.13 |    100.00
     Assoc |      8.00      44.00      20.00      12.00      16.00 |    100.00
        BA |     34.44      34.44       2.22      14.44      14.44 |    100.00
      Grad |     33.33      32.00      14.67      12.00       8.00 |    100.00
-----------+-------------------------------------------------------+----------
     Total |     30.74      34.07       8.89      12.22      14.07 |    100.00

          Pearson chi2(24) =  45.5157   Pr = 0.005
```

Suppose we find a significant association between level of education and most preferred mosquito repellant. It’s difficult to see any meaningful pattern, and doesn’t make a lot of sense.

Advanced Application

What if there is no relationship?

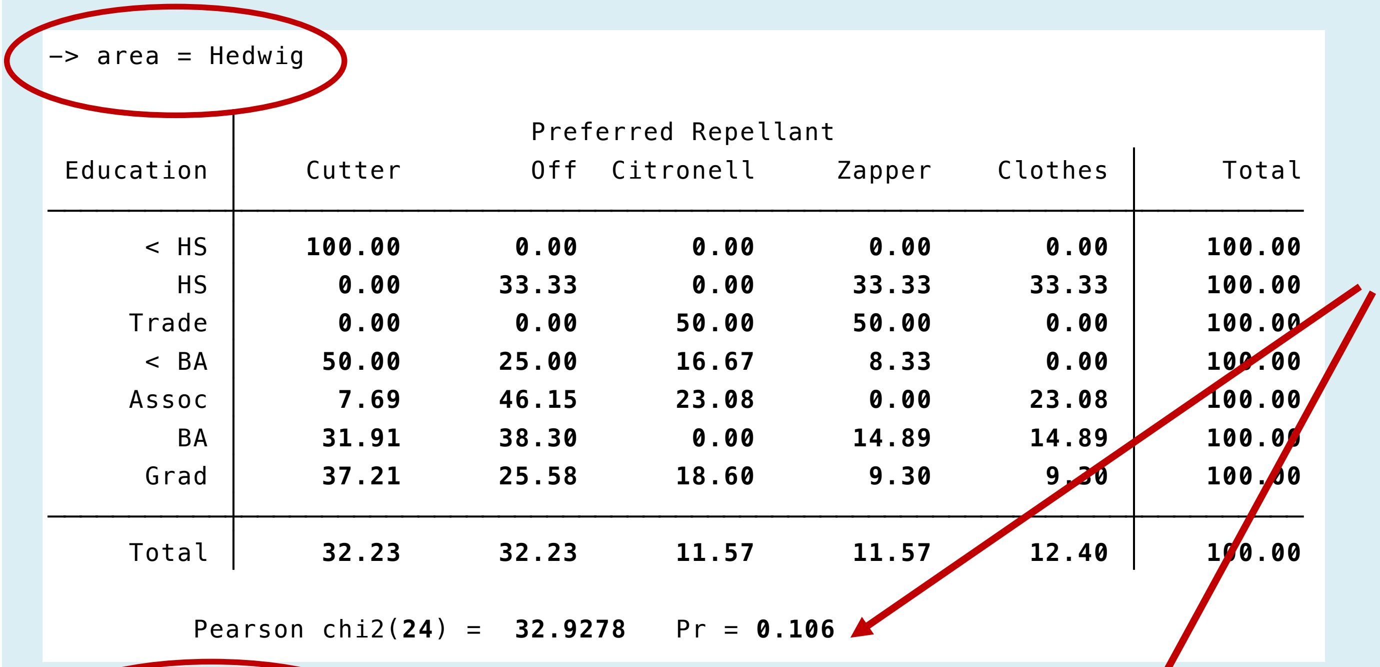
There is also an association between education and the area that the participant lives in, which is a plausible relationship given the different social and economic characteristics of the areas.



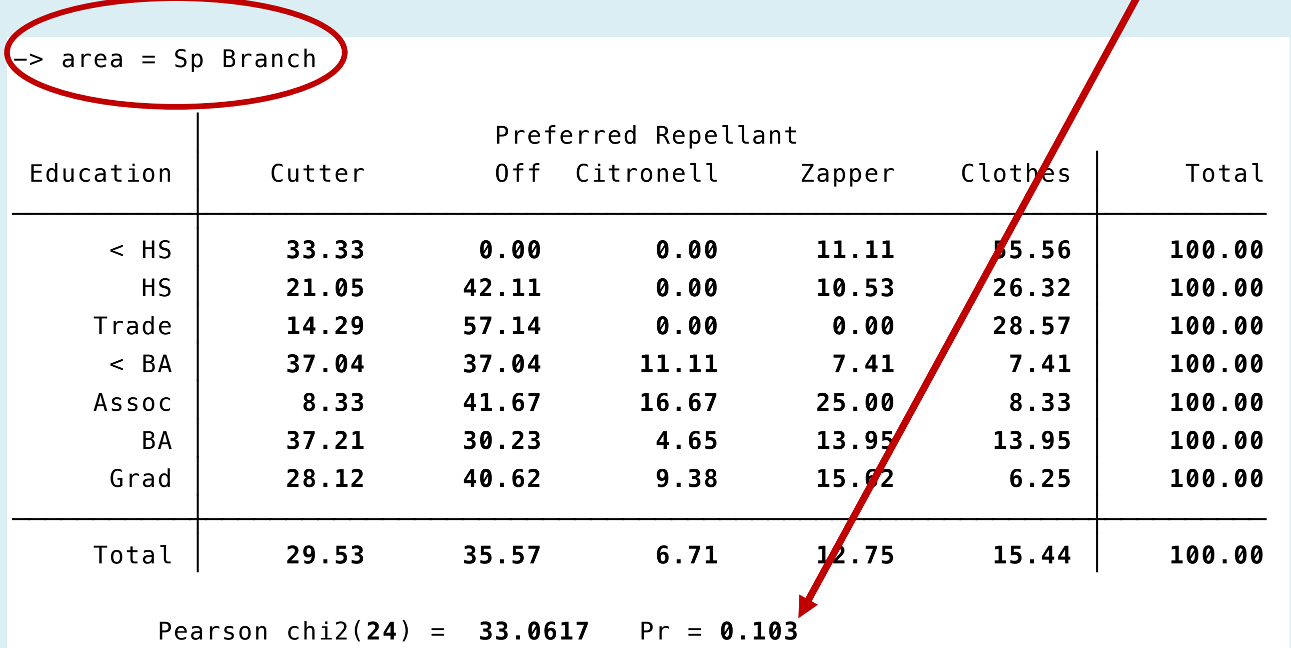
```
           |                              Education
 Community |      < HS         HS      Trade       < BA      Assoc         BA       Grad |     Total
-----------+-----------------------------------------------------------------------------+----------
    Hedwig |      0.83       2.48       1.65       9.92      10.74      38.84      35.54 |    100.00
 Sp Branch |      6.04      12.75       4.70      18.12       8.05      28.86      21.48 |    100.00
-----------+-----------------------------------------------------------------------------+----------
     Total |      3.70       8.15       3.33      14.44       9.26      33.33      27.78 |    100.00

          Pearson chi2(6) =  25.7881   Pr = 0.000
```

To Elaborate we break the analysis into two groups (areas) and see if the association between education and repellant is maintained.

Advanced Application



```
-> area = Hedwig

           |                    Preferred Repellant
 Education |    Cutter        Off  Citronell     Zapper    Clothes |     Total
-----------+-------------------------------------------------------+----------
      < HS |    100.00       0.00       0.00       0.00       0.00 |    100.00 
        HS |      0.00      33.33       0.00      33.33      33.33 |    100.00 
     Trade |      0.00       0.00      50.00      50.00       0.00 |    100.00 
      < BA |     50.00      25.00      16.67       8.33       0.00 |    100.00 
     Assoc |      7.69      46.15      23.08       0.00      23.08 |    100.00 
        BA |     31.91      38.30       0.00      14.89      14.89 |    100.00 
      Grad |     37.21      25.58      18.60       9.30       9.30 |    100.00 
-----------+-------------------------------------------------------+----------
     Total |     32.23      32.23      11.57      11.57      12.40 |    100.00 

          Pearson chi2(24) =  32.9278   Pr = 0.106
```



```
-> area = Sp Branch

           |                    Preferred Repellant
 Education |    Cutter        Off  Citronell     Zapper    Clothes |     Total
-----------+-------------------------------------------------------+----------
      < HS |     33.33       0.00       0.00      11.11      55.56 |    100.00 
        HS |     21.05      42.11       0.00      10.53      26.32 |    100.00 
     Trade |     14.29      57.14       0.00       0.00      28.57 |    100.00 
      < BA |     37.04      37.04      11.11       7.41       7.41 |    100.00 
     Assoc |      8.33      41.67      16.67      25.00       8.33 |    100.00 
        BA |     37.21      30.23       4.65      13.95      13.95 |    100.00 
      Grad |     28.12      40.62       9.38      15.62       6.25 |    100.00 
-----------+-------------------------------------------------------+----------
     Total |     29.53      35.57       6.71      12.75      15.44 |    100.00 

          Pearson chi2(24) =  33.0617   Pr = 0.103
```

It is not.

The association between education and repellant disappears when we look at both areas separately.

Advanced Application

A crosstabulation can also produce several other test statistics. These can be used to better describe the strength of the relationship and/or it's directionality (or trend).

These are described in the text (pp 363-373) but won't be examined for the course.

Key Terms

Bivariate analysis

Bivariate table

Cross-tabulation

Marginals

Cell

Column/Row variable

Chi-square (obtained)

Chi-square test

Expected frequencies

Observed frequencies

Statistical independence

Elaboration

Confounding

JTC280. Lab 7. Testing Independence.

Here is an example using the uranium mining data of a crosstabulation between sex and income, and how to write the results.

Note that by placing sex as the column variable, selecting column percentages and income as the row variables we are defining sex as the independent variable and income as the dependent variable (dependent variable is placed on the side of the table). Consider why this is the only logical arrangement!

```
. tabulate income sex, chi2 column nofreq

     Annual |          Sex
     Income |         f          m |     Total
------------+----------------------+----------
       <25K |     19.05       3.08 |      8.29 
     25-49K |     19.05      19.23 |     19.17 
     50-74K |     26.98      39.23 |     35.23 
    75-100K |     22.22      20.77 |     21.24 
      >100K |     12.70      17.69 |     16.06 
------------+----------------------+----------
      Total |    100.00     100.00 |    100.00 

          Pearson chi2(4) =  15.5642   Pr = 0.004
```

Interpretation: A cross-tabulation was run to evaluate the independence of respondent sex with annual gross income. Results show dependence between the variables in which the percentages of men and women in the 25-50K and 75-100K ranges are about equivalent but there are more women than men in the lowest income bracket and more men than women in the middle 50-75K and highest >100K income brackets. The chi-square test was significant at alpha = .05 (Chi-square = 15.6, *df* = 4, *p* =.004)

example of a **significant** finding	**Interpretation**: A cross-tabulation was run to evaluate the independence of respondent sex with annual gross income. Results show dependence between the variables in which the percentages of men and women in the 25-50K and 75-100K ranges are about equivalent but there are more women than men in the lowest income bracket and more men than women in the middle 50-75K and highest >100K income brackets. The chi-square test was significant at alpha = .05 (Chi-square = 15.6, *df* = 4, *p* =.004)
example of a **non-significant f**inding	**Interpretation**: A cross-tabulation was run to evaluate the independence of respondent sex with annual gross income. Results show independence between the variables in which the percentages of men and women at all salary ranges are about equivalent The chi-square test was not significant at alpha = .05 (Chi-square = 2.6, *df* = 4, *p* =.42)

Exercise

We want to examine some of the relationships among town, water, and actions.
We want to test the following hypotheses (stated as null):

$H1_o$: There is independence between town and water

Here is the frequency table, followed by two empty tables to help you organize your calculations.

```
      Water |         Town
     Source |     Rural  Wellingto |     Total
------------+----------------------+----------
       well |        69         22 |        91
    treated |        48         54 |       102
------------+----------------------+----------
      Total |       117         76 |       193
```

$$x^2 = \sum_{cells}^{all} \frac{(f_e - f_o)^2}{f_e}$$

$$f_e = \frac{(Col.\ marginal)(Row\ marginal)}{N}$$

First, percentage the table <u>for rows:</u>

	Town		Total
Water Source	Rural	Wellington	
Well			
Treated			
Total			

Next, calculate chi-square for the table, and obtain the p-value with the online calculator:

	fe	*fo*	$(fe - fo)$	$(fe - fo)^2$	$(fe - fo)^2 \div fe$
Rural/Well					
Rural/Treated					
Wellington/Well					
Wellington/Treated					

degrees of freedom = Chi-square = Σ

https://www.socscistatistics.com/pvalues/chidistribution.aspx

Q1: What is the value for chi-square? _____

Q2: What is the p-value for the test? _____

Q3: Which is the best interpretation for the test results:

- a) There's a 60/40 split overall with independence
- b) The difference in Wellington is 24/52 with independence
- c) Prevalence of wells is greater in Rural (76/24) with dependence
- d 100% are in Treated, with dependence

We'll use software to continue the analysis.
Use the commands below to run tests for the remaining 2 hypotheses:

$H2_o$: There is independence between water and decide

Statistics --> Summary tables and tests --> Frequency Tables --> Two-way table with measures of association
Row variable: decide Column variable: water
check boxes: Pearson's chi-squared, Within-row relative frequencies, Suppress frequencies

$H3_o$: There is independence between town and decide

Statistics --> Summary tables and tests --> Frequency Tables --> Two-way table with measures of association
Row variable: decide Column variable: town
check boxes: Pearson's chi-squared, Within-row relative frequencies, Suppress frequencies

Test	**Interpretation**
$H1_o$: independence between water and decide	Q4: What is the value for chi-square? 8.7 Q5: What is the p-value for the test? .013 Q6: Which is the best interpretation for the test results: a) non-significant p-value, dependence b) significant p-value, dependence c) non-significant p-value, independence d) significant p-value, independence
$H2_o$: independence between town and decide	Q7: What is the value for chi-square? 2.6 Q8: What is the p-value for the test? .274 Q9: Which is the best interpretation for the test results: a) non-significant p-value, dependence b) significant p-value, dependence c) non-significant p-value, independence d) significant p-value, independence

We also have a variable called actions that indicates how many civic actions the participant has taken on the issues (for or against). We wonder if this variable might control the relationship between water and decide as just observed. Here is the relevant output:

```
_____________________________________________________

-> actions = no actions

 Decision  |      Water Source
    Group  |       well    treated |     Total
-----------+-----------------------+-----------
  In favor |      24.14      75.86 |    100.00
 Undecided |      14.81      85.19 |    100.00
   Against |      41.38      58.62 |    100.00
-----------+-----------------------+-----------
     Total |      27.06      72.94 |    100.00

          Pearson chi2(2) =   5.1894   Pr = 0.075

_____________________________________________________

-> actions = 1 action

 Decision  |      Water Source
    Group  |       well    treated |     Total
-----------+-----------------------+-----------
  In favor |      61.54      38.46 |    100.00
 Undecided |      41.18      58.82 |    100.00
   Against |      54.17      45.83 |    100.00
-----------+-----------------------+-----------
     Total |      51.85      48.15 |    100.00

          Pearson chi2(2) =   1.3161   Pr = 0.518

_____________________________________________________

-> actions = 2+ actions

 Decision  |      Water Source
    Group  |       well    treated |     Total
-----------+-----------------------+-----------
  In favor |     100.00       0.00 |    100.00
 Undecided |      75.00      25.00 |    100.00
   Against |      72.50      27.50 |    100.00
-----------+-----------------------+-----------
     Total |      74.07      25.93 |    100.00

          Pearson chi2(2) =   0.7570   Pr = 0.685
```

Q10: What is the best conclusion about the question of whether actions control the association between water source and decision group?

a) all subgroups are significant, so there is control
b) all subgroups are non-significant, so there is no control
c) all subgroups are non-significant, so there is control

Analysis of Variance

or

ANOVA

ANOVA

An inferential technique designed to test for significant differences in means across three or more groups.

Dependent variable:

Interval level

Independent (Grouping) variable:

Categorical or Ordinal

ANOVA

One-way ANOVA

An analysis of variance procedure using one dependent and one independent variable.

Factorial ANOVA

An analysis of variance procedure using one dependent and multiple independent variables.

(more advanced technique)

ANOVA

Fundamental logic:

Significant differences in means across groups can be tested by partitioning the variance:

Calculating the portion of variance that can be seen **within** each of the groups and that which can be seen **between** groups.

ANOVA

Fundamental logic:

Variance within each of the groups: error, noise.

Variance between groups: explanatory, signal.

Variance Within + Variance Between = Total Variance

Hypothesis Testing with ANOVA

Stating the Research and Null Hypotheses

H_a:

At least *one mean* is different from the others.

H_0:

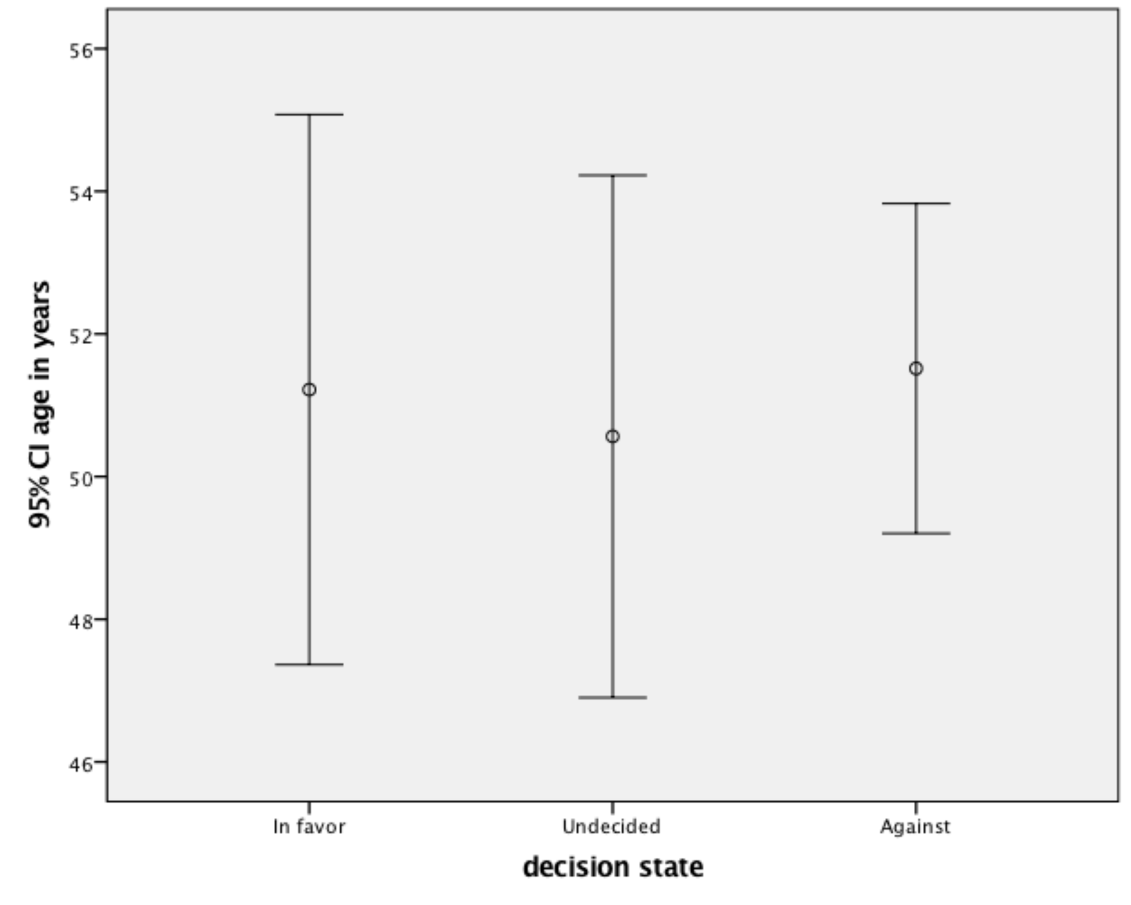
All means are equal: $\mu_1 = \mu_2 = \mu_3 = \mu_k$

ANOVA: a visual approach

Looking at the semester lab dataset …

Respondents reporting to be in one of three decision states concerning the mine --> categorical independent variable

Does the average age vary across these groups?



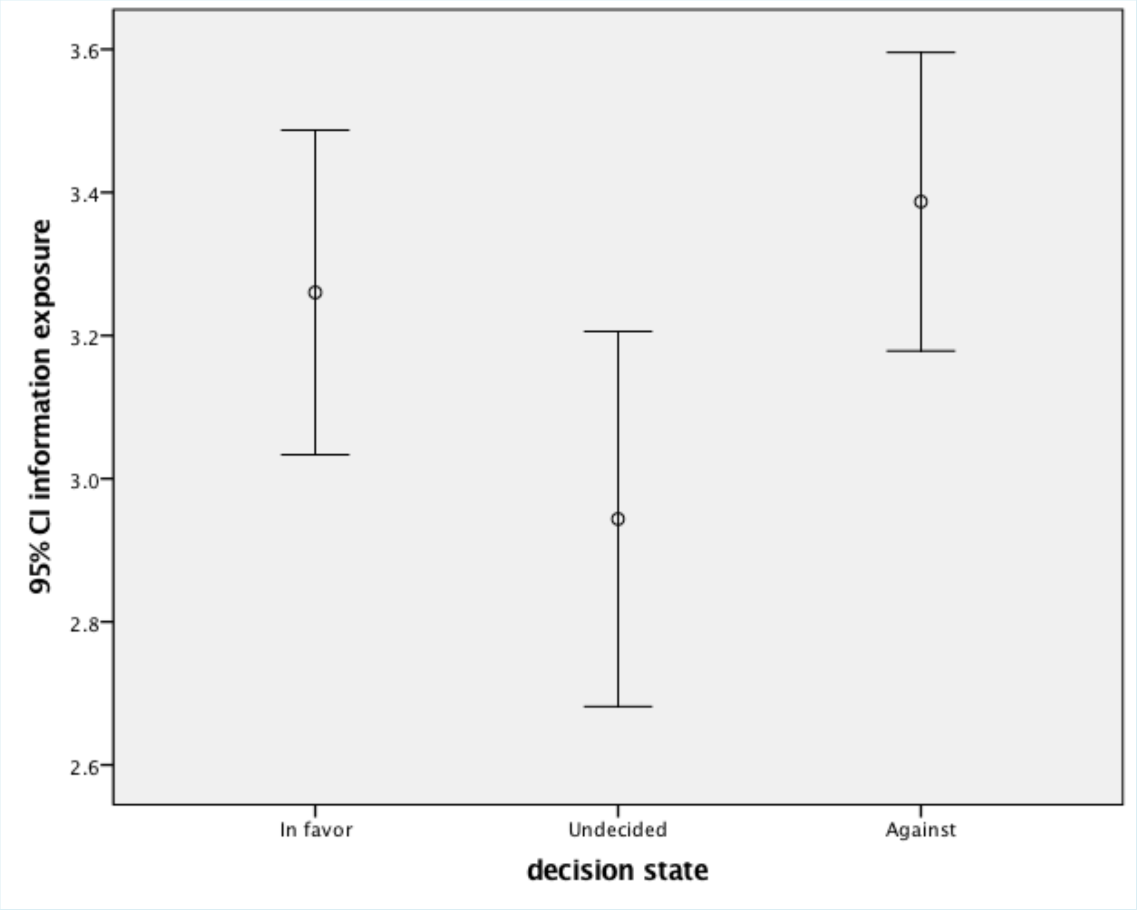
H_0: $\mu1 = \mu2 = \mu3$
do not reject the null

no evidence that ages are different.

ANOVA: a visual approach

Variable: how much information about the mine respondents report having encountered (scale, interval measure).

Does this vary across groups?



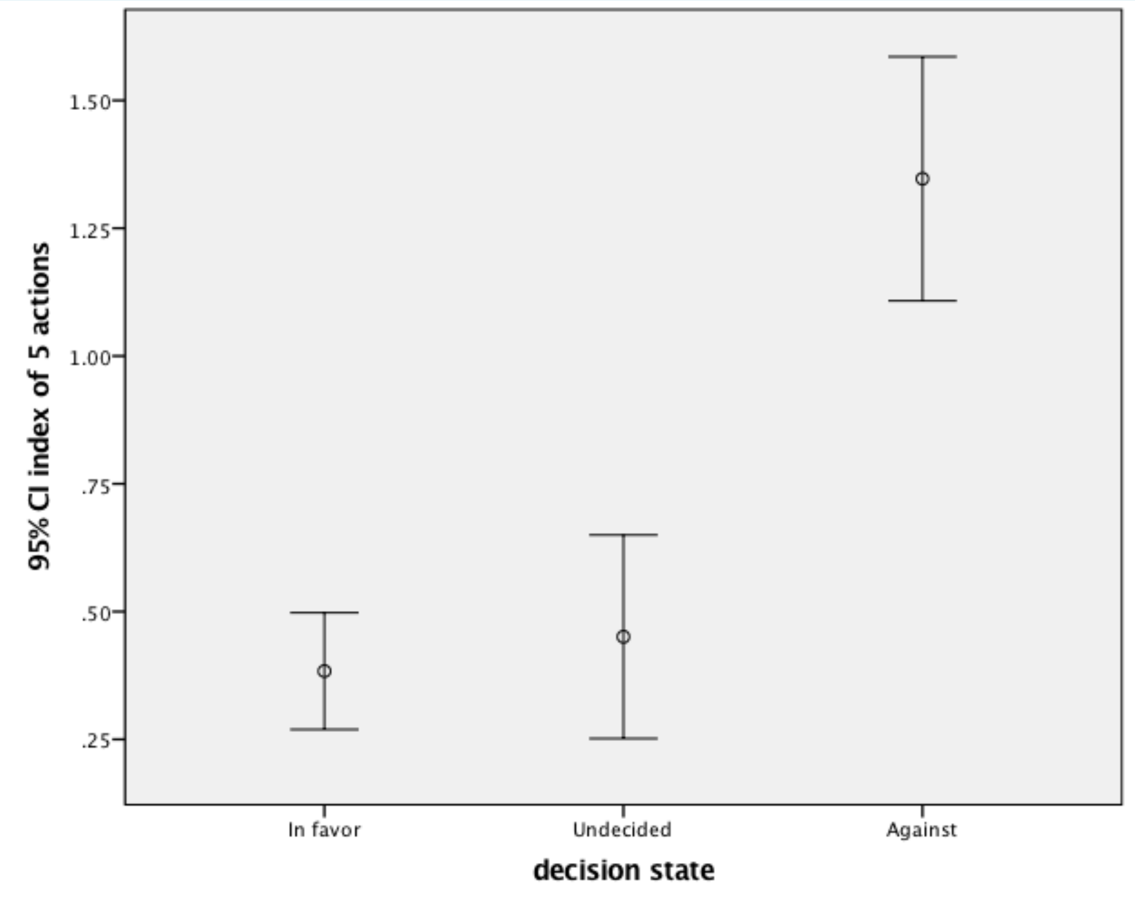
H_0: $\mu1 = \mu2 = \mu3$
do not reject the null

no evidence that information scores are different

ANOVA: a visual approach

Variable: how many actions respondents report having taken on the issue (scale, interval measure).

Does this vary across groups?

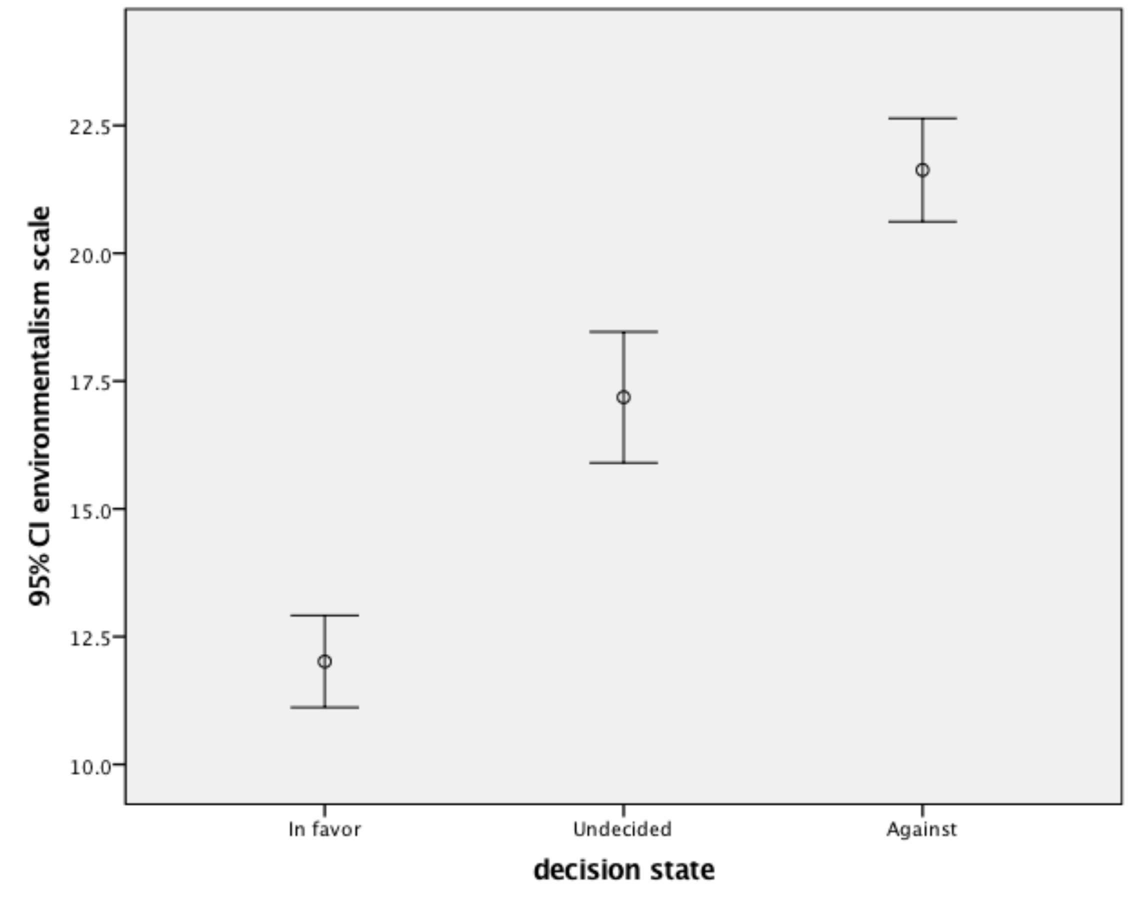


H_0: $\mu1 = \mu2 = \mu3$
reject the null

evidence that one group is different.

ANOVA: a visual approach

Variable: score on environmentalism (scale, interval measure). Does this vary across groups?

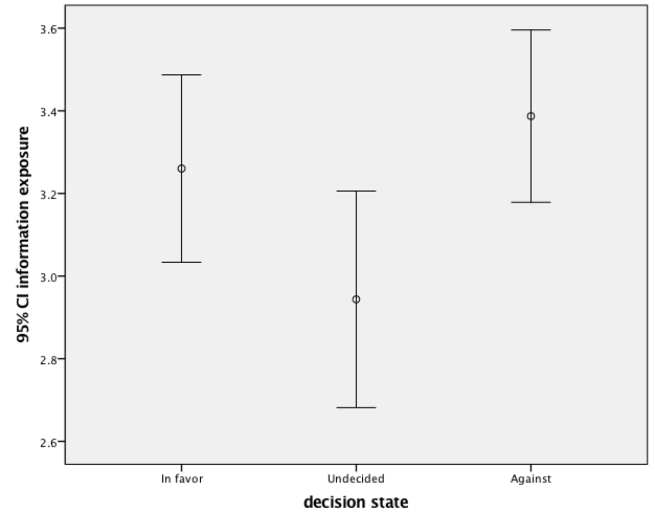


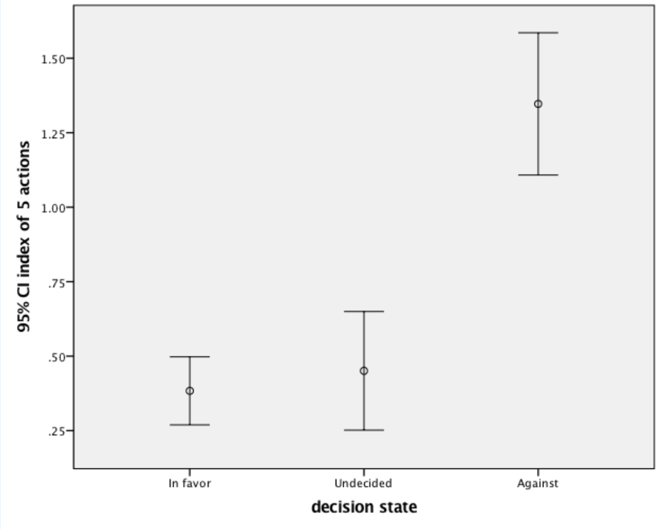
H_0: $\mu1 = \mu2 = \mu3$
reject the null

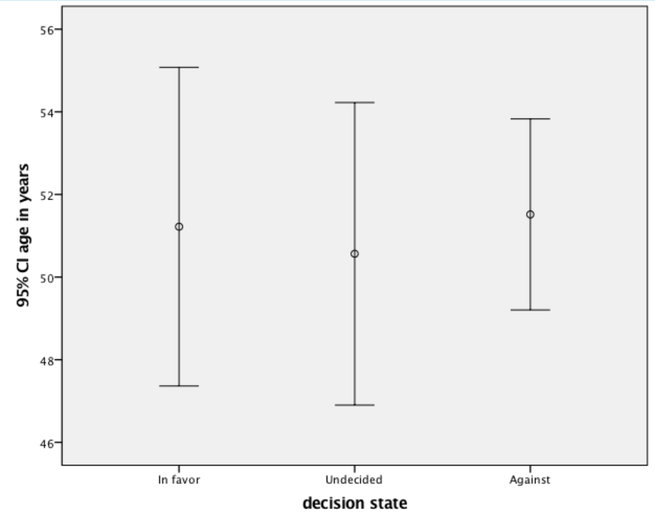
evidence that all groups are different.

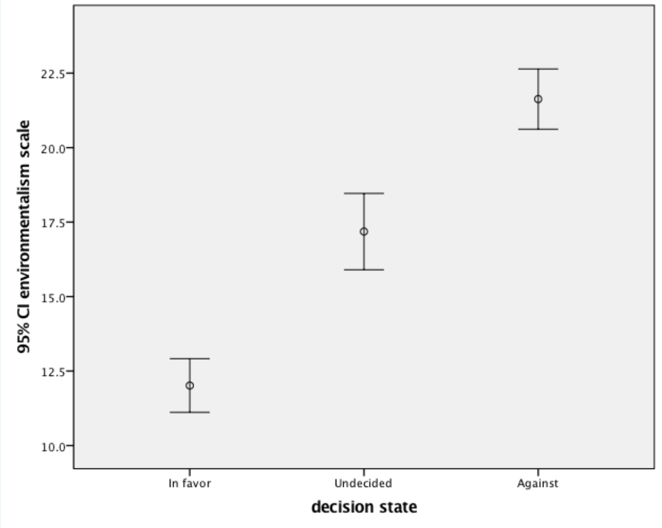
ANOVA: a visual approach

Overall, these demonstrate the range of possibilities. But just as with confidence intervals around two means or two proportions there can be ambiguity.



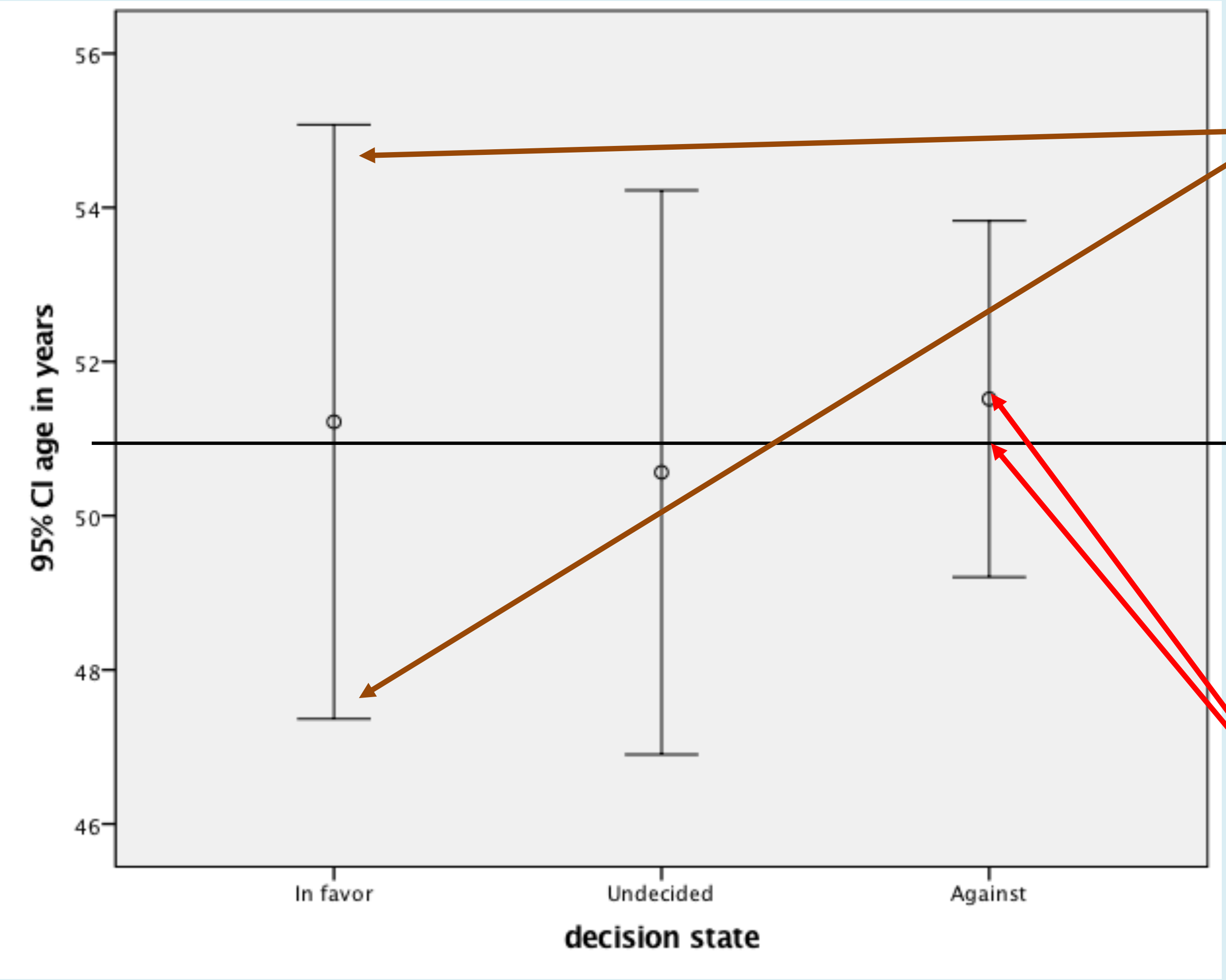






ANOVA: a visual approach

So, we examine the variance components to see if the differences within each group are greater than the differences across groups:



Variance within each group summed

Group 1 ~ 8
Group 2 ~ 8
Group 3 ~ 5 **= 21**

Overall average
“Grand Mean”
X-doublebar

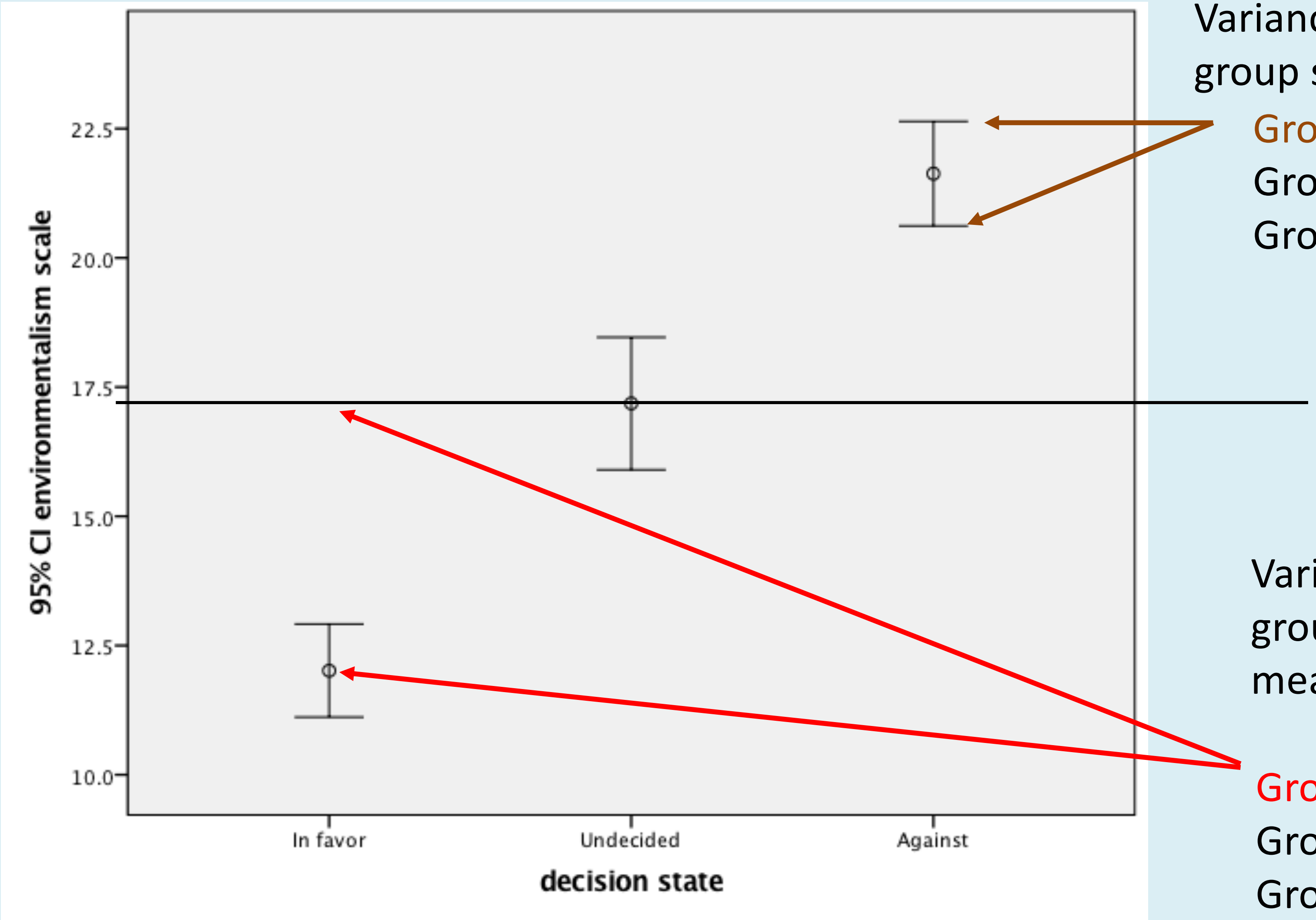
Variance between each group and the grand mean, summed

Group 3 ~ 1
Group 2 ~ 1
Group 1 ~ .5 **= 2.5**

Noise = 21 Signal = 2.5

ANOVA: a visual approach

So, we examine the variance components to see if the differences within each group are greater than the differences across groups:



Variance within each group summed

Group 3 ~ 2.5
Group 2 ~ 2.7
Group 1 ~ 2.3 **= 7.5**

Overall average
"Grand Mean"
X-doublebar

Variance between each group and the grand mean, summed

Group 1 ~ 5
Group 2 ~ 0
Group 3 ~ 5 **= 10**

Noise = 7.5 Signal = 10

Turning this into a test statistic: *F*

Ratio of between-group variance to within-group variance.

Technique: "Sum of Squares."

SSB: Sum of Squares Between

SSW: Sum of Squares Within

MSB: Mean Square Between

MSW: Mean Square within

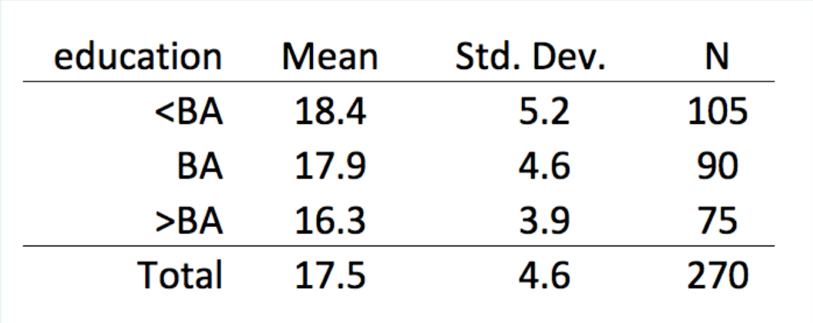
$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$

$$\mathrm{df}_\mathrm{b} = k - 1 \qquad \mathrm{df}_\mathrm{w} = N - k$$

where k = number of groups, N = total observations

Turning this into a test statistic: *F*

$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$

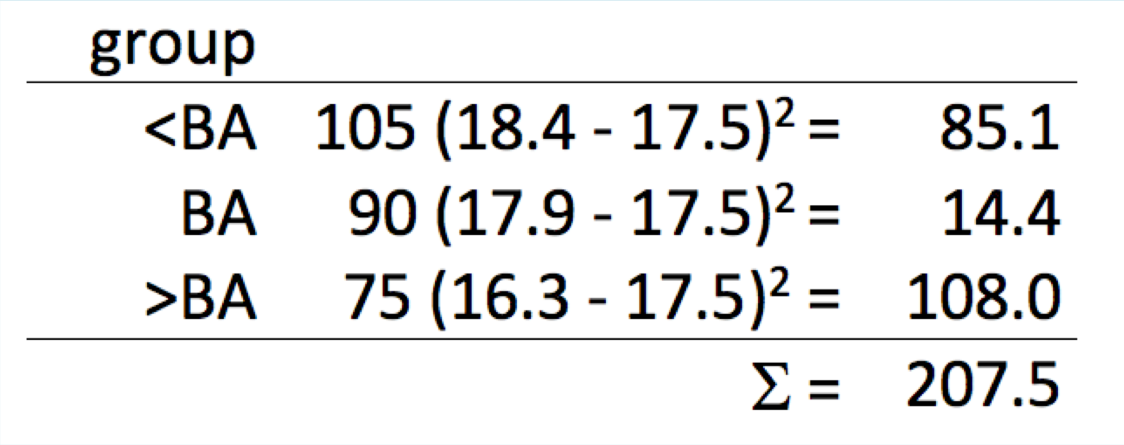


education	Mean	Std. Dev.	N
<BA	18.4	5.2	105
BA	17.9	4.6	90
>BA	16.3	3.9	75
Total	17.5	4.6	270

Work from a summary table, here we have have the “mosquito protection score” broken by three educational levels.

1. Find the Sum of Squares Between:

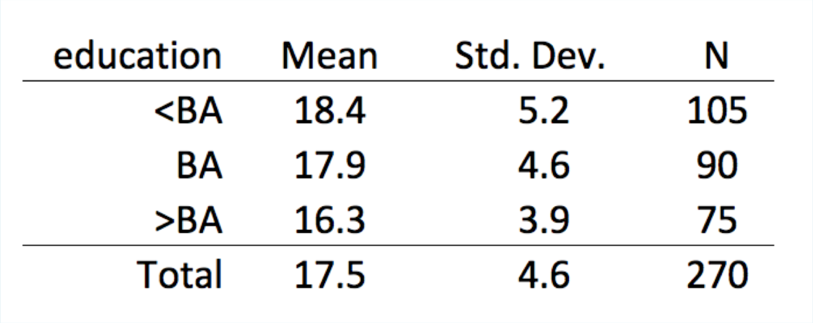
$$SSB = \sum_{i=1}^{k} n_i \, (\bar{Y}_i - \bar{\bar{Y}})^2$$



group		
<BA	105 (18.4 - 17.5)^2 =	85.1
BA	90 (17.9 - 17.5)^2 =	14.4
>BA	75 (16.3 - 17.5)^2 =	108.0
	Σ =	207.5

Turning this into a test statistic: *F*

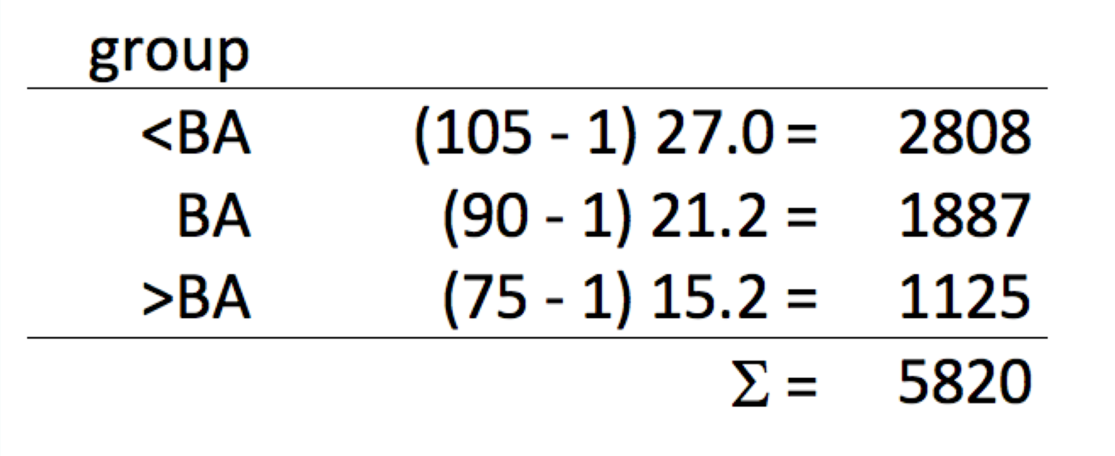
$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$



education	Mean	Std. Dev.	N
<BA	18.4	5.2	105
BA	17.9	4.6	90
>BA	16.3	3.9	75
Total	17.5	4.6	270

2. Find the Sum of Squares Within:

$$SSW = \sum_{i=1}^{k} (n_i - 1)s_i^2$$



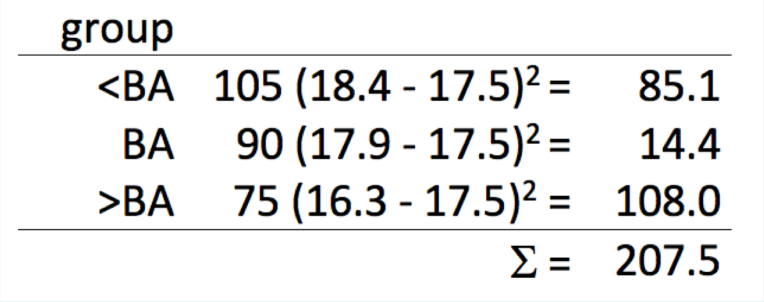
group		
<BA	(105 - 1) 27.0 =	2808
BA	(90 - 1) 21.2 =	1887
>BA	(75 - 1) 15.2 =	1125
	Σ =	5820

Turning this into a test statistic: *F*

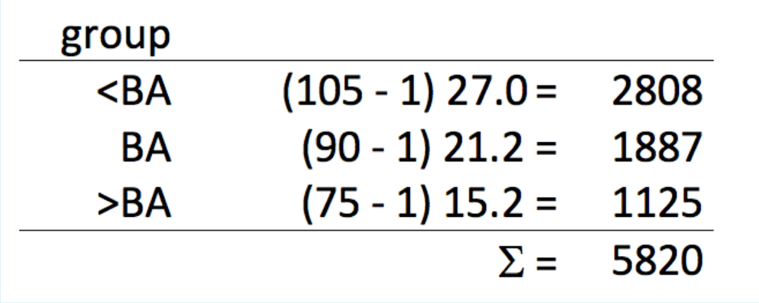
$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$

$$\mathrm{df}_\mathrm{b} = k - 1 \qquad \mathrm{df}_\mathrm{w} = N - k$$

where k = number of groups, N = total observations



group		
<BA	$105\,(18.4 - 17.5)^2 =$	85.1
BA	$90\,(17.9 - 17.5)^2 =$	14.4
>BA	$75\,(16.3 - 17.5)^2 =$	108.0
	$\Sigma =$	207.5



group		
<BA	(105 - 1) 27.0 =	2808
BA	(90 - 1) 21.2 =	1887
>BA	(75 - 1) 15.2 =	1125
	$\Sigma =$	5820

3. Find the Mean Squares and *F*:

$$\frac{207.5 \div 2 \qquad = 103.8}{5820 \div 267 \qquad = \;\; 21.8}$$

$$4.8 = F_{2,\,267}$$

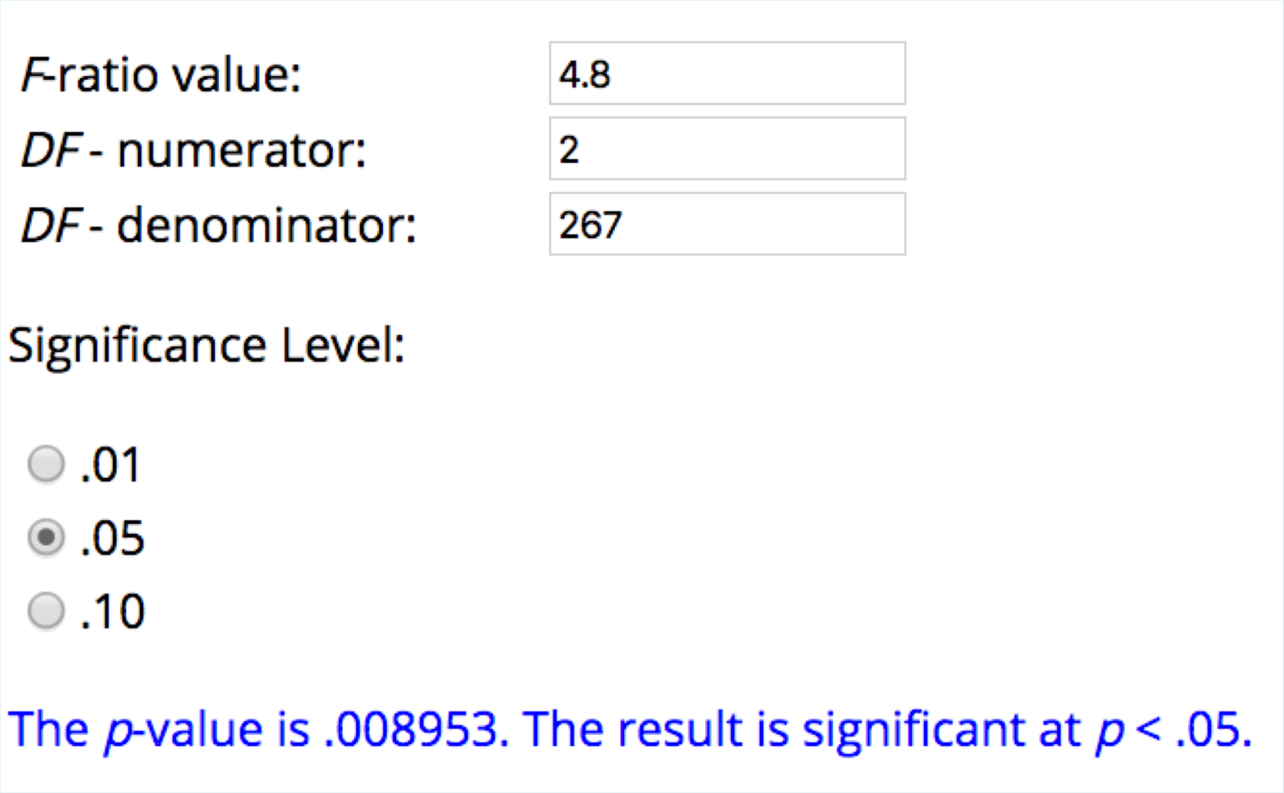
Turning this into a test statistic: *F*

4. Find the probability for the F score:

https://www.socscistatistics.com/pvalues/fdistribution.aspx

$$\frac{207.5 \div 2 \qquad = 103.8}{5820 \div 267 \quad = \quad 21.8}$$

$$4.8 = F_{2,\, 267}$$



F-ratio value:	4.8
DF - numerator:	2
DF - denominator:	267

Significance Level:

○ .01

● .05

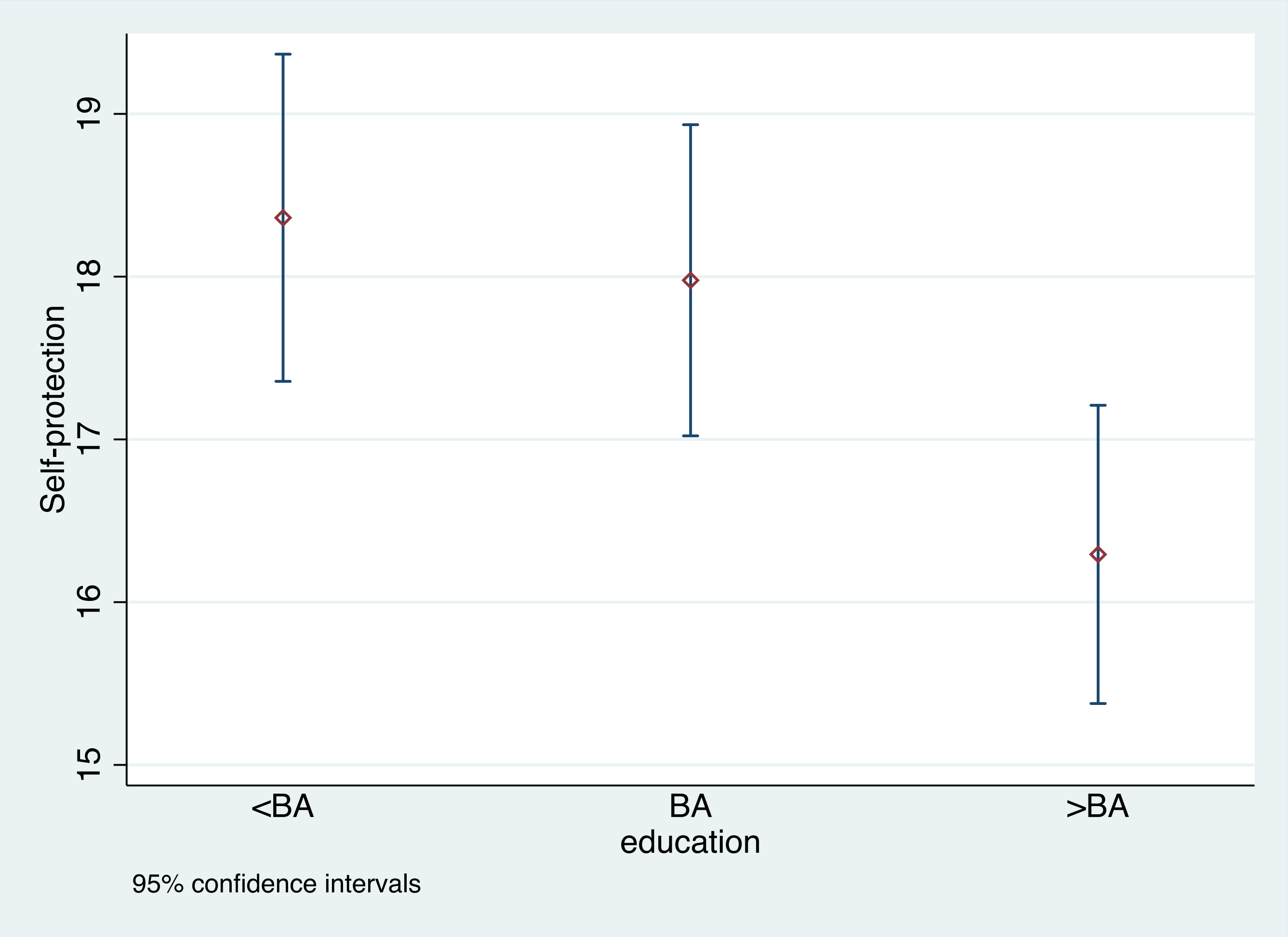
○ .10

The *p*-value is .008953. The result is significant at $p < .05$.

Reject the null hypothesis and conclude that at least two of the means are different at the 95% confidence level.

Turning this into a test statistic: *F*

Visualizing the differences:



It would appear that the self-protection score for the <BA and BA groups are about the same, but the >BA group is lower than both of those.

Organizing a source table

A handy way to calculate an ANOVA is with a “source table” that sets up the calculations for Sum of Squares, Mean Squares, and F:

	N	Mean	SD	s^2
Program A	30	50.26	10.45	109.2
Program B	30	45.32	12.76	162.8
Program C	30	53.67	11.47	131.6
Total	90	49.75		

$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$

Variance	Sum of Squares	df	MS	*F*	*p*
Between $SSB = \sum_{i=1}^{k} n_i\,(\bar{Y}_i - \bar{\bar{Y}})^2$	$30\,(50.26-49.75)^2 = 8$ $30\,(45.32-49.75)^2 = 589$ $30\,(53.67-49.75)^2 = 461$ $\Sigma = 1058$	2	529	3.92	.023
Within $SSW = \sum_{i=1}^{k} (n_i - 1)s_i^2$	29 * 109.2 = 3167 29 * 162.8 = 4721 29 * 131.6 = 3816 $\Sigma = 11704$	87	135		
Total	12,762	89			

Problem of Multiple Comparisons

The problem of Family-Wise Error is caused by the fact that we're doing multiple probability calculations simultaneously. This has the consequence of inflating the likelihood of making a Type I error (false positive).

So, all the ANOVA test can formally tell us is that at least two of the means are different.

But which two, or three?

Problem of Multiple Comparisons

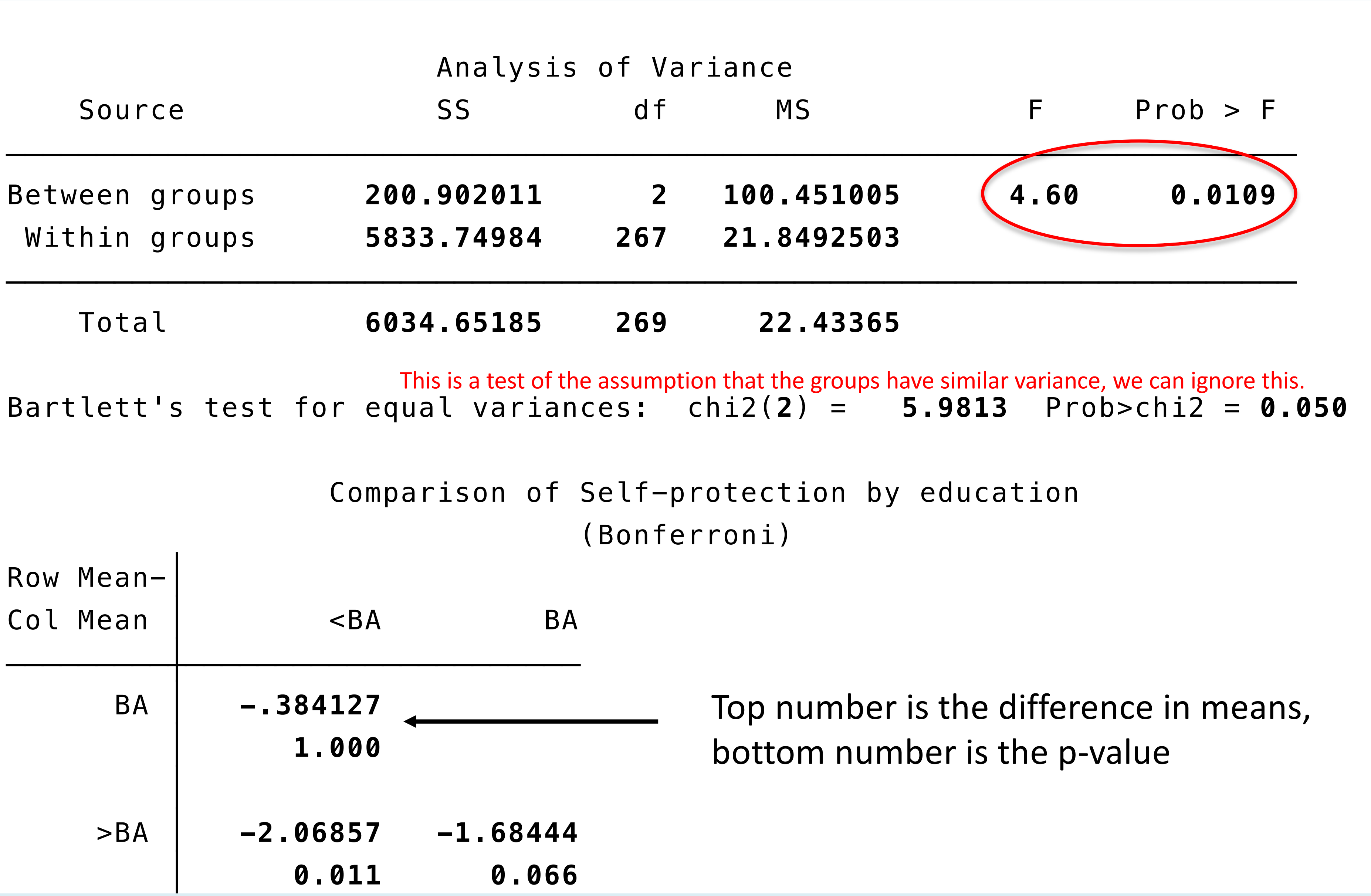
But which two, or three?

To answer this we must do a set of “post-hoc” t-tests. These are all of the possible pair-wise t-test that the ANOVA includes, but the formula used makes and adjustment for the Family-Wise Error.

For this we’ll use software.

Problem of Multiple Comparisons

Stata ANOVA output organizes the calculations (values differ due to rounding):

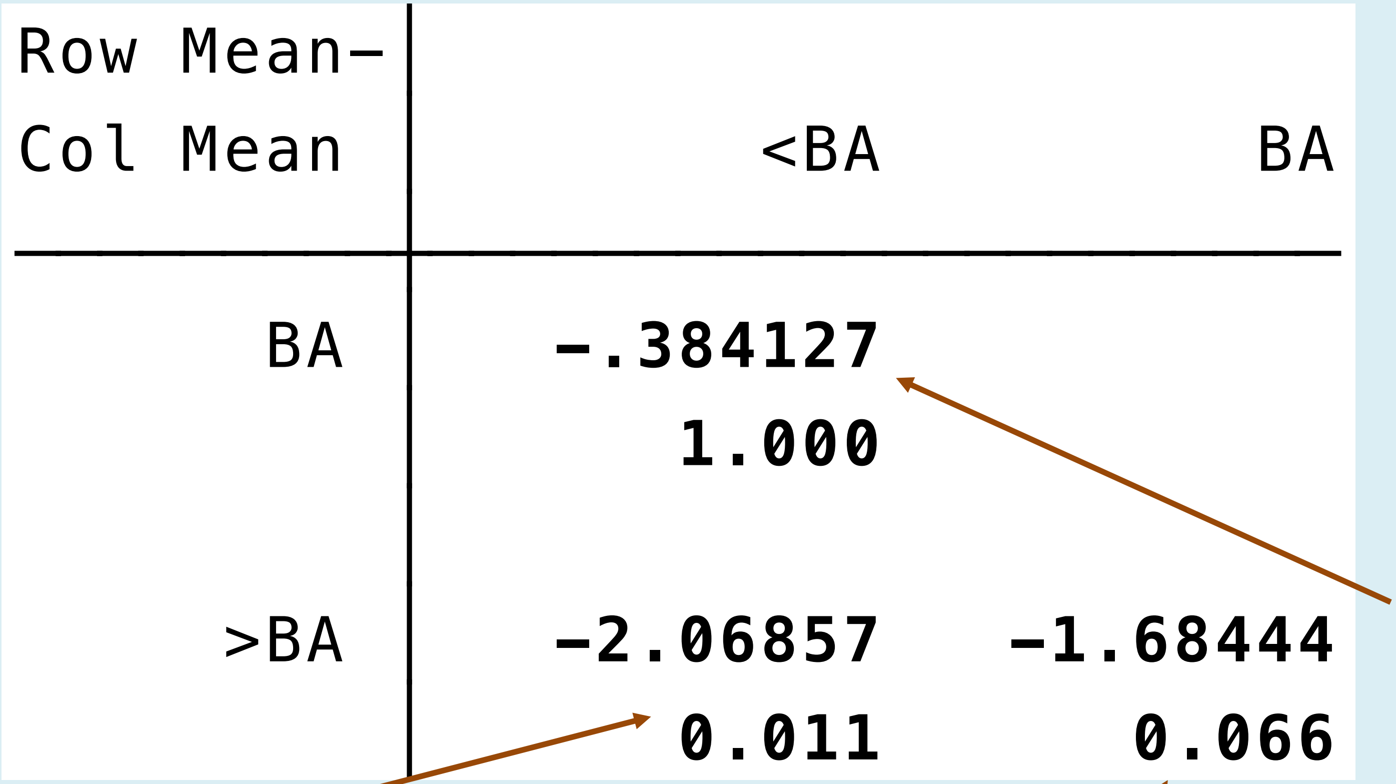


```
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      200.902011      2   100.451005      4.60     0.0109
 Within groups      5833.74984    267   21.8492503
------------------------------------------------------------------------
    Total           6034.65185    269     22.43365

Bartlett's test for equal variances:  chi2(2) =   5.9813  Prob>chi2 = 0.050

                  Comparison of Self-protection by education
                                (Bonferroni)
Row Mean-|
Col Mean |        <BA          BA
---------+----------------------
      BA |   -.384127
         |      1.000
         |
     >BA |   -2.06857    -1.68444
         |      0.011       0.066
```

This is a test of the assumption that the groups have similar variance, we can ignore this.

Top number is the difference in means, bottom number is the p-value

The cells in the table have the post-hoc t-test for each unique combination.



```
Row Mean-|
Col Mean |        <BA         BA
---------+----------------------
      BA |   -.384127
         |      1.000
         |
     >BA |   -2.06857   -1.68444
         |      0.011      0.066
```

Top left is the test for the difference in means between the <BA group and the BA group.

The difference is small and is non-significant.

Here we see what the chart illustrated. The mean for the >BA group is lower (by 2.06) than the <BA group (p = .01.)

But HERE we see that our visual evaluation of the ANOVA was incorrect. The difference between the >BA group and the BA group (1.68) is not large enough to be significant!

Key Terms

One-way ANOVA

Factorial ANOVA

Partitioning of variance

ANOVA hypothesis test

Within-group sum of squares (SSW)

Between-group sum of Squares (SSB)

Total sum of squares (SST)

Degrees of freedom between

Degrees of freedom within

Mean square between

Mean square within

F statistic

Family-wise error

Post-hoc t-tests

JTC270 Lab 8 ANOVA.

$$SSB = \sum_{i=1}^{k} n_i \, (\bar{Y}_i - \bar{\bar{Y}})^2 \qquad SSW = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

$$F = \frac{MSB}{MSW} = \frac{SSB \div df_b}{SSW \div df_w}$$

$$\mathrm{df_b} = k - 1 \qquad \mathrm{df_w} = N - k$$

where k = number of groups, N = total observations

To obtain p-value for F: http://www.socscistatistics.com/pvalues/fdistribution.aspx

Study data: From the uranium mine study we want to examine the dependent variable "actions" which is is coded as none, one, or two-or-more actions taken in the mine controversy. We will treat this as an ordinal variable.

We'll compare two independent variables. The first is simply the respondents' age. The other is "information exposure" which scores the individual on the amount of information actually found. We'll treat both of these as interval measures. For details on these variables and the survey questions used, see the semester dataset codebook.

Calculate an ANOVA for both of these summary tables and complete the questions in Canvas.

Round calculations to two decimal places.

Age

```
            |        Summary of Age
    Actions |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  no action |   50.541176   16.187556          85
   1 action |   54.296296   15.210014          54
  2+ action |   50.037037   12.435438          54
------------+------------------------------------
      Total |   51.450777    14.97981         193
```

These tables can help organize your calculations, then complete the ANOVA table.

group	SSB	
no action		=
1 action		=
2+ actions		=
	Σ =	

group	SSW	
no action		=
1 action		=
2+ actions		=
	Σ =	

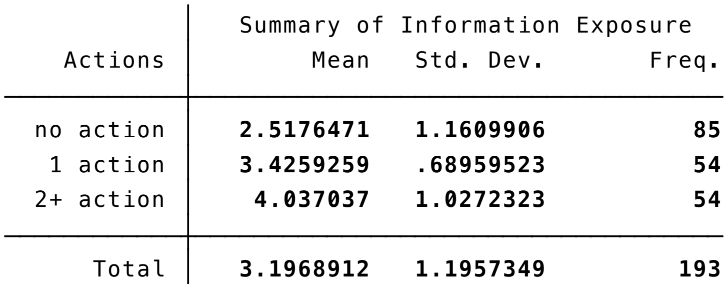
	df	Sum of Sq	Mean Sq	F	p-value
Between					
Within					

Q1. What is the achieved F-value? ___

Q2. What is the outcome of the test?

a) test is significant, all of the means are equal
b) test is non-significant, all means are equal
c) test is significant, at least one mean is different
d) test is non-significant, at least one mean is different

Information Exposure



```
            |   Summary of Information Exposure
    Actions |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  no action |   2.5176471   1.1609906          85
   1 action |   3.4259259   .68959523          54
  2+ action |    4.037037   1.0272323          54
------------+------------------------------------
      Total |   3.1968912   1.1957349         193
```

These tables can help organize your calculations, then complete the ANOVA table.

group	SSB	
no action		=
1 action		=
2+ actions		=
	Σ =	

group	SSW	
no action		=
1 action		=
2+ actions		=
	Σ =	

	df	Sum of Sq	Mean Sq	F	p-value
Between					
Within					

Q3. What is the achieved F-value? _____

Q4. What is the outcome of the test?

a) test is significant, all of the means are equal

b) test is non-significant, all means are different

c) test is significant, at least one mean is different

d) test is non-significant, at least one mean is different

Finally, we'll use Stata to analyze the mean values of the variable "information seeking" across the three levels of the variable "decide" and also include the post-hoc t-tests.

"Information seeking" is a measure of how much effort the participant has made looking for information on the mine issue (1 = little, 5 = a lot). "Decide" indicates the group that the participants place themselves in: for the mine, neutral, or against.

To run the analysis open the semester lab dataset and execute these commands:

Statistics --> Linear Models and Related --> ANOVA/MANOVA --> one-way ANOVA
Response variable: infoseek   Factor variable: decide
check: Bonferroni   check: Produce summary table   submit

To plot the 95% confidence intervals for the groups, in the Command window at the bottom of the Stata frame enter this code: ciplot infoseek, by(decide)   return

The post-hoc tests are at the bottom, and these are described near the end of the lecture slides.

Q5. What is the achieved F-value? _____

Q6. What is the outcome of the F-test?
a) test is significant, all of the means are equal
b) test is non-significant, all means are different
c) test is significant, at least one mean is different
d) test is non-significant, at least one mean is different

Q7. What does the post-hoc t-test tell us?
a) all three means are significantly different
b) all three means are equal
c) the mean for the "in favor" group is different from the "undecided" group.
d) the mean for against is different from the other two groups, which are the same

Regression and Correlation

Simple Regression

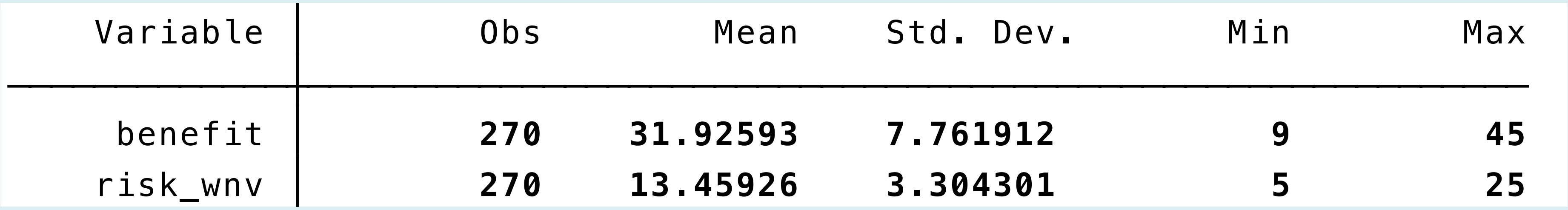
Visual analysis: The Scattergram

Used to display a relationship between two interval-ratio variables.

Let’s look at a plot of two interval variables

Benefit: How much benefit the person senses by avoiding mosquitoes, high values → greater benefit. Measured with multiple questions that are then summed for the scale.

Risk: How much health risk the person senses from West Nile virus, high values → greater risk. Measured with multiple questions that are then summed for the scale.

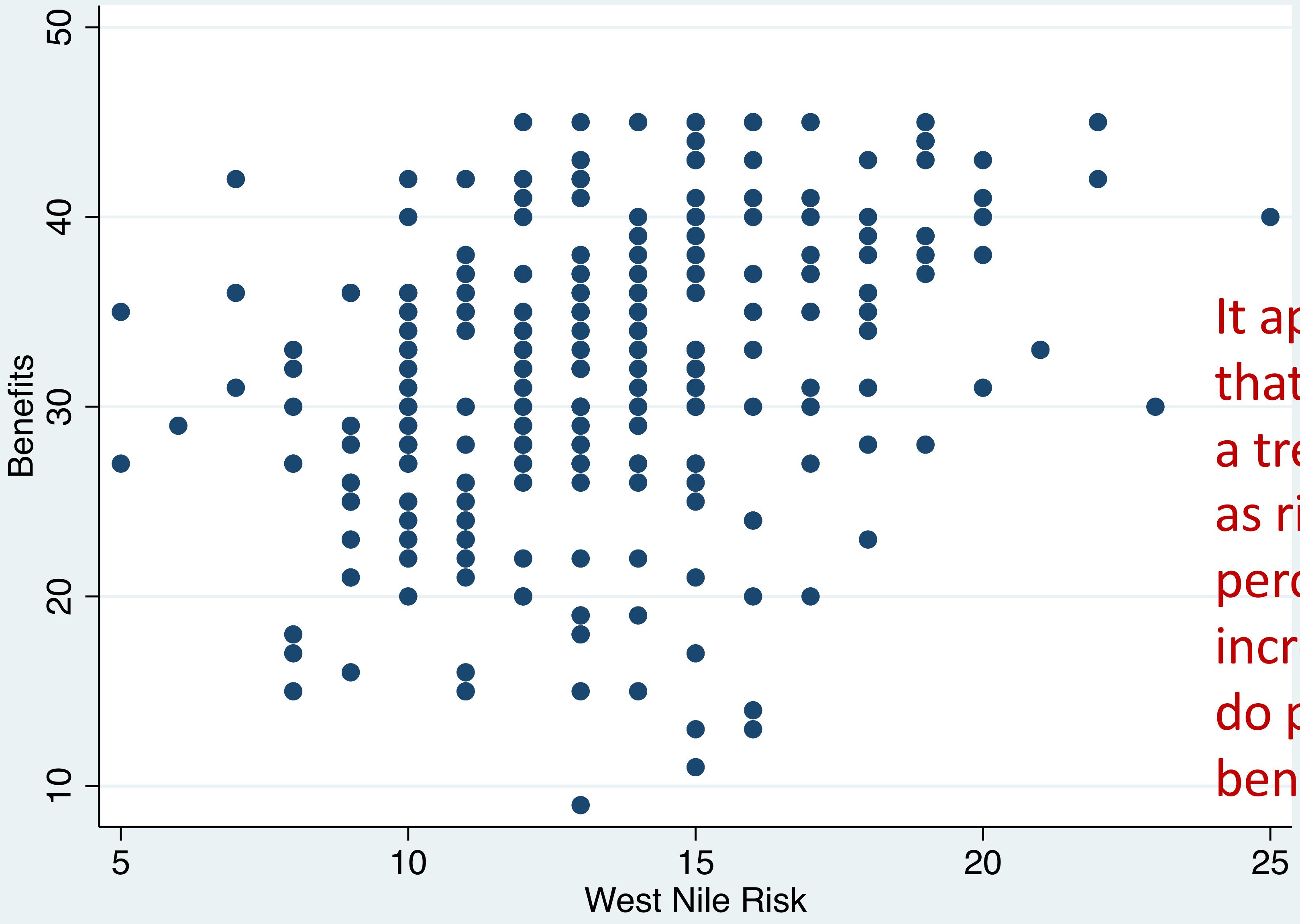


```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     benefit |        270    31.92593    7.761912          9         45
    risk_wnv |        270    13.45926    3.304301          5         25
```

Theory would suggest that those who see great risk from WNv will also perceive greater benefits from avoiding mosquitoes.

Simple Regression

Visual analysis: The Scattergram

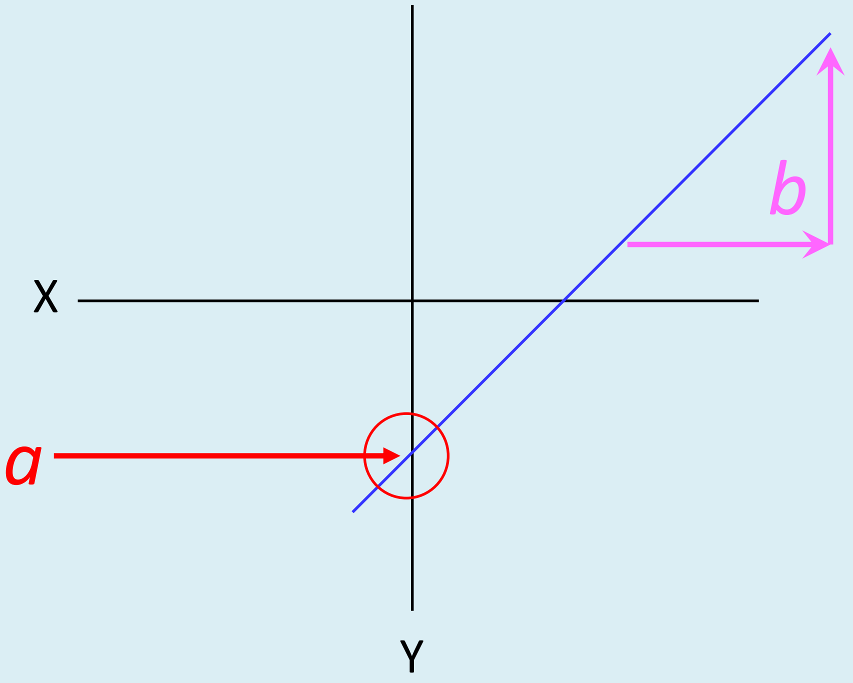


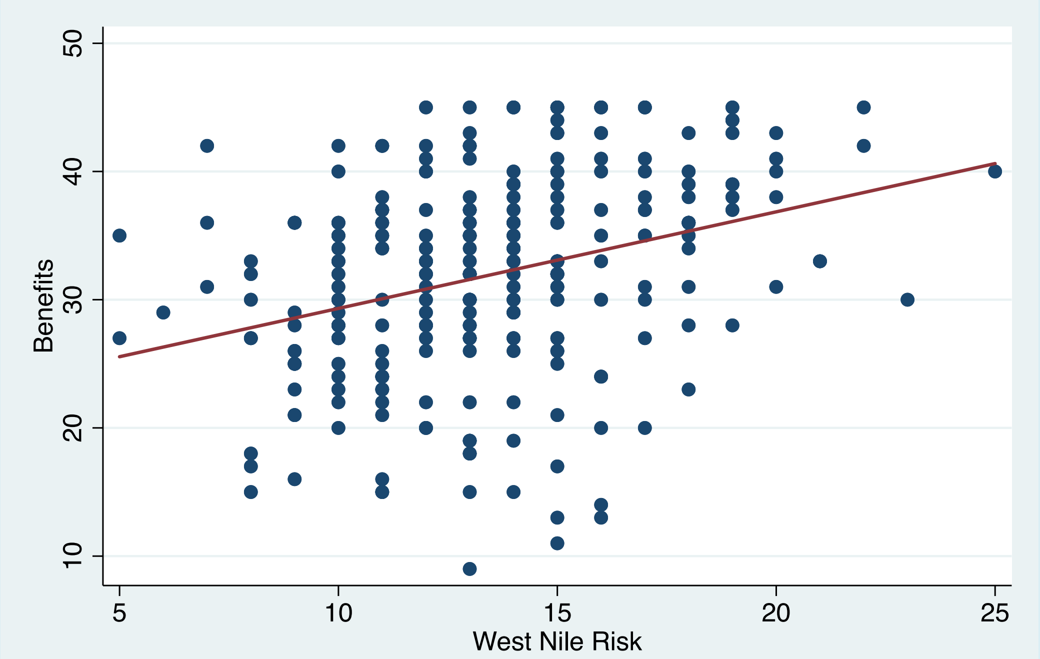
It appears that there is a trend that as risk perception increases so do perceived benefits.

Simple Regression

This can be called a Linear Relationship.

A relationship between two interval-ratio variables in which the observations displayed in a scatter diagram can be approximated with a straight line.





This line has the equation $Y = a + b(X)$

Simple Regression

Let’s explore this further using a much smaller dataset.

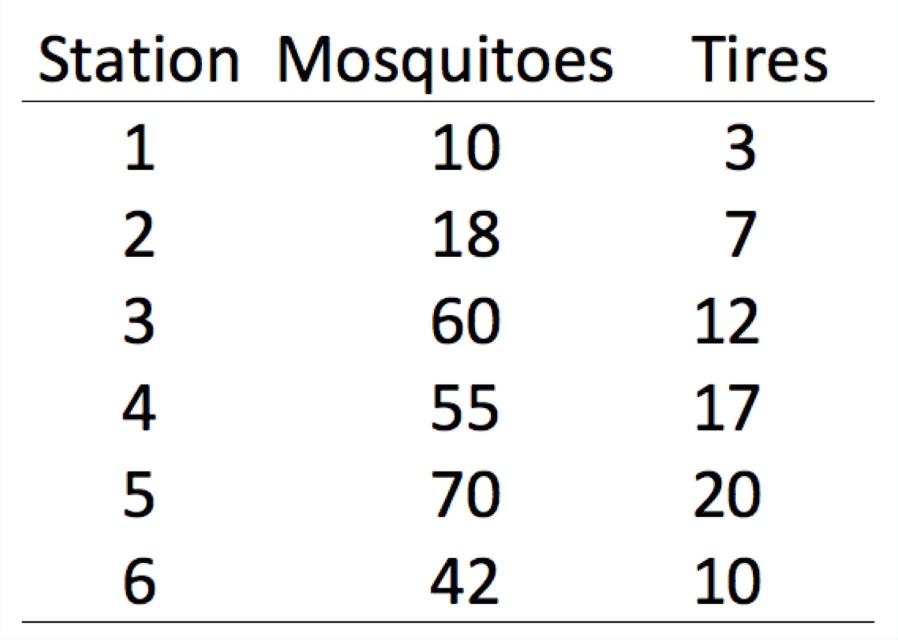
It is known that the presence of abandoned tires increases the breeding area for mosquitoes. Scientists in the study area want to see if abandoned tires also favor the presence of one mosquito species, Culex. This is the one that transmits West Nile.

Y axis, **Dependent variable:** Percentage of mosquitoes that are Culex (as opposed to other species) caught in six monitoring stations.

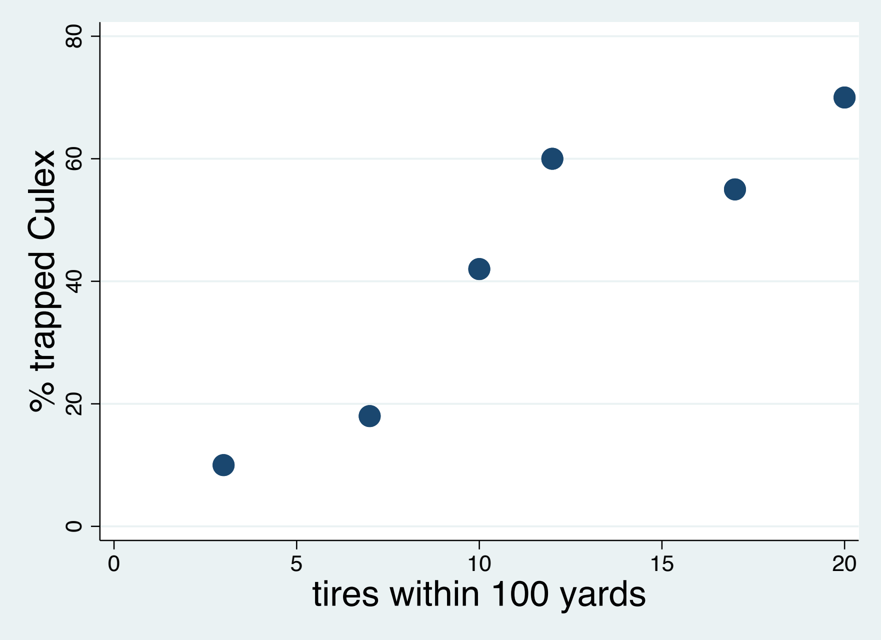
X axis, **Independent variable:** Number of abandoned tires found within 100 yards of each monitoring station.

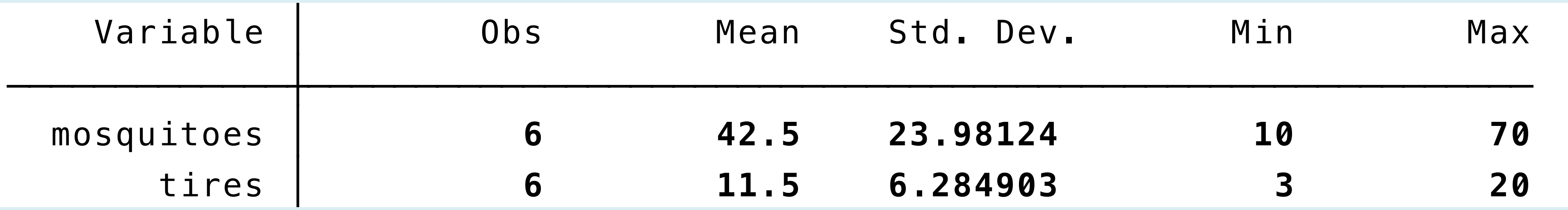
Simple Regression

Our data



Station	Mosquitoes	Tires
1	10	3
2	18	7
3	60	12
4	55	17
5	70	20
6	42	10





```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  mosquitoes |          6        42.5    23.98124         10         70
       tires |          6        11.5    6.284903          3         20
```

Simple Regression

Finding the best line:
Least Squares Method

The technique that produces the best possible line. Our equation will allow us to evaluate the linear relationship and make a prediction of the Culex percentage based on the presence of tires.

Simple Regression

Finding the best line:

Least Squares Method

$$\widehat{Y} = a + b(X)$$

$$b = \frac{S_{XY}}{S_x^2} \quad \textit{where} \quad S_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

and

$$a = \bar{Y} - b\bar{X}$$

Simple Regression

Least Squares Method

$$S_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

1. Calculate the covariance S_{xy} 698.5 ÷ 5 = 139.7

X(tires)	Y(mosquitoes)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
3	10	-8.5	-32.5	276.3
7	18	-4.5	-24.5	110.3
12	60	0.5	17.5	8.8
17	55	5.5	12.5	68.8
20	70	8.5	27.5	233.8
10	42	-1.5	-0.5	0.8
M = 11.5	M = 42.5			Σ = 698.5

Simple Regression

Least Squares Method

$$b = \frac{S_{XY}}{S_x^2}$$

2. Calculate the slope *b*

S_{XY} = 139.7

S_x given = 6.28

$S_x{}^2$ = 39.5

139.7 ÷ 39.5 = 3.5

Simple Regression

Least Squares Method

$$a = \bar{Y} - b\bar{X}$$

3. Calculate the intercept α

$b =$ 3.5

Mean X given = 11.5 42.5 – (3.5 * 11.5) = 2.25

Mean Y given = 42.5

Simple Regression

Least Squares Method

$$\widehat{Y} = a + b(X)$$

4. State the linear equation

$b = 3.5$

$a = 2.25$

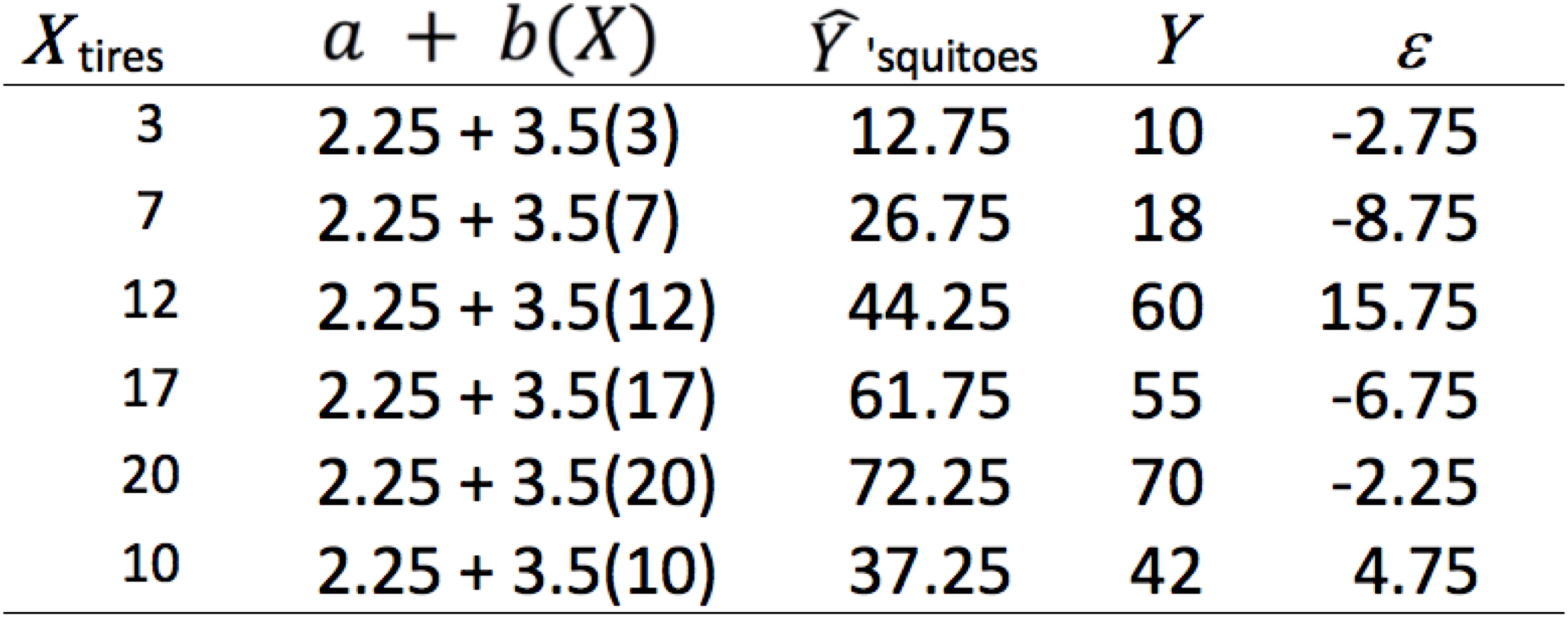
$\widehat{Y} = 2.25 + 3.5(X)$

Simple Regression

Least Squares Method

$$\widehat{Y} = 2.25 + 3.5(X)$$

5. Calculate the predicted values of Y

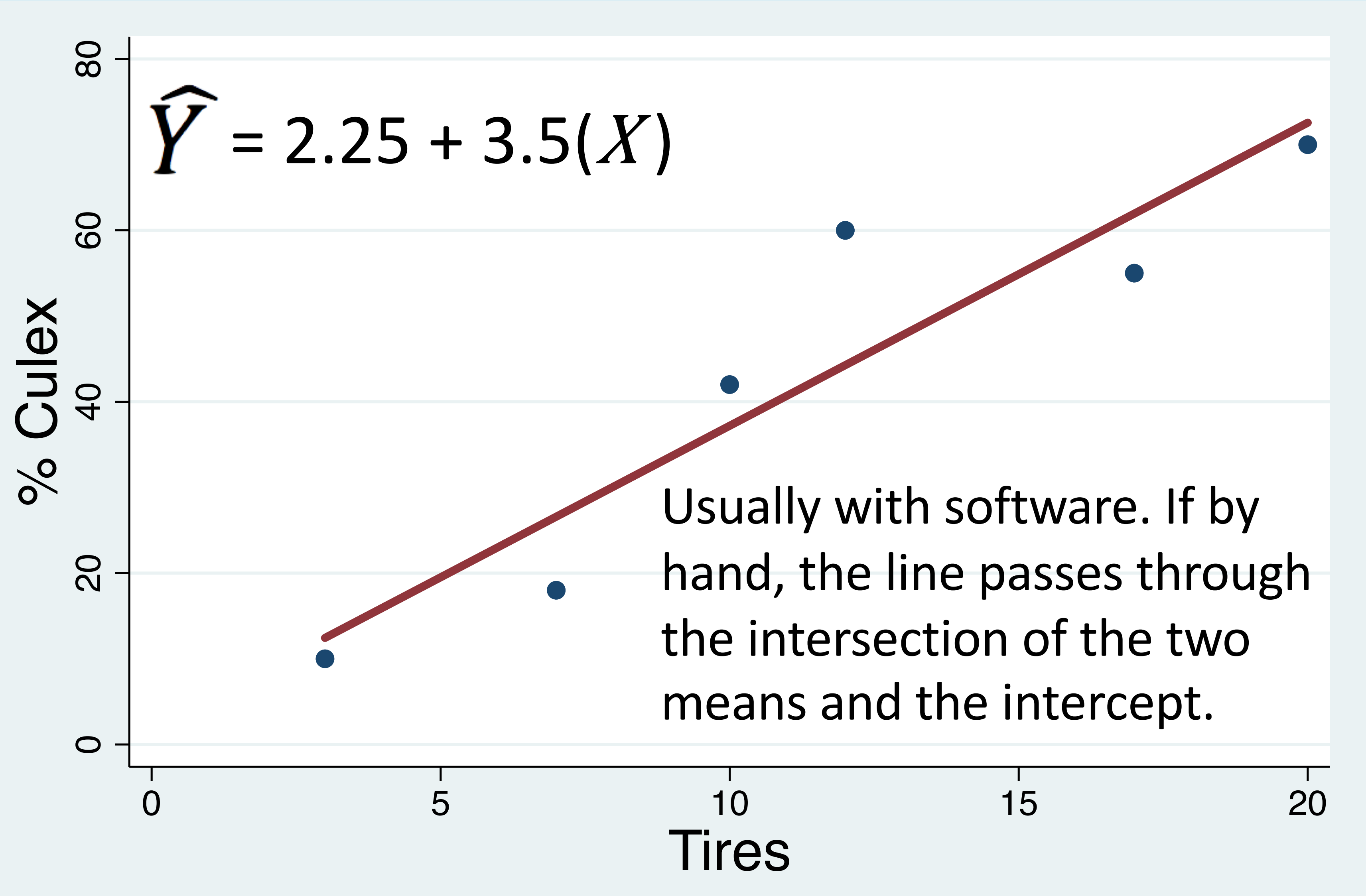


X_{tires}	$a + b(X)$	$\widehat{Y}_{\text{'squitoes}}$	Y	ε
3	2.25 + 3.5(3)	12.75	10	-2.75
7	2.25 + 3.5(7)	26.75	18	-8.75
12	2.25 + 3.5(12)	44.25	60	15.75
17	2.25 + 3.5(17)	61.75	55	-6.75
20	2.25 + 3.5(20)	72.25	70	-2.25
10	2.25 + 3.5(10)	37.25	42	4.75

Simple Regression

Least Squares Method

6. Plot data with regression line



$$\hat{Y} = 2.25 + 3.5(X)$$

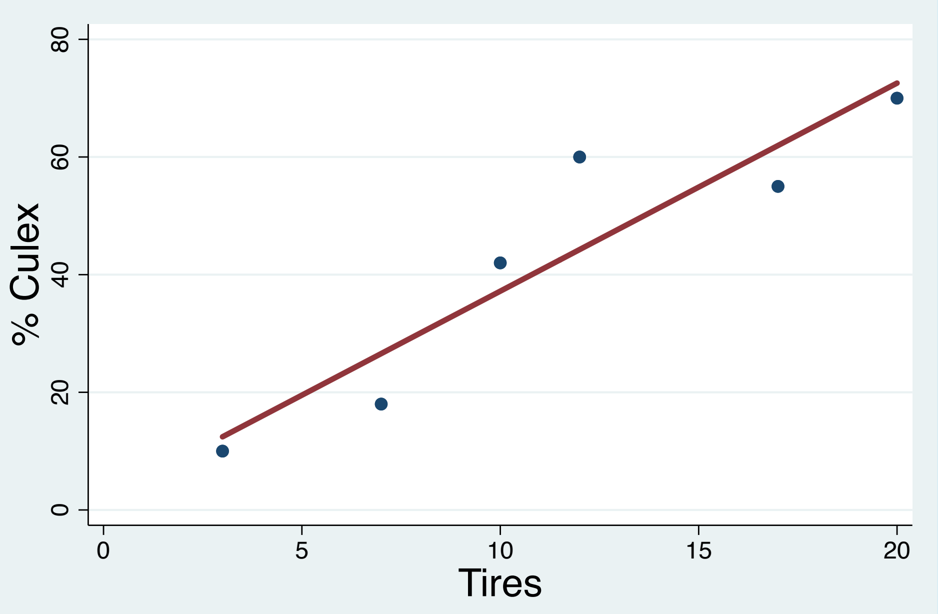
Usually with software. If by hand, the line passes through the intersection of the two means and the intercept.

Simple Regression

Least Squares Method

7. Interpret the slope *b*

$$\hat{Y} = 2.25 + 3.5(X)$$



For each additional tire the should be, on average, a 3.5% increase in the concentration of Culex mosquitoes found.

Apply Prediction:

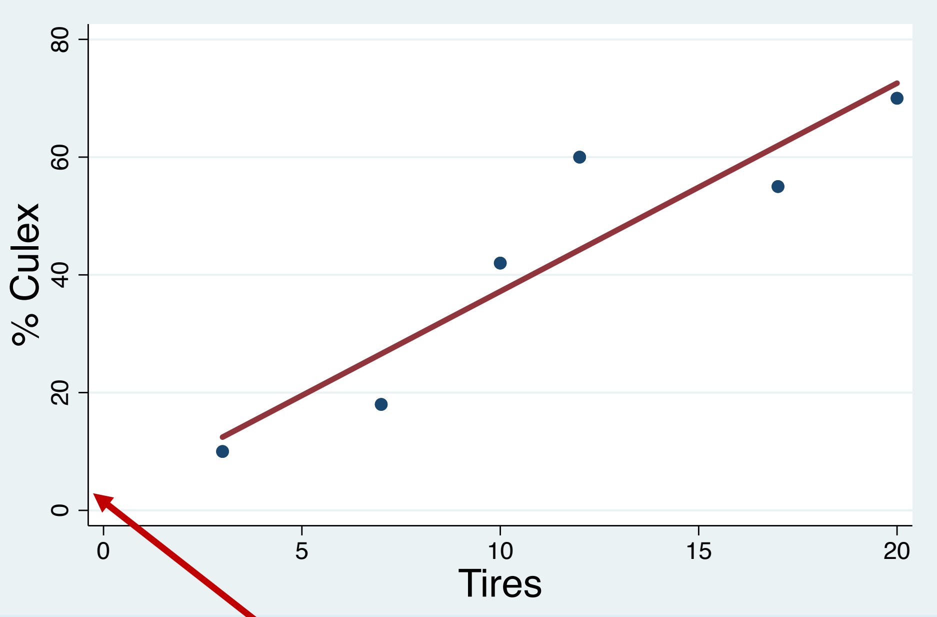
If there are 15 tires expect about 55% Culex.

2.25 + 3.5(15)

Simple Regression

Least Squares Method

8. Interpret the intercept *a* $\widehat{Y} = 2.25 + 3.5(X)$



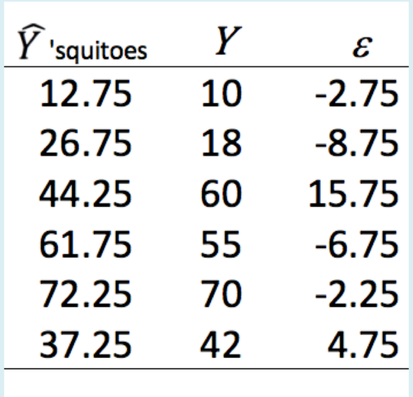
This is not typically done because the interpretation can be a value that is not possible. However: the intercept is the predicted value for Y when X = 0.

So, if there are zero tires, we would still expect 2.25% Culex.

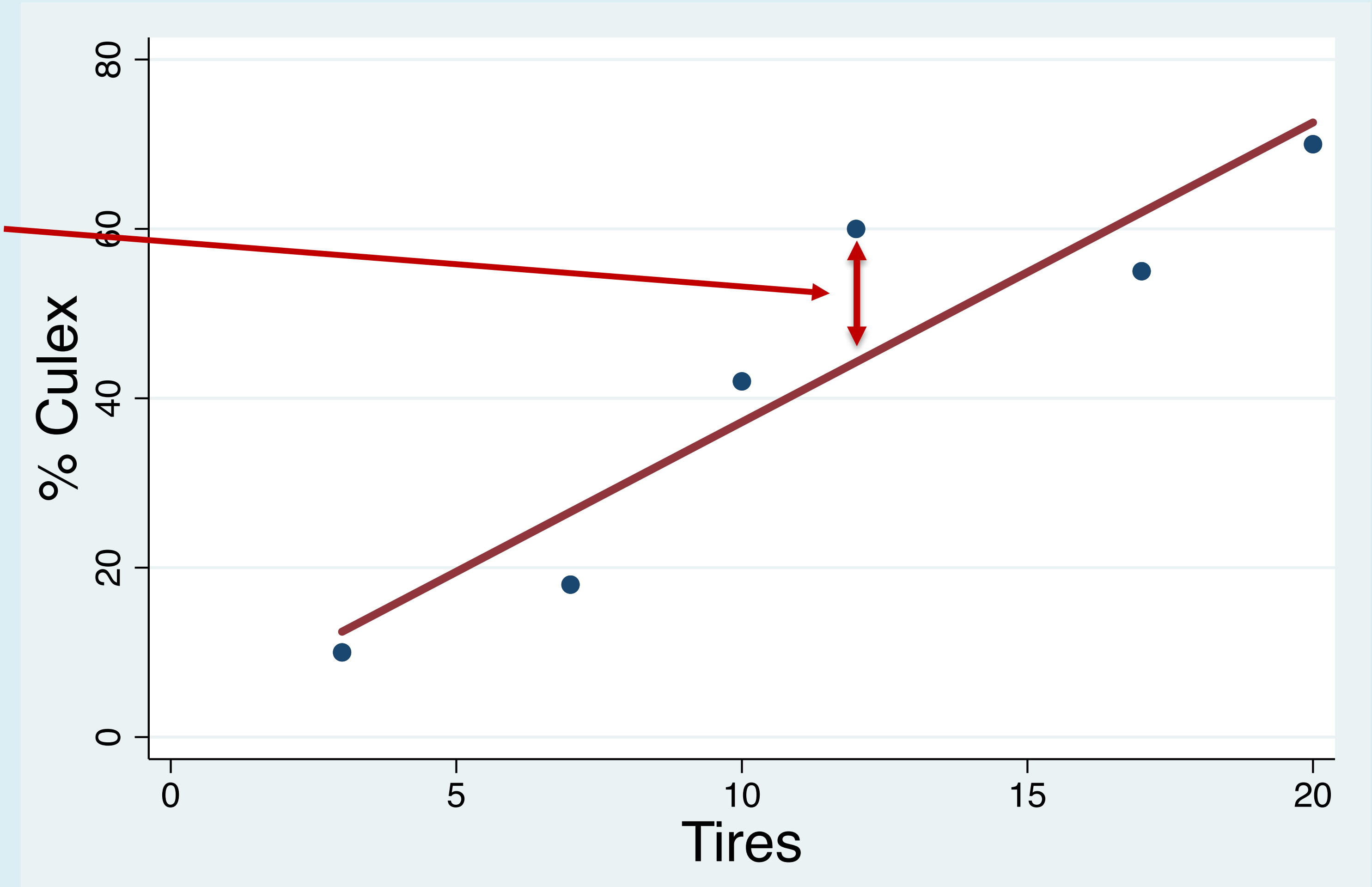
Simple Regression

Least Squares Method

9. Observe the error



$\hat{Y}$ 'squitoes	Y	ε
12.75	10	-2.75
26.75	18	-8.75
44.25	60	15.75
61.75	55	-6.75
72.25	70	-2.25
37.25	42	4.75



Recall this is *Estimation,* so there is error. Our prediction for 12 tires was quite a bit lower than what was observed.

Simple Regression

Least Squares Method

10. Assess the strength and direction of the association

$$r = \frac{s_{XY}}{s_X s_Y}$$

We already have all of these values. We've calculated the covariance and the standard deviations are provided in the summary, so …

r = 139.7 ÷ (7.8 * 3.3) = .97

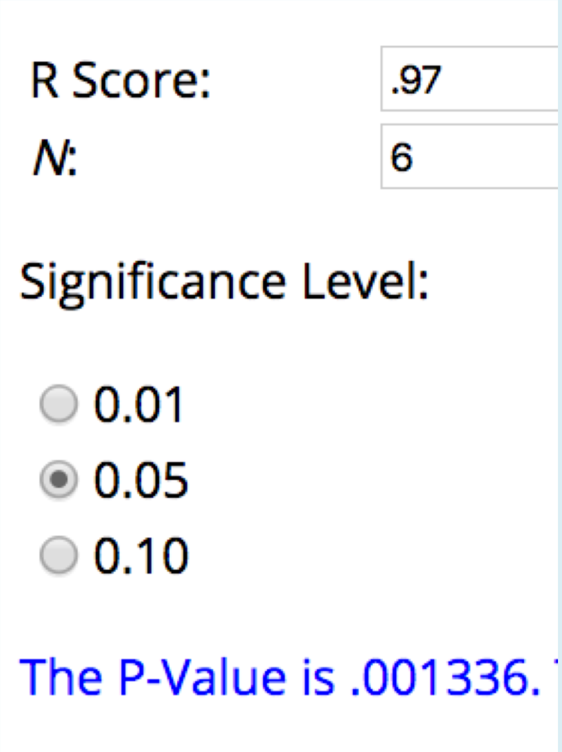
Simple Regression

Least Squares Method

11. Assess the significance of the association

$$r = 139.7 \div (7.8 * 3.3) = .97$$

All that's needed to get a p-value for r is the value of r and the N. Then use a table or software.



R Score: .97

N: 6

Significance Level:

○ 0.01
● 0.05
○ 0.10

The P-Value is .001336.

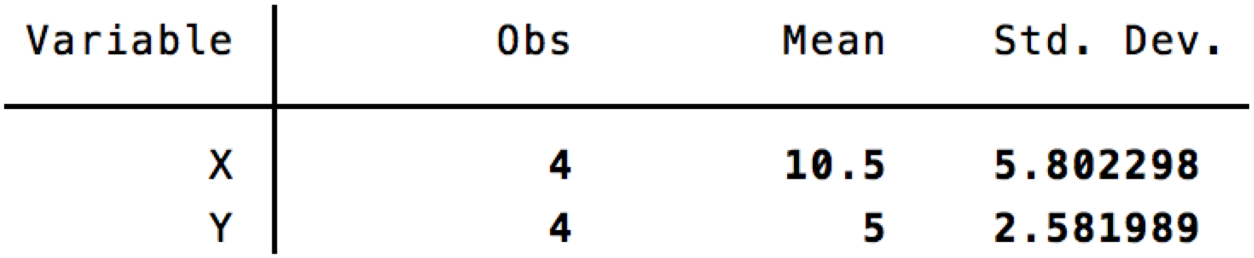
https://www.socscistatistics.com/pvalues/pearsondistribution.aspx

Simple Regression

Another example, all parts:

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
5	2	-5.5	-3	16.5
7	4	-3.5	-1	3.5
12	6	1.5	1	1.5
18	8	7.5	3	22.5

$\Sigma =$ 44.0



```
Variable |       Obs        Mean    Std. Dev.
---------+------------------------------------
       X |         4        10.5    5.802298
       Y |         4           5    2.581989
```

$$S_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

44 ÷ 3 = 14.7

$$b = \frac{S_{XY}}{S_x^2}$$ **14.7 ÷ 33.6 = .44**

$$a = \bar{Y} - b\bar{X}$$ **5 - .44(10.5) = .38**

$$\hat{Y} = a + b(X)$$ **.44 + .38(X)**

$$r = \frac{S_{XY}}{S_X S_Y}$$ **14.7 ÷ (5.8 * 2.6) = .97**

Simple Regression

Least Squares Method

Interpreting r

r is a standardized coefficient that ranges from -1 to 1

For r = -1 or r = +1 all of the observations fall exactly on the regression line. This is called a deterministic relationship.

For the case r = 0 the is no relationship.

The hypothesis test for r is H_o: r = 0. Given the sample size, we ask if the size of r is sufficiently different from zero as to have not occurred by chance.

Simple Regression

Least Squares Method

Interpreting r

If we square the value of r then we can express it as the percentage of “overlap” or “common variance” between X and Y.

r = .97 r^2 = .94 or 94% common variance

Simple Regression

Least Squares Method

Interpreting r

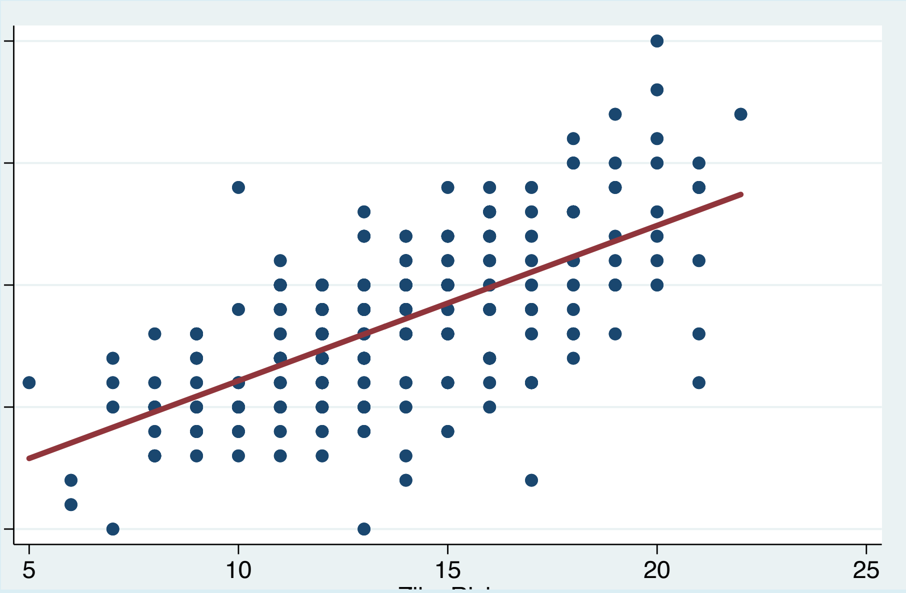
When r is positive we have a positive association. As the value of X increases so does the value of Y.

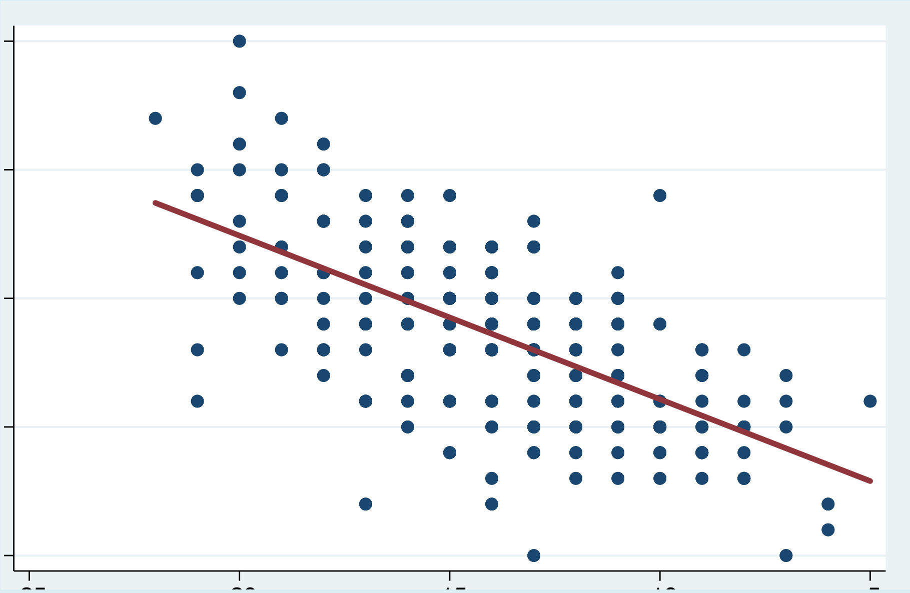
When r is negative we have a negative, or inverse association. As the value of X increases the value of Y decreases.

Simple Regression

Least Squares Method

Interpreting r





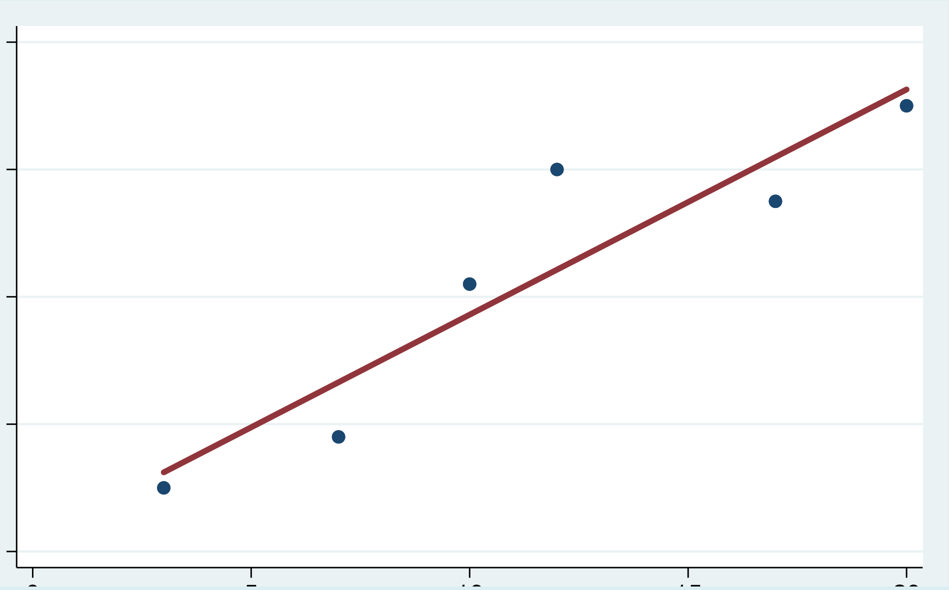
Positive Correlation

Negative Correlation

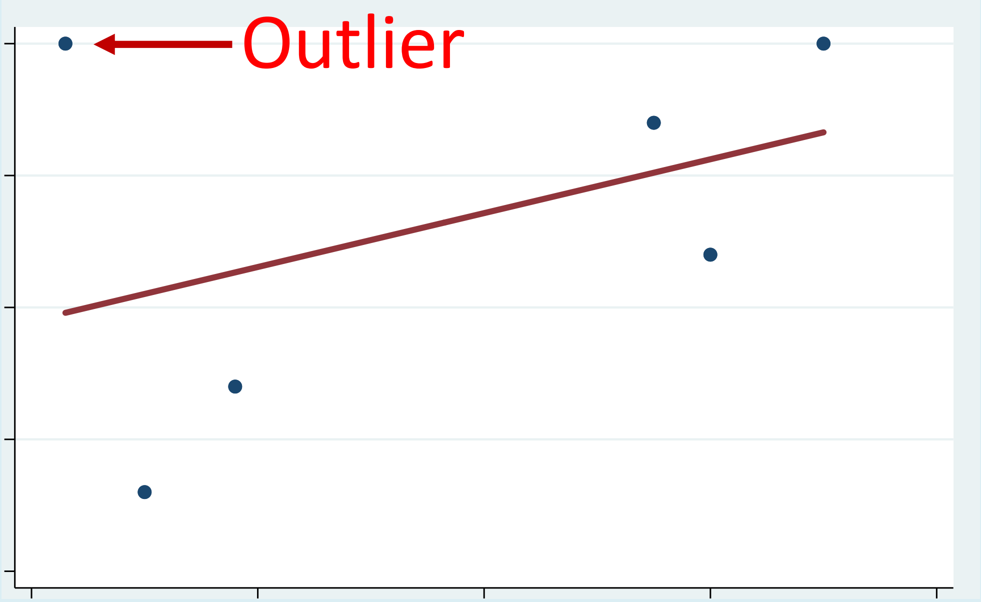
Simple Regression

Least Squares Method

Outliers are values that fall well beyond the overall range of the data. They can have a strong effect on the slope. This is one good reason to examine scatterplots.

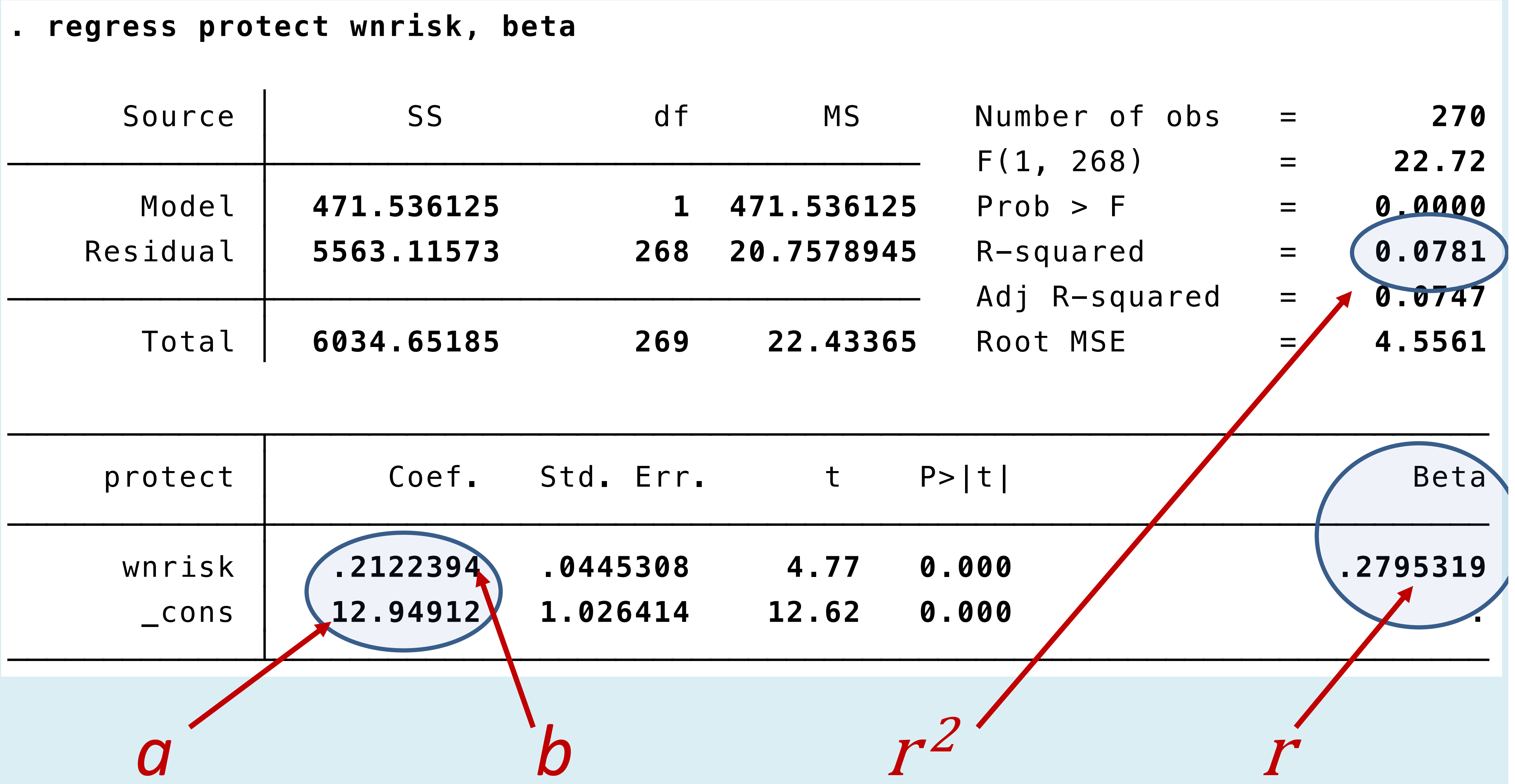


$r = .93$



$r = .42$

Regression in Stata

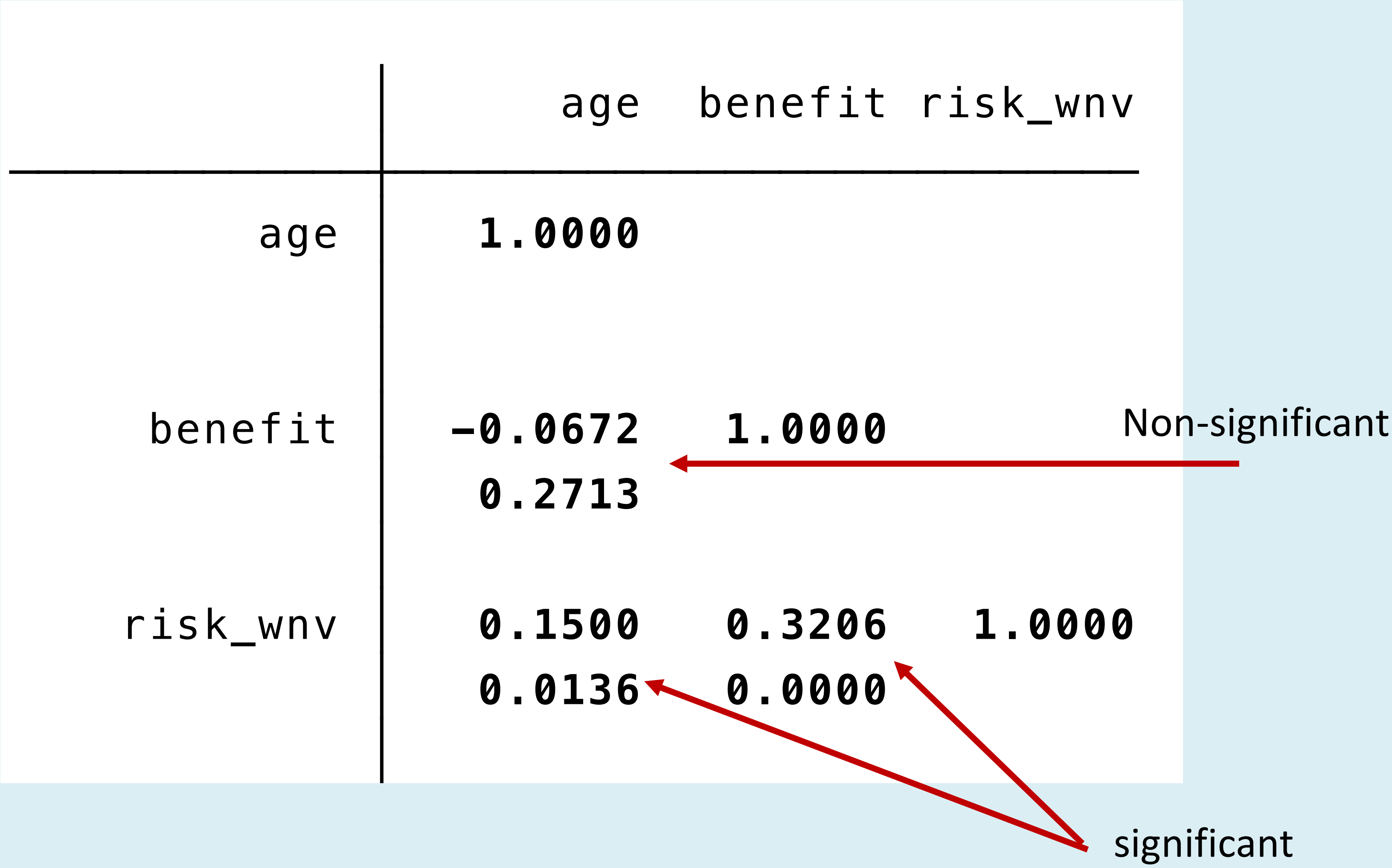


```
. regress protect wnrisk, beta

      Source |       SS           df       MS      Number of obs   =       270
-------------+----------------------------------   F(1, 268)       =     22.72
       Model |  471.536125         1  471.536125   Prob > F        =    0.0000
    Residual |  5563.11573       268  20.7578945   R-squared       =    0.0781
-------------+----------------------------------   Adj R-squared   =    0.0747
       Total |  6034.65185       269    22.43365   Root MSE        =    4.5561

------------------------------------------------------------------------------
     protect |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      wnrisk |   .2122394   .0445308     4.77   0.000                 .2795319
       _cons |   12.94912   1.026414    12.62   0.000                        .
------------------------------------------------------------------------------
```

Correlation Matrix in Stata

Top number: *r*

Bottom number: *p-value*



```
             |      age  benefit risk_wnv
-------------+---------------------------
         age |   1.0000
             |
             |
     benefit |  -0.0672   1.0000
             |   0.2713
             |
    risk_wnv |   0.1500   0.3206   1.0000
             |   0.0136   0.0000
             |
```

Non-significant

significant

Key Terms

Pearson’s correlation coefficient (r)
Coefficient of determination (r2)
Scatter diagram (scatterplot)
Linear relationship
Deterministic relationship
Null hypothesis for r
Linear equation
Covariance
Regression line
Least squares line (best-fitting line)
Slope (b)
Y-intercept (a)
Positive and Negative association
Outliers
Predicted value (Y-hat)
Error

Important Note on Text:

The significance of a regression can also be tested using ANOVA, as described on pages 441-443. You’ll see this in Stata output as well. While it’s the standard approach because it can be extended to more complicated designs, we’re just using the simpler approach described here. So skip those pages

Likewise, the text starts a very basic but not very complete introduction to the next topic, Multiple Regression (pp. 450-453). We’ll also not cover this topic.

JTC270 Lab 9 Regression and Correlation. $\widehat{Y} = a + b(X)$

$$S_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1} \qquad b = \frac{S_{XY}}{S_x^2} \qquad a = \bar{Y} - b\bar{X} \qquad r = \frac{S_{XY}}{S_X S_Y}$$

```
    Variable |        Obs        Mean    Std. Dev.
-------------+------------------------------------
       wells |          5          12     5.87367
       mines |          5         5.6    3.435113
```

X(wells)	Y(mines)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
5	1			
8	3			
12	7			
15	9			
20	8			

$\Sigma =$

https://www.socscistatistics.com/pvalues/pearsondistribution.aspx

During the controversy over the mining operation concern was raised about the possibility that locations that were most favorable for water wells might also be most favorable for mine injection locations. This would be an important association to investigate since a positive finding would support the argument that risks to groundwater could be magnified if the mining operation expanded or intensified.

An investigation was done focusing on the rural area of the study. A random selection of five two square mile parcels was made. The number of water wells in each parcel was recorded. The number of mine injection locations proposed by the mining company was obtained as well. The data are presented above, with a blank table that might help organize your calculations.

Q1. What is the value for the covariance? ____

Q2. What is the value for the slope *b* ? ____

Q3. What is the value for the intercept, *a* ? ____

Q4. What is the value of *r* ? _____

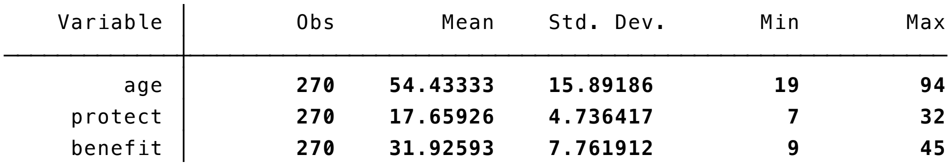
Q5. What is the p-value for r *?* _____

Q6. What is the conclusion of the test?

a) non-significant finding, no association
b) significant finding, positive association
c) significant finding, negative association
d) significant finding, null association

Q7. How many mine sites would you predict there to be in an area with 25 wells? _______

Turning to our lecture dataset about West Nile, we'll look at three variables: the participant's age, the score for protective actions against mosquitoes, and the score for perception of benefits from taking protective actions. The summary statistics are:



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |        270    54.43333    15.89186         19         94
     protect |        270    17.65926    4.736417          7         32
     benefit |        270    31.92593    7.761912          9         45
```

First, we want to see what correlations might be present. We can do this directly without regression.

Statistics --> Summary tables and tests --> Summary and descriptive statistics --> pairwise correlations
enter variables: age protect benefit check: Print significance level for each entry submit

Examine the output

Q8. Of the set of correlations, which are significant?

a) age with protect
b) age with benefit
c) protect with benefit
d) all are significant

We have a theory that believing that protective behaviors are beneficial may lead to taking protective actions. Thus, we might explore that relationship in a regression with protect at the dependent variable and benefit as the independent variable.

Statistics --> Linear models and related --> Linear regression
dependent variable = protect Independent variables = benefit
Reporting tab: check Standardized beta coefficients submit

Ignore the top part of the output, just examine the bottom section that reports the regression.

Q9. Which of these is the regression equation?

a) .03 + 11.5(X)
b) .19 - 11.5(Y)
c. 11.5 + .19(X)
d) .31 + 1.16(X)

Q10. In the summary statistics we see that the highest score on benefit is 45. What would we predict the individual's score to be on protect at that level? _____

Happy eating: Ingredient to a long life, in your cereal bowl

By Jen Christensen, CNN

Updated 10:22 AM ET, Fri March 27, 2015

SHUTTERSTOCK

High-fiber cereal may help you keep diabetes, cancer, obesity, heart disease and early death at bay.

Story highlights

- An observational study found people who ate at least 10.22 grams of cereal fiber had a 19% lower risk of death
- People who ate at least 34g of whole wheat and fiber reduced risk of death by 17%
- Early studies have shown eating a diet rich in cereal fiber can help reduce risk for diabetes, cancer, inflammation and obesity

(CNN)—There's good news for those of you who wake up to a bowl of cereal every morning, especially if your go-to choice is high in fiber. Hidden in that favorite box may be a prize better than a plastic toy. It just might hold the key to a longer life, according to a new study.

No, this study wasn't done by Snap, Crackle or Pop. Tony the Tiger was not involved in the making of this research.

How to choose a healthy breakfast cereal

Some scientists at the Harvard School of Public Health have been researching the impact of cereal fiber on diet for years. They found that people who reported in

Focus on Health

13 ways to stop drinking soda

Cereal could be the key to long life

Sleep better with six minutes of yoga before bed

Instilling empathy among doctors pays off

Can't fully expect when expecting? Accurate gender

10 nutritionists share the McDonald's meals they'd order

surveys a diet rich in cereal fiber lived longer than those who chose less well in the morning. They had a 19% reduced risk of death, compared to those who ate the least amount of cereal fiber.

Crunching the numbers even further, the authors found that high fiber cereal eaters had a 34% lower risk of death from diabetes and a 15% reduced risk of death from cancer. People who ate a lot of whole grains and dietary fiber had a 17% lower risk of all-cause mortality.

Cereal fiber, they conclude, is one "potentially protective component" of a really healthy, premature death-preventing diet.

The study was published in the latest issue of BMC Medicine.

It drew from the NIH-AARP Diet and Health Study and included more than 566,000 AARP members ageds 50 to 71 from six states and two large cities. It excluded individuals who reported extreme-energy intake, which is common, since scientists believe these survey takers are not totally accurate in what they report. That left them with over 367,000 people.

This new study builds on others that have shown that cereal fiber and whole grains have a positive impact on your life if you want to avoid cancers, inflammation and obesity as well.

Does that mean that eating a daily bowl of your favorite purple horseshoe marshmallow-sprinkled cereal is doing your body right?

Well, don't court the leprechaun quite yet, dietitians say. Those cereals have sugar among their top ingredients, so dietitians suggest you avoid those.

If you want to get the daily serving that these researchers say showed a difference in risk reduction, you need to eat at least 10.22 grams of cereal fiber per day based on a 1,000 kcal daily diet.

If you want to get your fill with just one serving of cereal, aim for those that have "fiber" in the title or list at least 10 grams of fiber per serving on the label. Fiber One lists 14g per serving. Kashi GoLean lists 10g. Mini-Wheats lists 8g.

If you can't stand the taste of high-fiber cereals, don't worry. Other popular cereals such as Cheerios have about 3g of fiber per serving, as do Honey Bunches of Oats. Oatmeal is a good source of fiber too.

Whole grains and regular dietary fiber also may help reduce your mortality risks, the study found, and those can be found in a large number of products.

Consider oatmeal or a non-high fiber cereal (3 to 5 grams), eat a piece of wheat bread (about 5 grams) or a whole wheat tortilla (about 5 grams). Black beans are a rich source for dietary fiber (19.5 grams). And add even some fruit like apples (a large one has 4.5 grams) or a half a cup of blackberries (4.4 grams) all of that would add up to this 'higher' total that may improve your odds of living a longer life.

And if you don't want to spend hours reading labels at the grocery store, dietitian Lori Zanini said she tells her clients this fiber rule of thumb: "No animal product will naturally have this," Zanini said. "Plants are where you should go to find fiber. It only comes from the cell walls of plants."

Most Americans, she said probably don't get the amount of cereal fiber or whole grains recommended as advantageous in this study. "But once you consciously seek it out, it does become easier," Zanini said "And with the wide variety of ways you can get fiber into your diet it isn't hard, especially if you know it may help your health."

The operative word is "may," study author Dr. Lu Qi said. Keep in mind the study looks at connections; it doesn't show causality. To definitively show cereal is the key to long life, the

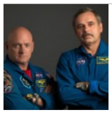
More Top Stories

Co-pilot had medical note for day of crash, hid illness

Saudis won't rule out nukes

Harry Reid, Senate minority leader, to retire

'Family Matters' star arrested


Astronauts leaving for a year in space

Apple CEO to give away his fortune

Eat cereal and live a long time

More from CNN

Black man found hanging from a tree in Mississippi

This video is no longer available

Woman shot in the head in rush hour road rage incident

A relic of the 'True Cross' is carbon dated

Germanwings crash: A pilot's take (Opinion)

Islamic ring in Viking grave sheds new light

professor at the Harvard medical school and Harvard's school of public health said, you'd need a clinical trial that would look at this specific issue.

That said, Qi personally is a believer in the breakfast food. He said he eats cereal regularly to start his day. Harvard even provides breakfast for free to the faculty. And if it's good enough for doctors at Harvard, they may just be on to something.

More from CNN

Did Jesus reject his family? ▶

Panel: Jews are latest group King has offended ▶

Mom's 911 call leads to tragedy ▶

Why people on this island live past 100 ▶

Why 'SNL' ISIS skit was brilliant

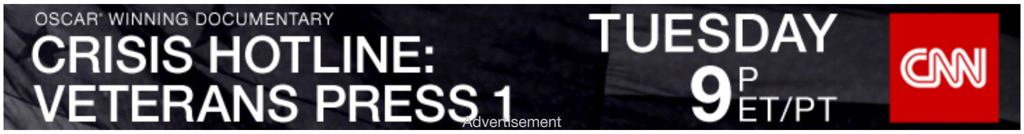
Putin: Russia was ready for nuclear alert over Crimea

More from Health

Stop drinking soda, for (your own) good

Fit Nation: Change up your living space to get in shape

Indiana governor declares public health emergency due to HIV epidemic



Huang *et al. BMC Medicine* (2015) 13:59
DOI 10.1186/s12916-015-0294-7

BMC Medicine

RESEARCH ARTICLE

Open Access

Consumption of whole grains and cereal fiber and total and cause-specific mortality: prospective analysis of 367,442 individuals

Tao Huang[1], Min Xu[1], Albert Lee[2], Susan Cho[3] and Lu Qi[1,4*]

Abstract

Background: Intakes of whole grains and cereal fiber have been inversely associated with the risk of chronic diseases; however, their relation with total and disease-specific mortality remain unclear. We aimed to prospectively assess the association of whole grains and cereal fiber intake with all causes and cause-specific mortality.

Methods: The study included 367,442 participants from the prospective NIH-AARP Diet and Health Study (enrolled in 1995 and followed through 2009). Participants with cancer, heart disease, stroke, diabetes, and self-reported end-stage renal disease at baseline were excluded.

Results: Over an average of 14 years of follow-up, a total of 46,067 deaths were documented. Consumption of whole grains were inversely associated with risk of all-cause mortality and death from cancer, cardiovascular disease (CVD), diabetes, respiratory disease, infections, and other causes. In multivariable models, as compared with individuals with the lowest intakes, those in the highest intake of whole grains had a 17% (95% CI, 14–19%) lower risk of all-cause mortality and 11–48% lower risk of disease-specific mortality (all *P* for trend <0.023); those in the highest intake of cereal fiber had a 19% (95% CI, 16–21%) lower risk of all-cause mortality and 15–34% lower risk of disease-specific mortality (all *P* for trend <0.005). When cereal fiber was further adjusted, the associations of whole grains with death from CVD, respiratory disease and infections became not significant; the associations with all-cause mortality and death from cancer and diabetes were attenuated but remained significant (*P* for trend <0.029).

Conclusions: Consumption of whole grains and cereal fiber was inversely associated with reduced total and cause-specific mortality. Our data suggest cereal fiber is one potentially protective component.

Keywords: Cereal fiber, Mortality, Whole grains

Background

Grains, also called cereals, are the seeds of plants cultivated for food. When whole, they include the germ, bran, and endosperm [1]. Most whole grains are abundant sources of dietary fiber and other nutrients, such as minerals and antioxidants, which have shown beneficial effects on human health including improvement of weight loss, insulin sensitivity, and lipid profile, as well as inhibition of systemic inflammation [2-4].

In epidemiology studies, evidence is accumulating indicating that consumption of whole grain products or their effective components, especially dietary fiber found in the grain, i.e., cereal fiber, may reduce the risk of chronic disease. Several recent meta-analyses taking into account a large number of subjects and prospective studies showed significant and consistent protective effects of high intake of whole grains and cereal fiber on type 2 diabetes [5], cardiovascular disease (CVD) [6], and certain cancers (e.g., colorectal cancer) [7]. In our earlier analysis in the Nurses' Health Study, we observed potential protective effects of whole grains on total or cardiovascular mortality in diabetic women [8]. Although the National Institutes of Health (NIH)-AARP Diet and Health Study previously reported that dietary

* Correspondence: nhlqi@channing.harvard.edu
[1] Department of Nutrition, Harvard School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA
[4] Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 75 Francis St, Boston, MA 02115, USA
Full list of author information is available at the end of the article

fiber intake was inversely associated with the risk of total death and death from CVD, infectious diseases, and respiratory diseases [9], few studies have prospectively examined the associations of whole grains and its components, such as cereal fiber, with total or disease-specific mortality.

In the present study, we used data from 367,442 people who were at risk for a total of 12.3 million person-years. We aimed to provide reliable estimates of independent associations between baseline whole grains and cereal fiber intake and the risk of total or cause-specific death from CVD, cancers, diabetes, and other diseases.

Methods

Study population

The NIH-AARP Diet and Health Study included 566,399 AARP members aged 50 to 71 from six US states (California, Florida, Louisiana, New Jersey, North Carolina, and Pennsylvania) and two metropolitan areas (Atlanta, Georgia, and Detroit, Michigan) [10]. Participants responded to a questionnaire mailed in October 1995 and December 1997. Details of the NIH-AARP Study have been previously described [11]. Among participants who returned the questionnaires with satisfactory dietary data, we excluded individuals who indicated that they were proxies for the intended respondent (n = 15,760) as well as those who had cancer (n = 50,591), heart disease (n = 80,254), stroke (n = 12,812), diabetes (n = 52,647), or self-reported end-stage renal disease at baseline (n = 1,299). We also excluded those who reported extreme consumption (>2 times the interquartile ranges of Box-Cox transformed intake) of total energy (n = 3,771) and dietary fiber (n = 3,324). Exclusion of individuals reporting extreme energy intake is widely used in nutritional epidemiology studies since these participants are more likely to over- or under-report their intake [12]. After exclusions (n = 198,957), the analytic cohort included 367,442 individuals. The NIH-AARP Diet and Health study was approved by the Special Studies Institutional Review Board of the US National Cancer Institute. All participants provided written informed consent.

Assessment of dietary exposures

At baseline, dietary intake was assessed with a self-administered 124-item food frequency questionnaire (FFQ), which was an early version of the Diet History Questionnaire developed at the National Cancer Institute [13,14]. Participants were asked to report their usual frequency of intake and portion size over the past 12 months using 10 predefined frequency categories ranging from never to 6+ times per day for beverages and from never to 2+ times per day for solid foods with three portion size categories. The food items, portion sizes, and nutrient database were constructed using the US Department of Agriculture's 1994–1996 Continuing Survey of Food Intakes by Individuals. The FFQ used in the study was calibrated using two non-consecutive 24-hour dietary recalls in 1953 NIH-AARP study participants. The nutrient database for dietary fiber was based on AOAC methods.

The whole grains were defined as the whole grain part of each product. The US Department of Agriculutre's Pyramid Servings Database enabled us to accurately estimate whole-grain intake from all foods in the FFQ. The sources of whole-grain intake in the FFQ used in our study were ready-to-eat cereals, high-fiber cereals, other fiber cereals, whole-grain breads or dinner rolls, cooked cereal, popcorn, pancakes, waffles, French toast or crepes, rice or other cooked grains, bagels, English muffins, tortillas, pasta, crackers, chips, cookies or brownies, sweet pastries, and pies. In this Continuing Survey of Food Intakes by Individuals dataset., whole grain foods were defined as those containing at least 25% whole grains and/or bran. Main fibers are from fruit, grains, vegetables, and beans in the present study. Cereal fiber was defined as fiber from all cereals (e.g., ready-to-eat cereals, high-fiber cereals, cooked cereal, and other fiber cereals) and grain-based products.

In specifying numbers of deaths by intake quintile, the number of deaths is determined by the energy-adjusted intake quintile for the entire population; when deaths are specific to a sex, we used the quintiles within the sex. We also collected demographic, anthropometric, and lifestyle information, including history of smoking (the number of cigarettes smoked per day), time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline), physical activity (never, rare, 1 to 2, 3 to 4, ≥5 hours/week), alcohol intake (g/day) family history of cancers, menopausal hormone therapy use in women, and some medical conditions at baseline.

Ascertaining mortality

The AARP dataset denotes date of death and cause of death. There are 22 broad categories for cause of death. The modeling analysis for specific cause of death is performed with the study end date of 2008. For total mortality, models for study end dates in 2008 and 2009 were designed. Subjects with date of death after the study's end date are treated as alive at the end of the study, with no death or cause of death in the model. When the study end date was 2008, and there was a date of death but no cause of death for 2008 or earlier, the subject was not included in the modeling for cause of death, but only total mortality. Thus, when specifying numbers of deaths it depends on both the study end date and whether the cause of death field is missing. Vital status was determined through a periodic linkage of the cohort to the Social Security Administration Death Master File and follow-up searches of the National Death Index Plus for participants who matched the Social Security Administration Death Master File, cancer registry linkage, questionnaire responses, and responses to other mailings. The *International Classification of Diseases, Ninth Revision* [15] and the *International Statistical Classification of Diseases, 10th Revision* [16] were used to define death as follows: cancer (ICD-9, 140–239; ICD-10, C00–C97 and D00–D48), CVD (ICD-9, 390–398, 401–404, 410–429, and 440–448; ICD-10, I00–I13, I20–I51, and I70–I78), diabetes (ICD-9, 250; ICD-10, E10–E14), respiratory disease (ICD-9, 480–487 and 490–496; ICD-10, J10–J18 and J40–J47), infections (ICD-9, 001–139; ICD-10, A00–B99), and all other/unknown causes.

Statistical analysis

We used the Cox proportional hazards model to estimate hazard ratios (HRs) and two-sided 95% confidence intervals (CIs) using the SAS PROC PHREG procedure (Version 9.1; SAS Institute Inc., Cary, NC, USA). Person-years of follow-up were calculated from the date of the baseline questionnaire until the date of death or the end of follow-up (December 31, 2009), whichever occurred first. Intake of whole grains and cereal fiber were adjusted for total energy intake using the residual method [17], and were categorized into quintiles.

We presented the results from four analysis models. Model 1, adjusted for age and gender; Model 2, adjusted for age, gender, the number of cigarettes smoked per day, and time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline); Model 3, adjusted for age, gender, the number of cigarettes smoked per day, time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline), race or ethnicity group, alcohol intake, education level, marital status (yes, no), health status (poor, fair, good, very good), obesity (underweight, overweight, obesity), physical activity, consumption of red meat (processed and fresh meat), total fruit and total vegetables, total energy intake, and hormone usage; and Model 4, based on Model 3 further adjusted for cereal fiber (whole grains analysis).

For missing data in each covariate, we created indicator variables. Overall, missing data was less than 5%. The model results summary includes the results of statistical tests for trend in the response for the risk variable. Quintiles Trend *P* denotes the *P* value when the median value within the risk variable quintile is included in the hazard model as linear.

Results

Table 1 shows baseline characteristics of study participants (n = 367,442), according to intake of whole grains and cereal fiber. During an average of 14 years of follow-up (total person-years, 5,148,760), we documented 46,067 deaths, among them 11,283 from CVD, 19,043 from cancer, 371 from diabetes, 3,796 from respiratory disease, 922 from infection, and 5,223 from other causes. At baseline, intakes of whole grains and cereal fiber were inversely correlated with prevalence of overweight, obesity, and current smoking, as well as intake of red meat. The levels of moderate and vigorous physical activity were higher among participants with higher intakes of whole grains or cereal fiber than those with lower intakes.

Whole grains and cereal fiber intake with total mortality

In age- and gender-adjusted analysis (Model 1), we found that intake of whole grains were inversely associated with all-cause mortality (Table 2). As compared with the lowest quintile, the HRs across increasing quintiles of whole grain intake were 0.78 (95% CI, 0.76–0.80), 0.70 (95% CI, 0.68–0.72), 0.63 (95% CI, 0.61–0.65), and 0.61 (95% CI, 0.59–0.62) (*P* trend <0.0001). Further adjustment for smoking status and time since smoking cessation (Model 2) did not appreciably change the associations. When the models further included race/ethnicity, education, marital status, self-rated health status, obesity (underweight, overweight, and obesity), physical activity, use of menopausal hormone therapy, and intake of alcohol, red meat, fruits, vegetables, and total energy (Model 3), the highest quintile of whole grain intake was associated with 17% (95% CI, 14–19%) lower risk of all-cause mortality (*P* trend <0.0001). The associations between whole grain intake and all-cause mortality was attenuated, the highest quintile of whole grain intake was associated with 6% (95% CI, 3–10%) lower risk, but remained significant when cereal fiber was additionally adjusted (Model 4; *P* trend = 0.002). These results suggested that the protective effects of whole grain may be due, at least in the main part, to its cereal fiber component.

Similarly, we found that cereal fiber intake was significantly associated with all-cause mortality in age- and gender-adjusted and multivariate-adjusted models (Models 1, 2 and 3; all *P* trend <0.0001; Table 3). In model 3, the highest quintile of cereal intake was associated with 19% (16–21%) lower risk of all-cause mortality (*P* trend <0.0001).

Whole grains and cereal fiber intake with cause-specific mortality

We next tested the associations for cause-specific mortalities. In age- and gender-adjusted and multivariate adjusted models (Models 1, 2 and 3), intakes of whole grains or cereal fiber were inversely associated with risk of death from CVD, cancer, diabetes, respiratory disease, infections, and other/unknown causes (all *P* trend <0.023). In Model 3, as compared with the lowest quintiles, people in the highest quintile of whole grain intake had 11%

Table 1 Baseline characteristics of the study participants according to intake of whole grains and cereal fiber

	Total participants	Whole grains			Cereal fiber		
		Q1	Q3	Q5	Q1	Q3	Q5
n	367,442	73,488	73,489	73,489	73,488	73,489	73,489
Age, mean years	61.7	61.1	61.7	62.1	61.0	61.7	62.2
Female, %	43.9	30.7	51.9	40.4	35.0	49.9	40.2
Whites, %	92.9	92.6	92.9	93.2	91.2	93.0	94.4
College graduate, %	41.0	35.2	41.4	44.9	36.2	40.4	46.4
Married, %	68.1	72.7	65.2	68.9	69.8	66.1	69.9
Moderate physical activity (3–4 times/week), %	27.3	23.0	27.8	30.6	23.6	27.2	30.9
Vigorous physical activity (≥5 times/week), %	19.2	17.5	18.0	23.6	18.1	17.6	24.1
Overweight, %	42.4	44.5	42.0	41.0	43.9	42.2	41.1
Obesity, %	19.6	22.7	19.7	16.6	23.6	20.0	15.0
Very good or excellent self-report health, %	61.4	56.5	62.2	64.3	58.8	61.4	65.0
Previous or current use of postmenopausal hormone therapy, %	55.0	47.9	55.6	58.1	45.5	54.4	59.2
Former smoker, %	48.7	48.5	48.2	49.5	47.6	48.4	50.2
Current smoker, %	12.8	21.4	11.3	8.0	21.4	11.5	7.0
Median total energy intake (kcal/d)	1,805	2,394	1,527	1,855	2,330	1,563	1,832
Median alcohol intake (g/d)	14.7	32.5	10.0	8.8	27.5	11.6	9.9
Median servings of food							
Red meat (oz/d)	2.0	3.1	1.6	1.7	3.1	1.71	1.6
Fruits (cup eq/d)	2.0	2.1	1.8	2.2	2.3	1.8	2.2
Vegetables (cup eq/d)	1.9	2.2	1.7	2.0	2.3	1.7	2.0

(respiratory disease) to 48% (diabetes) lower risk of cause-specific mortality, while people in the highest quintile of cereal fiber intake had 15% (cancer) to 34% (diabetes) lower risk of cause-specific mortality.

When cereal fiber was further adjusted, the associations of whole grains with death from CVD, respiratory disease, infections, and other causes became non-significant; however, the associations with death from cancer and diabetes remained significant (*P* trend <0.029).

Discussion

In this large prospective cohort study of the US population, we found that high consumption of whole grains or cereal fiber was significantly associated with reduced risk of all-cause mortality and death from CVD, cancer, diabetes, respiratory disease, infections, and other causes. As compared with individuals with the lowest intake of whole grains, those in the highest intake group had a 17% lower risk of all-cause mortality and 11 to 48% lower risk of disease-specific mortality. As compared with individuals with the lowest intake of cereal fiber, those in the highest intake group had a 19% lower risk of all-cause mortality and 15 to 34% lower risk of disease-specific mortality. Furthermore, our results suggested that the protective effects of whole grains may due, at least in the main part, to its cereal fiber component.

To the best of our knowledge, the present study is, thus far, the largest in size regarding deaths in a prospective setting. Our findings are concordant with previously observed protective effects of whole grain intake on CVD, diabetes, and certain cancers [18,19]. Based on a meta-analysis of six cohort studies including 286,125 participants and 10,944 cases, a two servings per day increment in whole grain consumption was associated with a 21% (95% CI, 13–28%) decrease in risk of type 2 diabetes after adjustment for potential confounders and BMI [5]. These findings were confirmed by Ye et al.'s meta-analysis [20], in which it was also reported that compared with never/rare consumers of whole grains, individuals consuming 48 to 80 g of whole grains per day (3 to 5 serving/day) had a 21% lower risk of CVD (relative risk = 0.79; 95% CI, 0.74–0.85). Inverse associations were also reported between intake of whole grains and incident hypertension [21]. In a meta-analysis of 25 prospective studies, the summary relative risk of developing colorectal cancer for 10 g daily of cereal fiber was 0.90 (95% CI, 0.83–0.97), pooled results from six studies showed the relative risk for an increment of three servings daily of whole grains was 0.83 (95% CI, 0.78–0.89) [7]. High whole grain intakes have been related to reduced risk of other cancers, such as digestive cancer, in prospective studies, although the protective effects were not consistently observed [22,23].

Table 2 Association of whole grain intake with total and cause-specific mortality

	All participants	**Whole grains (oz eq/d)**					***P* trend**
		Q1 (n = 41,248)	**Q2 (n = 41,248)**	**Q3 (n = 41,249)**	**Q4 (n = 41,248)**	**Q5 (n = 41,249)**	
		0.13	**0.30**	**0.47**	**0.69**	**1.20**	
Causes of death							
All cause							
No. of deaths	46,067	11,845	9,450	8,694	8,054	8,024	
Model 1		1.00	0.78 (0.76–0.80)	0.70 (0.68–0.72)	0.63(0.61–0.65)	0.61 (0.59–0.62)	<0.0001
Model 2		1.00	0.88 (0.86–0.91)	0.83 (0.81–0.85)	0.78 (0.75–0.80)	0.77 (0.75–0.79)	<0.0001
Model 3		1.00	0.93 (0.90–0.95)	0.89 (0.87–0.92)	0.85 (0.82–0.87)	0.83 (0.81–0.86)	<0.0001
Model 4		1.00	0.96 (0.93–0.99)	0.95 (0.92–0.98)	0.92 (0.89–0.96)	0.94 (0.90–0.97)	0.002
Cardiovascular disease							
No. of deaths	11,283	2,921	2,330	2,121	1,914	1,997	<0.0001
Model 1		1.00	0.78 (0.74–0.83)	0.69 (0.65–0.73)	0.60 (0.57–0.64)	0.60 (0.57–0.64)	<0.0001
Model 2		1.00	0.88 (0.83–0.93)	0.81 (0.77–0.86)	0.73 (0.69–0.77)	0.75 (0.71–0.80)	<0.0001
Model 3		1.00	0.93 (0.88–0.98)	0.88 (0.83–0.93)	0.81 (0.77–0.86)	0.83 (0.78–0.88)	<0.0001
Model 4		1.00	0.96 (0.91–1.02)	0.95 (0.89–1.01)	0.90 (0.84–0.97)	0.95 (0.88–1.03)	0.188
Cancer							
No. of deaths	19,043	4,836	3,912	3,616	3,388	3,291	
Model 1		1.00	0.79 (0.76–0.83)	0.71 (0.68–0.75)	0.65 (0.62–0.68)	0.61 (0.59–0.64)	<0.0001
Model 2		1.00	0.91 (0.87–0.95)	0.86 (0.83–0.90)	0.82 (0.79–0.86)	0.80 (0.76–0.84)	<0.0001
Model 3		1.00	0.94 (0.90–0.98)	0.91 (0.87–0.95)	0.88 (0.84–0.92)	0.85 (0.81–0.89)	<0.0001
Model 4		1.00	0.96 (0.92–1.00)	0.95 (0.90–0.99)	0.93 (0.88–0.98)	0.93 (0.88–0.99)	0.025
Diabetes							
No. of deaths	371	113	72	73	66	47	
Model 1		1.00	0.62 (0.46–0.84)	0.62 (0.46–0.83)	0.54 (0.40–0.73)	0.37 (0.27–0.52)	<0.0001
Model 2		1.00	0.67 (0.49–0.90)	0.67 (0.50–0.91)	0.60 (0.44–0.82)	0.42 (0.30–0.60)	<0.0001
Model 3		1.00	0.71 (0.53–0.96)	0.76 (0.56–1.03)	0.72 (0.53–0.99)	0.52 (0.37–0.75)	0.0009
Model 4		1.00	0.74 (0.54–1.00)	0.81 (0.59–1.13)	0.78 (0.55–1.13)	0.57 (0.37–0.89)	0.029
Respiratory disease							
No. of deaths	3,796	1,123	802	673	606	592	
Model 1		1.00	0.69 (0.63–0.75)	0.56 (0.50–0.61)	0.48 (0.43–0.53)	0.45 (0.41–0.50)	<0.0001
Model 2		1.00	0.88 (0.80–0.96)	0.79 (0.72–0.87)	0.74 (0.67–0.82)	0.74 (0.67–0.82)	<0.0001
Model 3		1.00	0.99 (0.90–1.09)	0.94 (0.85–1.03)	0.91 (0.82–1.01)	0.89 (0.80–0.98)	0.0099
Model 4		1.00	1.02 (0.93–1.12)	1.00 (0.90–1.11)	1.01 (0.90–1.14)	1.03 (0.91–1.18)	0.67
Infections							
No. of deaths	922	251	184	163	161	163	
Model 1		1.00	0.71 (0.59–0.86)	0.61 (0.50–0.75)	0.58 (0.48–0.71)	0.57 (0.47–0.70)	<0.0001
Model 2		1.00	0.78 (0.64–0.94)	0.69 (0.57–0.84)	0.68 (0.55–0.83)	0.68 (0.55–0.83)	0.0002
Model 3		1.00	0.84 (0.70–1.02)	0.78 (0.64–0.96)	0.79 (0.65–0.97)	0.77 (0.62–0.95)	0.02
Model 4		1.00	0.87 (0.71–1.06)	0.83 (0.67–1.04)	0.87 (0.69–1.10)	0.89 (0.68–1.16)	0.54

Table 2 Association of whole grain intake with total and cause-specific mortality *(Continued)*

Other/unknown causes							
No. of deaths	5,223	1,206	1,058	1,038	940	981	
Model 1		1.00	0.86 (0.79–0.93)	0.82 (0.76–0.90)	0.71 (0.66–0.78)	0.72 (0.66–0.78)	<0.0001
Model 2		1.00	0.91 (0.84–0.99)	0.90 (0.83–0.98)	0.80 (0.73–0.87)	0.81 (0.74–0.88)	<0.0001
Model 3		1.00	0.97 (0.88–1.06)	0.97 (0.89–1.06)	0.87 (0.79–0.96)	0.86 (0.78–0.94)	0.0001
Model 4		1.00	0.99 (0.91–1.09)	1.03 (0.93–1.13)	0.96 (0.86–1.06)	0.98 (0.87–1.09)	0.54

Data are hazard ratios (HR) and 95% confidence interval (CI). The numbers of deaths are for participants who died during follow-up. The co-variables are baseline assessments.

Model 1, Adjusted for age and gender; Model 2, Adjusted for age, gender, the number of cigarettes smoked per day, and time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline); Model 3, Adjusted for age, gender, the number of cigarettes smoked per day, time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline), race or ethnicity group, alcohol intake, education level, marital status (yes, no), health status (poor, fair, good, very good), obesity (underweight, overweight, obesity), physical activity, consumption of red meat, total fruit and total vegetables, total energy intake, and hormone usage. Model 4, Based on Model 3 further adjusted for cereal fiber.

Very few previous studies have examined the relationship between whole grains and their components with mortality in humans. Our findings are consistent with the results reported in the Nurses' Health Study, in which whole grain intake, especially bran, was associated with lower all-cause and CVD mortality in diabetic women [24]. Similarly, higher fiber intake was associated with lower total mortality, particularly mortality from circulatory, digestive, and non-CVD non-cancer inflammatory diseases in a large European prospective study of 452,717 men and women [25]. In a previous analysis among our study samples, it was found that intake of fiber from grains but not from other sources was inversely related to all-cause mortality and death from cancer, CVD, infections, and respiratory disease [9]. In this updated analysis, we found cereal fiber intake was inversely associated with death from diabetes. However, we did not report the associations of specific types of whole grain foods/products with mortality and cause-specific mortality, since it is hard to further differentiate such food groups; this presents a limitation of this observational study.

In addition, we found that the associations of whole grains with death from CVD, respiratory disease, and infections became non-significant after adjustment for cereal fiber intake. The associations with total mortality and death from cancer and diabetes were also largely attenuated, although they remained significant after adjustment for cereal fiber intake. These observations suggest that the protective effects of whole grains on mortality are at least partly mediated by its cereal fiber component. Such a postulation is supported by previous evidence that shows cereal fiber intake is related to an improvement of insulin sensitivity and lipid profile, an increase in protective molecules such as adiponectin, and a reduction in inflammation markers [26-28].

The protective effect of whole grains and fiber consumption on risk of mortality is biologically plausible. Dietary fiber intake is associated with lower levels of inflammation markers, such as C-reactive protein, and tumor necrosis factor α receptor 2, which play key roles in chronic inflammatory conditions [29,30]. Whole grain foods are rich in fiber. Therefore, the anti-inflammatory effects of dietary fiber may help explain, at least in part, the inverse associations of whole grains and fiber consumption with chronic disease death. Moreover, whole grains and cereal fiber have a high content of antioxidants, vitamins, trace minerals, phenolic acids, lignans, and phytoestrogens, which have been associated with a reduced risk of colorectal cancer [31] and lower risk of death from non-cardiovascular, non-cancer inflammatory diseases and respiratory system diseases [32]. In addition, dietary fibers have specific and unique impacts on intestinal microbiota composition and metabolism [33,34]. Additionally, recent studies have related gut microbiota with various chronic diseases such as obesity, CVD, diabetes, and cancer [34,35]. Further functional investigations are warranted to verify these potential mechanisms.

Strengths and limitations of the study

In our study cohort, both whole grains and cereal fiber were correlated with high levels of physical activity and better health status, as well as with low BMI, low levels of smoking, and low intakes of alcohol and red meat. However, our results were less likely due to the potential confounding of these factors because careful adjustment for these factors in our analyses did not significantly change the results. Nevertheless, we acknowledge that the positive associations may still be related to residual confounding of non-measured covariates. Reverse causality might also affect the associations, since people with chronic disease might modify their eating habits by consuming healthy foods including those rich in whole grains and cereal fiber. In our analyses, however, we have excluded patients with cancer, heart disease, and diabetes at baseline and only analyzed the associations with incident cases. Whole grains and cereal fiber intakes were evaluated by self-report at a single time point. It is

Table 3 Association of cereal fiber intake with total and cause-specific mortality

	All participants	**Cereal fiber (g/d)**					***P* trend**
		Q1 (n = 73,488)	**Q2 (n = 73,488)**	**Q3 (n = 73,489)**	**Q4 (n = 73,488)**	**Q5 (n = 73,489)**	
		2.02	**4.15**	**5.27**	**6.65**	**10.22**	
Causes of death							
All cause							
No. of deaths	46,067	11,700	9,652	8,664	8,133	7,918	
Model 1		1.00	0.80 (0.78–0.83)	0.70 (0.68–0.72)	0.64 (0.62–0.66)	0.59 (0.58–0.61)	<0.0001
Model 2		1.00	0.89 (0.87–0.92)	0.83 (0.81–0.85)	0.78 (0.76–0.81)	0.76 (0.73–0.78)	<0.0001
Model 3		1.00	0.93 (0.90–0.95)	0.87 (0.85–0.90)	0.84 (0.81–0.86)	0.81 (0.79–0.84)	<0.0001
Cardiovascular disease	11,283						
No. of deaths		2,901	2,368	2,094	1,986	1,934	
Model 1		1.00	0.80 (0.76–0.85)	0.69 (0.65–0.73)	0.63 (0.59–0.66)	0.57 (0.54–0.61)	<0.0001
Model 2		1.00	0.88 (0.83–0.93)	0.80 (0.76–0.85)	0.76 (0.71–0.80)	0.72 (0.68–0.76)	<0.0001
Model 3		1.00	0.93 (0.87–0.98)	0.86 (0.81–0.91)	0.83 (0.78–0.88)	0.80 (0.75–0.85)	<0.0001
Cancer							
No. of deaths	19,043	4,772	3,974	3,616	3,391	3,290	
Model 1		1.00	0.81 (0.78–0.85)	0.72 (0.69–0.76)	0.66 (0.63–0.69)	0.61 (0.59–0.64)	<0.0001
Model 2		1.00	0.91 (0.87–0.95)	0.87 (0.83–0.91)	0.83 (0.79–0.86)	0.80 (0.77–0.84)	<0.0001
Model 3		1.00	0.94 (0.90–0.98)	0.90 (0.86–0.95)	0.87 (0.83–0.91)	0.85 (0.81–0.89)	<0.0001
Diabetes							
No. of deaths	371	92	92	72	57	51	
Model 1		1.00	0.92 (0.69–1.22)	0.70 (0.52–0.95)	0.54 (0.39–0.74)	0.45 (0.32–0.64)	<0.0001
Model 2		1.00	0.97 (0.73–1.29)	0.77 (0.56–1.04)	0.60 (0.43–0.83)	0.52 (0.37–0.73)	<0.0001
Model 3		1.00	1.00 (0.74–1.35)	0.83 (0.60–1.15)	0.70 (0.50–0.99)	0.66 (0.46–0.94)	0.005
Respiratory disease							
No. of deaths	3,796	1,082	866	652	649	547	
Model 1		1.00	0.75 (0.68–0.82)	0.54 (0.49–0.59)	0.52 (0.47–0.57)	0.42 (0.38–0.46)	<0.0001
Model 2		1.00	0.91(0.83–1.00)	0.76 (0.69–0.84)	0.80 (0.72–0.88)	0.70 (0.63–0.78)	<0.0001
Model 3		1.00	0.98 (0.89–1.08)	0.82 (0.74–0.91)	0.89 (0.80–0.99)	0.79 (0.71–0.88)	<0.0001
Infections							
No. of deaths	922	230	210	165	157	160	
Model 1		1.00	0.87 (0.72–1.05)	0.66 (0.54–0.81)	0.61 (0.50–0.75)	0.60 (0.49–0.74)	<0.0001
Model 2		1.00	0.94 (0.78–1.13)	0.75 (0.61–0.91)	0.71 (0.58–0.87)	0.71 (0.58–0.88)	0.0002
Model 3		1.00	1.04 (0.85–1.27)	0.84 (0.68–1.05)	0.82 (0.66–1.02)	0.83 (0.67–1.03)	0.023
Other/unknown causes							
No. of deaths	5,223	1,215	1,050	1,044	963	951	
Model 1		1.00	0.84 (0.77–0.91)	0.81 (0.74–0.88)	0.72 (0.66–0.78)	0.67 (0.62–0.73)	<0.0001
Model 2		1.00	0.88 (0.81–0.96)	0.88 (0.81–0.95)	0.79 (0.73–0.87)	0.76 (0.70–0.83)	<0.0001
Model 3		1.00	0.90 (0.83–0.99)	0.91 (0.83–0.99)	0.83 (0.76–0.91)	0.79 (0.72–0.87)	<0.0001

Data are hazard ratios (HR) and 95% confidence interval (CI). The numbers of deaths are for participants who died during follow-up. The co-variables are baseline assessments.

Model 1, Adjusted for age and gender; Model 2, Adjusted for age, gender, the number of cigarettes smoked per day, and time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline); Model 3, Adjusted for age, gender, the number of cigarettes smoked per day, time of smoking cessation (<1 years, 1 to 5 years, 5 to 10 years, or ≥10 years before baseline), race or ethnicity group, alcohol intake, education level, marital status (yes, no), health status (poor, fair, good, very good), obesity (underweight, overweight, obesity), physical activity, consumption of red meat, total fruit and total vegetables, total energy intake, and hormone usage.

likely that dietary habits might change during the long (14 years on average) follow-up period, and such temporal patterns were not reflected in our analysis. Moreover, the observational nature of our study limits causality inference between intakes of whole grains or cereal fiber and mortality.

Conclusions

Data from our study indicate that intake of whole grains and cereal fiber may reduce the risk of all-cause mortality and death from chronic diseases such as cancer, CVD, diabetes, respiratory disease, infections, and other causes. Disappearance or attenuation of whole grain associations with total mortality and death from chronic diseases after adjustment for cereal fiber intake suggests that cereal fiber partly accounts for the protective effects of whole grains on mortality.

Abbreviations

CI: Confidence intervals; CVD: Cardiovascular diseases; FFQ: Food frequency questionnaire; HR: Hazard ratio; NIH: National Institutes of Health.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TH and LQ conceived the study. TH, MX, and LQ analyzed the data and wrote the draft of the paper. MX, AL, SC, and LQ contributed to writing, reviewing, and revising of the paper. LQ is the guarantor. All authors read and approved the final manuscript.

Acknowledgments

We thank all the participants in the NIH-AARP Diet and Health Study. We also thank Dr. David Hasza for statistical assistance.

Sources of support

This study is funded by an unrestricted research fund from NutraSource. Dr. Qi was supported by grants from the National Heart, Lung, and Blood Institute (HL071981), the National Institute of Diabetes and Digestive and Kidney Diseases (DK091718), the Boston Obesity Nutrition Research Center (DK46200), and United States–Israel Binational Science Foundation Grant 2011036. Dr. Qi was a recipient of the American Heart Association Scientist Development Award (0730094 N). Funding from NutraSource. There were no other relationships or activities that could appear to have influenced the submitted work.

Author details

[1]Department of Nutrition, Harvard School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA. [2]NutraSource (AWL), Royal Oak, MI 48073, USA. [3]NutraSource (SSC), Clarksville, MD 21029, USA. [4]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 75 Francis St, Boston, MA 02115, USA.

Received: 29 October 2014 Accepted: 13 February 2015
Published online: 24 March 2015

References

1. Slavin J. Whole grains and human health. Nutr Res Rev. 2004;17:99–110.
2. Qi L, Hu FB. Dietary glycemic load, whole grains, and systemic inflammation in diabetes: the epidemiological evidence. Curr Opin Lipidol. 2007;18:3–8.
3. Qi L, van Dam RM, Liu S, Franz M, Mantzoros C, Hu FB. Whole-grain, bran, and cereal fiber intakes and markers of systemic inflammation in diabetic women. Diabetes Care. 2006;29:207–11.
4. Harland JI, Garton LE. Whole-grain intake as a marker of healthy body weight and adiposity. Public Health Nutr. 2008;11:554–63.
5. de Munter JS, Hu FB, Spiegelman D, Franz M, van Dam RM. Whole grain, bran, and germ intake and risk of type 2 diabetes: a prospective cohort study and systematic review. PLoS Med. 2007;4:e261.
6. Mellen PB, Walsh TF, Herrington DM. Whole grain intake and cardiovascular disease: a meta-analysis. Nutr Metab Cardiovasc Dis. 2008;18:283–90.
7. Aune D, Chan DS, Lau R, Vieira R, Greenwood DC, Kampman E, et al. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose–response meta-analysis of prospective studies. BMJ. 2011;343:d6617.
8. Pajvani UB, Du X, Combs TP, Berg AH, Rajala MW, Schulthess T, et al. Structure-function studies of the adipocyte-secreted hormone Acrp30/ adiponectin: Implications fpr metabolic regulation and bioactivity. J Biol Chem. 2003;278:9073–85.
9. Park Y, Subar AF, Hollenbeck A, Schatzkin A. Dietary fiber intake and mortality in the NIH-AARP diet and health study. Arch Intern Med. 2011;171:1061–8.
10. Adams KF, Schatzkin A, Harris TB, Kipnis V, Mouw T, Ballard-Barbash R, et al. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. N Engl J Med. 2006;355:763–78.
11. Lacey Jr JV, Brinton LA, Leitzmann MF, Mouw T, Hollenbeck A, Schatzkin A, et al. Menopausal hormone therapy and ovarian cancer risk in the National Institutes of Health-AARP Diet and Health Study Cohort. J Natl Cancer Inst. 2006;98:1397–405.
12. Willett W. Nutritional epidemiology: issues and challenges. Int J Epidemiol. 1987;16:312–7.
13. Schatzkin A1, Mouw T, Park Y, Subar AF,Kipnis V, Hollenbeck A, et al. Dietary fiber and whole-grain consumption in relation to colorectal cancer in the NIH-AARP Diet and Health Study. Am J Clin Nutr. 2007;85:1353–60.
14. Thompson FE, Kipnis V, Subar AF, Krebs-Smith SM, Kahle LL, Midthune D, et al. Evaluation of 2 brief instruments and a food-frequency questionnaire to estimate daily number of servings of fruit and vegetables. Am J Clin Nutr. 2000;71:1503–10.
15. International Classification of Diseases, Ninth Revision (ICD-9). World Health Organization. http://www.who.int/classifications/icd/en/
16. International Classification of Diseases (ICD). World Health Organization. Retrieved 23 November 2010. http://apps.who.int/classifications/icd10/browse/2010/en
17. Willett W. Nutritional epidemiology. 2nd ed. New York: Oxford University Press; 1998.
18. Borneo R, Leon AE. Whole grain cereals: functional components and health benefits. Food Funct. 2012;3:110–9.
19. Slavin J. Why whole grains are protective: biological mechanisms. Proc Nutr Soc. 2003;62:129–34.
20. Ye EQ, Chacko SA, Chou EL, Kugizaki M, Liu S. Greater whole-grain intake is associated with lower risk of type 2 diabetes, cardiovascular disease, and weight gain. J Nutr. 2012;142:1304–13.
21. Flint AJ, Hu FB, Glynn RJ, Jensen MK, Franz M, Sampson L, et al. Whole grains and incident hypertension in men. Am J Clin Nutr. 2009;90:493–8.
22. Egeberg R, Olsen A, Christensen J, Johnsen NF, Loft S, Overvad K, et al. Intake of whole-grain products and risk of prostate cancer among men in the Danish Diet, Cancer and Health cohort study. Cancer Causes Control. 2011;22:1133–9.
23. Schatzkin A, Park Y, Leitzmann MF, Hollenbeck AR, Cross AJ. Prospective study of dietary fiber, whole grain foods, and small intestinal cancer. Gastroenterology. 2008;135:1163–7.
24. He M, van Dam RM, Rimm E, Hu FB, Qi L. Whole-grain, cereal fiber, bran, and germ intake and the risks of all-cause and cardiovascular disease-specific mortality among women with type 2 diabetes mellitus. Circulation. 2010;121:2162–8.
25. Chuang SC, Norat T, Murphy N, Olsen A, Tjonneland A, Overvad K, et al. Fiber intake and total and cause-specific mortality in the European Prospective Investigation into Cancer and Nutrition cohort. Am J Clin Nutr. 2012;96:164–74.
26. Qi L, Meigs JB, Liu S, Manson JE, Mantzoros C, Hu FB. Dietary fibers and glycemic load, obesity, and plasma adiponectin levels in women with type 2 diabetes. Diabetes Care. 2006;29:1501–5.
27. Esposito K, Giugliano D. Whole-grain intake cools down inflammation. Am J Clin Nutr. 2006;83:1440–1. Author reply 1441–2.
28. Weickert MO, Mohlig M, Schofl C, Arafat AM, Otto B, Viehoff H, et al. Cereal fiber improves whole-body insulin sensitivity in overweight and obese women. Diabetes Care. 2006;29:775–80.
29. Ma Y, Griffith JA, Chasan-Taber L, Olendzki BC, Jackson E, Stanek 3rd EJ, et al. Association between dietary fiber and serum C-reactive protein. Am J Clin Nutr. 2006;83:760–6.

30. Wannamethee SG, Whincup PH, Thomas MC, Sattar N. Associations between dietary fiber and inflammation, hepatic function, and risk of type 2 diabetes in older men: potential mechanisms for the benefits of fiber on diabetes risk. Diabetes Care. 2009;32:1823–5.
31. Slavin JL. Mechanisms for the impact of whole grain foods on cancer risk. J Am Coll Nutr. 2000;19:300S–7.
32. Jacobs Jr DR, Andersen LF, Blomhoff R. Whole-grain consumption is associated with a reduced risk of noncardiovascular, noncancer death attributed to inflammatory diseases in the Iowa Women's Health Study. Am J Clin Nutr. 2007;85:1606–14.
33. Yang J, Martinez I, Walter J, Keshavarzian A, Rose DJ. In vitro characterization of the impact of selected dietary fibers on fecal microbiota composition and short chain fatty acid production. Anaerobe. 2013;23:74–81.
34. Parnell JA, Reimer RA. Prebiotic fibres dose-dependently increase satiety hormones and alter Bacteroidetes and Firmicutes in lean and obese JCR:LA-cp rats. Br J Nutr. 2012;107:601–13.
35. Saura-Calixto F. Dietary fiber as a carrier of dietary antioxidants: an essential physiological function. J Agric Food Chem. 2011;59:43–9.

The New York Times

THE NEW HEALTH CARE

Congratulations. Your Study Went Nowhere.

Researchers should embrace negative results instead of accentuating the positive, which is one of several biases that can lead to bad science.

By Aaron E. Carroll

Sept. 24, 2018

A review of research on antidepressants showed how biases can harm science. Lucy Nicholson/Reuters

When we think of biases in research, the one that most often makes the news is a researcher's financial conflict of interest. But another bias, one possibly even more pernicious, is how research is published and used in supporting future work.

A recent study in Psychological Medicine examined how four of these types of biases came into play in research on antidepressants. The authors created a data set containing 105 studies of antidepressants that were registered with the Food and Drug Administration. Drug companies are required to register trials before they are done, so the researchers knew they had more complete information than what might appear in the medical literature.

Publication bias refers to the decision on whether to publish results based on the outcomes found. With the 105 studies on antidepressants, half were considered "positive" by the F.D.A., and half were considered "negative." Ninety-eight percent of the positive trials were published; only 48 percent of the negative ones were.

Outcome reporting bias refers to writing up only the results in a trial that appear positive, while failing to report those that appear negative. In 10 of the 25 negative studies, studies that were considered negative by the F.D.A. were reported as positive by the researchers, by switching a secondary outcome with a primary one, and reporting it as if it were the original intent of the researchers, or just by not reporting negative results.

Spin refers to using language, often in the abstract or summary of the study, to make negative results appear positive. Of the 15 remaining "negative" articles, 11 used spin to puff up the results. Some talked about statistically nonsignificant results as if they were positive, by referring only to the numerical outcomes. Others referred to trends in the data, even though they lacked significance. Only four articles reported negative results without spin.

Spin works. A randomized controlled trial found that clinicians who read abstracts in which nonsignificant results for cancer treatments were rewritten with spin were more likely to think the treatment was beneficial and more interested in reading the full-text article.

It gets worse. Research becomes amplified by citation in future papers. The more it's discussed, the more it's disseminated both in future work and in practice. Positive studies were cited three times more than negative studies. This is **citation bias**.

Only half of the research was positive. Almost no one would know that. Even thorough reviews of the literature would find that nearly all studies were positive, and those that were negative were ignored. This is one reason you wind up with 10 percent of Americans on antidepressants when good research shows the efficacy of many of the drugs is far less than believed.

The preregistration of trials is supposed to help control for these biases. It works sporadically. In 2011, researchers examined cohorts of randomized controlled trials to see how well the published research matched what scientists said it was going to do beforehand. In some studies, they found, eligibility criteria for participants differed greatly from what was published.

In some, they found that procedures had changed for how to conduct analyses. In almost all, the sample size calculations had changed. Almost none reported on all the outcomes that were noted in the protocols or registries. Primary outcomes were changed or dropped in up to half of publications. This isn't to say secondary outcomes don't matter; they're often very important. It's also possible that some of these decisions were made for legitimate reasons, but, too often, there are no explanations.

In 2012, researchers re-analyzed 42 meta-analyses for nine drugs in six classes that had been approved by the F.D.A. In their re-analyses, they included data from the F.D.A. that was not in the medical literature. The addition of the new data changed the results in more than 90 percent of the studies. In those where efficacy went down, it did so by a median 11 percent. When efficacy went up — about the same rate that it went down — it did so by a median 13 percent.

This problem is worldwide. In 2004 in JAMA, a study reviewed more than 100 trials approved by a scientific-ethical committee in Denmark that resulted in 122 publications and more than 3,700 outcomes. But a great deal went unreported: about half of the outcomes on whether the drugs worked, and about two-thirds of the outcomes on whether the drugs caused harm. Positive outcomes were more likely to be reported. More than 60 percent of trials had at least one primary outcome changed or dropped.

But when the researchers surveyed the scientists who conducted the trials and published the results, 86 percent reported that there were no unpublished outcomes.

There has even been a systematic review of the many studies of these types of biases. It provides empirical evidence that the biases are widespread and cover many domains.

A modeling study published in BMJ Open in 2014 showed that if a publication bias caused positive findings to be published at four times the rate of negative ones for a particular treatment, 90 percent of large meta-analyses would later conclude that the treatment worked when it actually didn't.

This doesn't mean we should discount all results from medical trials. It means that we need, more than ever, to reproduce research to make sure it's robust. Dispassionate third parties who attempt to achieve the same results will fail to do so if the reported findings have been massaged in some way.

Further, there are things we can do to fix this problem. We can demand that trial results be published, regardless of findings. To that end, we can encourage journals to publish negative results as doggedly as positive ones. We can ensure that preregistered protocols and outcomes are the ones that are finally reported in the literature. We can hold authors to more rigorous standards when they publish, so that results are accurately and transparently reported. We can celebrate and elevate negative results, in both our arguments and reporting, as we do positive ones. Unfortunately, getting such research published is harder than it should be.

These actions might make for more boring news and more tempered enthusiasm. But they might also lead to more accurate science.

Aaron E. Carroll is a professor of pediatrics at Indiana University School of Medicine who blogs on health research and policy at The Incidental Economist and makes videos at Healthcare Triage. He is the author of "The Bad Food Bible: How and Why to Eat Sinfully." @aaronecarroll

READ 11 COMMENTS

The New York Times

PHYS ED

Dog Owners Get More Exercise

Dog owners spent close to 300 minutes each week walking with their dogs, about 200 more minutes of walking than people without dogs.

By Gretchen Reynolds

May 29, 2019

Dog owners are about four times more likely than other people to meet today's physical activity guidelines, according to a large-scale new study of dogs and exercise.

The study, which involved hundreds of British households, suggests that having a dog can strongly influence how much people exercise. But it also raises questions about why some dog owners never walk their pets or otherwise work out and whether any of us should acquire a dog just to encourage us to move.

Most people who live with dogs, including me, are familiar with their joy at ambling along paths, trails and sidewalks. We also have to deal with their jowly dejection when our work deadlines or other issues interfere with walks.

Few of us would be surprised that past studies have found links between dog ownership and frequent walking. But many of those studies have been small and relied solely on people's sometimes-unreliable recall of their exercise routines. They also have not looked at whether walking a dog might displace other kinds of physical activity, which would mean that dog owners were not exercising more, in total, than other people; only that they were exercising more often with a dog.

Those issues prompted exercise scientists from the University of Liverpool and other institutions to decide recently to undertake one of the most comprehensive comparisons yet of how often, whether and in what ways dog owners and their dog-less neighbors exercise.

So, for the new study, which was published in April in Scientific Reports, they first turned to a neighborhood near Liverpool and began asking families in the area about their lives and pets. The researchers focused on a single community, so that everyone involved should share approximately the same local environment with similar access to sidewalks, parks or other amenities that might affect their exercise routines.

They wound up with almost 700 participants from 385 neighboring households, half of them women and most middle-aged, although about 70 children also participated. About a third of these people owned a dog.

The scientists asked everyone in these households, including the children, to complete lengthy questionnaires about how much and in what ways they moved each week. They also provided activity monitors to a few of the families and asked the members to wear them for a week while exercising as usual.

Then they collected and compared data.

It was immediately apparent that people who owned dogs walked far more often than those without dogs, says Carri Westgarth, a lecturer in human-animal interaction at the University of Liverpool, who led the new study.

In general, according to both the questionnaires and activity monitors, most dog owners spent close to 300 minutes each week walking with their dogs, which was about 200 more minutes of walking per week than people without dogs.

Due primarily to these walks, most dog owners met or exceeded the standard guidelines for exercising for health, which call for at least 150 minutes of moderate exercise every week.

More unexpected, dog owners also spent slightly more time than other people jogging, cycling and visiting the gym solo, without their dogs, indicating that walking Fido had not bumped other activities from their lives.

The influence of dogs extended also to the young, the scientists found. Children whose families owned dogs walked for about 100 minutes each week and played and romped with their pets for another 200 minutes, making them substantially more active than children in homes without dogs.

At the same time, though, and to the puzzlement of the researchers, a small portion of dog owners never walked their dogs, and those owners almost all were young, healthy and female.

Taken as a whole, the results suggest that people with dogs are more physically active than those without, Dr. Westgarth says.

But the findings also show that dog owners can remain sedentary and their reasons should be investigated, she says. The women in this study who did not walk their pets may have worried about controlling their animals or their safety on the streets, or they may simply have disliked walking.

Such concerns need to be acknowledged, understood and addressed if having a dog is to be promoted as a way of increasing exercise, she says.

Of course, this kind of observational study cannot tell us whether dog ownership actually causes people to move more, or if active people also own dogs. The study also did not account for differences in pets' sizes, breeds, temperament or training and whether those affect owners' willingness to walk, although the researchers plan to look at those issues in future studies.

For now, Dr. Westgarth says, she would not advise anyone to buy a dog only in hopes that, like a furry Fitbit, it will prod us to move.

"A dog is not a tool just to make us more physically active," she says. "But if you feel that you have the time, inclination and finances to take on the responsibility of having a dog, they are a great motivator to get out walking when you otherwise would have made excuses not to."

READ 204 COMMENTS

The Washington Post
Democracy Dies in Darkness

Dog owners are much happier than cat owners, survey finds

By **Christopher Ingraham**

April 5, 2019 at 4:00 a.m. MDT

The well-respected survey that's been a barometer of American politics, culture and behavior for more than four decades finally got around to the question that has bedeviled many a household.

Dog or cat?

In 2018, the General Social Survey for the first time included a battery of questions on pet ownership. The findings not only quantified the nation's pet population — nearly 6 in 10 households have at least one —they made it possible to see how pet ownership overlaps with all sorts of factors of interest to social scientists.

Like happiness.

For starters, there is little difference between pet owners and non-owners when it comes to happiness, the survey shows. The two groups are statistically indistinguishable on the likelihood of identifying as "very happy" (a little over 30 percent) or "not too happy" (in the mid-teens).

But when you break the data down by pet type — cats, dogs or both — a stunning divide emerges: Dog owners are about twice as likely as cat owners to say they're very happy, with people owning both falling somewhere in between.

Dog people, in other words, are slightly happier than those without any pets. Those in the cat camp, on the other hand, are significantly less happy than the pet-less. And having both appears to cancel each other out happiness-wise. (Since someone's bound to ask, it isn't possible to do this same type of analysis for say, rabbit owners or lizard owners or fish owners, since there aren't enough of those folks in the survey to make a statistically valid sample).

These differences are quite large: The happiness divide between dog and cat owners is bigger than the one between people who identify as middle and upper class, and nearly as large as the gap between those who say they're in "fair" versus "good or excellent" health.

However, correlation doesn't equal causation, and there are probably a number of other differences between dog and cat owners that account for some of the differences. The General Social Survey data show that dog owners, for instance, are more likely to be married and own their own homes than cat owners, both factors known to affect happiness and life satisfaction.

Previous research on this topic yielded mixed results. In 2006, the Pew Research Center found no significant differences in happiness between pet owners and non-pet owners, or cat and dog owners. However, that survey did not distinguish between people who owned “only” a dog or a cat, and those who owned “either” a dog or a cat, potentially muddying the distinctions between exclusive dog and cat owners.

A 2016 study of dog and cat owners, on the other hand, yielded greater happiness ratings for dog owners relative to cat people. It attributed the contrast, at least in part, to differences in personality: Dog owners tended to be more agreeable, more extroverted and less neurotic than cat owners. And a 2015 study linked the presence of a cat in the home to fewer negative emotions, but not necessarily an increase in positive ones.

Other research makes the case that some of the pet-happiness relationship is causal, at least when it comes to canines. A 2013 study found, for instance, that dog owners are more likely to engage in outdoor physical activity than people who don’t own dogs, with obvious benefits for health and happiness.

Research also has shown that dog owners are more likely than other folks to form friendships with people in their neighborhoods on the basis of the random encounters that happen when they’re out walking their pets. Those social connections likely contribute to greater well-being among dog owners.

The General Social Survey also asked a number of questions about how people interact with their pets, and the answers may also explain some of the happiness gap. Dog owners, for instance, are more likely to seek comfort from their pet in times of stress, more likely to play with their pet, and more likely to consider their pet a member of their family. Those differences suggest a stronger social bond with their pets, which could create a greater sense of well-being.

Stepping away from the data, cat owners might protest that ownership isn't about "happiness" at all: There's something about felines that is grander and more mysterious — something that can't be captured in a public opinion poll.

"A cat has absolute emotional honesty," as Ernest Hemingway put it. "Human beings, for one reason or another, may hide their feelings, but a cat does not."

Christopher Ingraham

Christopher Ingraham

Christopher Ingraham writes about all things data. He previously worked at the Brookings Institution and the Pew Research Center. Follow

Your profile is incomplete

Before you can contribute to our community, please visit your Profile page in order to complete your profile.

Comments

Comments are now closed

All comments sections automatically close 14 days after the story has published. For more details, please see our **discussion guidelines**.

All Comments (1.1k) Viewing Options ▾

blipper 5 months ago

Mr. Peanutbutter versus Princess Carolyn.

Like 👍 1 Link Report ⚑

Liz Harris 6 months ago

Yes and people who wear blue on Tuesdays are happier than people who wear yellow on Thursdays.

Like 👍 Link Report ⚑

Alex Grin 6 months ago

Interesting article, was useful to know. I'm fond of dogs too. I have English Bulldog and I realized that if you're getting a dog, expect to spend as much time with it as you would a baby. I checked all info about best food for dogs on https://petstiger.com/best-dog-food-for-english-bulldogs/ to keep my pet healthy. He's on a healthy diet right now.

Like 👍 Link Report ⚑

Mandy Cat 6 months ago

Perhaps dog owners are happier because they enjoy the slavish adulation dogs tend to provide. We cat folks prefer no nonsense feline commentary. Painful at times, yes.

Like 👍 2 Link Report ⚑

Cyclopsina 6 months ago

I have a cat and a dog. The dog was trained to help us with foster cats from the Humane Society. We adopted our last foster cat because she was terminally ill and had six months to live, and she required daily IV treatments that we had to administer at home. We figured that no one would adopt her so we did. This was four years ago. She miraculously recovered the minute we signed the adoption papers.

Our dog and the cat sleep next to each other, and play together. It is very sweet.

Like 4 Link Report

> **Dr.Who3** 6 months ago
>
> You and your partner sound like very good heart-ed people.
>
> Like 1 Link Report

Jamie Rash III 6 months ago

I'm just going to say it: We cat people are more intelligent than dog people. And as everyone knows, the smarter you are, the less happy. ;p

Like 6 Link Report

> **Stressor** 6 months ago
>
> I have two dogs, but you might be on to something. ;)
>
> Like 2 Link Report

> **ExcuuuuseMe** 6 months ago
>
> Don't I know it! There are mornings I wake up and think I just can't go on ... then, my rescue cat starts meowing at me that he wants to have some breakfast and play. So, I go on...
>
> Like 2 Link Report

Jim Meyerling 6 months ago

This article is mostly about the maudlin nonsense of the immature. Dogs...cats...parakeets....idiotic. Those animals are parasites who exploit stupid humans for food, shelter and such so they can lay around leading a cushy life. My pet spider Ethel displays none of that. She's a black widow who doesn't believe in divorce. No. Her husband Wilson was much smaller than Ethel it's true....but....Ethel tired of his attentions so rather than divorce him....she had him for lunch. She's independent, obtains her own food....she's fond of bugs trapped in her web...and hangs out in her web. The perfect pet. And...she gets rid of house flies as well.

Like 👍 4 Link 🔗 Report ⚑

Dr.Who3 6 months ago

Next up, do blondes or brunettes make better wives? Tune in tomorrow for another debate.

Like 👍 2 Link 🔗 Report ⚑

Dr.Who3 6 months ago

A choice of pet is like a religion, it should be between the owner and the pet, unless the pet is dangerous.

Like 👍 Link 🔗 Report ⚑

Dr.Who3 6 months ago

I am not convinced that it is having a dog that makes the owner happier than owning a cat. Correlation does not equal causation, and the differences are not that great anyway. So really, what should one conclude from this 'study'?

Any thoughts on this question?

Like 👍 Link 🔗 Report ⚑

Stressor 6 months ago

Sometimes, articles appear for entertainment purposes. And you know what Twain said about statistics.

Like 1 Link Report

Barry-NJ 6 months ago

I don't think that's because happier people buy dogs.hat the article ever implied causation. It just said that dog owners were happier. For all we know,

Like Link Report

bannedindc 6 months ago

The article addresses the correlation/causation point.

Like Link Report

View More Comments

The New York Times | http://nyti.ms/1IlyFyK

TECHNOLOGY

Facebook Use Polarizing? Site Begs to Differ

By FARHAD MANJOO MAY 7, 2015

For years, political scientists and social theorists have fretted about the Internet's potential to flatten and polarize democratic discourse.

Because so much information now comes through digital engines shaped by our preferences — Facebook, Google and others suggest content based on what consumers previously enjoyed — scholars have theorized that people are building an online echo chamber of their own views.

But in a peer-reviewed study published on Thursday in the journal Science, data scientists at Facebook report the echo chamber is not as insular as many might fear — at least not on the social network. While independent researchers said the study was important for its scope and size, they noted several significant limitations.

After analyzing how 10.1 million of the most partisan American users on Facebook navigated the site over a six-month period last year, researchers found that people's networks of friends and the stories they see are in fact skewed toward their ideological preferences. But that effect is more limited than the worst case some theorists had predicted, in which people would see almost no information from the other side.

On average, about 23 percent of users' friends are of an opposing political affiliation, according to the study. An average of almost 29 percent of the news stories displayed by Facebook's News Feed also appear to present views that conflict with the user's own ideology.

In addition, researchers said that individuals' choices about which stories to click on had a larger effect than Facebook's filtering mechanism in determining whether people encountered news that conflicted with their professed ideology.

"This is the first time we've been able to quantify these effects," said Eytan Bakshy, a data scientist at Facebook who led the study. He said that he began the work in 2012 out of an interest in the way social networks shape how the public gets news. "You would think that if there was an echo chamber, you would not be exposed to any conflicting information," he added, "but that's not the case here."

Facebook's findings run counter to a longstanding worry about the potential for digital filtering systems to shape our world. For Facebook, the focus is on the algorithm that the company uses to decide which posts people see, and which they do not, in its News Feed.

Eli Pariser, chief executive of the viral content website Upworthy, labeled this effect the "Filter Bubble." Some Facebook users have said they unfollow those who post content with which they disagree. And with political discussions being increasingly pitched in the run-up to next year's presidential election, in which the Internet will be used as a primary campaign tool, the problem appeared to be getting worse.

"This shows that the effects that I wrote about exist and are significant, but they're smaller than I would have guessed," Mr. Pariser said in an interview about Facebook's study.

Natalie Jomini Stroud, a professor of communications studies at the University of Texas at Austin, who was not involved in the study, said the results were "an important corrective" to the conventional wisdom.

The study adds to others that debate whether the Internet creates an echo chamber. A Pew Research Center report last year found that media outlets people name as prime information sources about politics and news are strongly correlated with their political views. Another study last year published as a working paper in the National Bureau of Economic Research analyzed Twitter usage during the 2012 election and found social media often exposed users only to opinions that match their own.

Dr. Stroud and other researchers note that Facebook's study has limitations. It arrived at 10.1 million users by screening only for Americans older than 18 who log on to Facebook at least four out of seven days of the week, and who interacted with at least one news link during the second half of 2014. Importantly, all self-identified as liberal or conservative in their profiles. Most Facebook users do not post their political views, and Dr. Stroud cautioned those users might be more or less accepting of conflicting political views.

Criticism of the study was swift. In an article responding to the study, Zeynep Tufekci, a professor at the University of North Carolina, Chapel Hill, said, “People who self-identify their politics are almost certainly going to behave quite differently, on average, than people who do not.” She added, “The study is still interesting, and important, but it is not a study that can generalize to Facebook users.”

Facebook’s researchers said studying those who self-report their politics was the most technically feasible way to determine users’ political affiliations, and trying to infer those from other methods would have led to greater uncertainty.

The findings are convenient for Facebook. With more than 1.3 billion users, the social network is effectively the world’s most widely read daily newspaper. About 30 percent of American adults get their news from Facebook, according to the Pew Research Center. But its editorial decisions are drafted with little transparency using the News Feed algorithm. Facebook could use the study’s results to show that the algorithm is not ruining national discourse.

Facebook said its researchers had wide latitude to pursue their research interests and to present whatever they found. The results were reviewed before publication in Science, with the journal selecting an anonymous panel of scholars unaffiliated with Facebook. Science does not disclose the identity of experts and warns reviewers to declare any financial ties that might be perceived as a conflict of interest with the study being reviewed.

Facebook also noted that this study was substantively different from one that caused an outcry last year, in which the company’s scientists altered the number of positive and negative posts that some people saw to examine the effects on mood. This study did not involve an experiment that changed users’ experience of Facebook; researchers analyzed how people use Facebook as it stands today.

For Facebook’s study, researchers first determined the point of view of a given article by looking at whether liberals or conservatives had shared it most. They found unsurprising partisan attitudes about some news sources, with Fox News stories shared mainly by conservatives and Huffington Post articles shared by liberals.

Then they measured how often feeds of users, whose identifying details had been taken out, displayed stories that conflicted with their professed ideologies, and how often they clicked on those stories.

Some academics said Facebook was always tweaking the News Feed and could easily make changes that would create a more sealed echo chamber.

“A small effect today might become a large effect tomorrow,” David Lazer, a political scientist at Northeastern University who studies social networks, wrote in a commentary on the Facebook study also published in Science. “The deliberative sky is not yet falling, but the skies are not completely clear either.”

The study — also written by Solomon Messing and Lada Adamic, who are data scientists at Facebook — presented other findings on how people of different political persuasions use the world’s largest social network.

One is that liberals live in a more tightly sealed echo chamber than conservatives, though conservatives were more selective about what they click on when they see ideologically challenging views.

About 22 percent of the news stories that Facebook presents to liberals is of a conservative bent, while 33 percent of the stories shown to conservatives presents a liberal point of view. The difference, researchers said, is that liberal users are connected to fewer friends who share views from the other side.

But liberals were only 6 percent less likely to click on ideologically challenging articles instead of ideologically consistent articles that appeared in their feed. Conservatives were 17 percent less likely to click, meaning they appeared more reluctant to indulge opposing views.

The study also raised — but did not answer — the question of what happens after people click on an article with an opposing view: Are they being persuaded by its arguments, or are they dismissing it out of hand?

“People who are really into politics expose themselves to everything,” said Diana C. Mutz, a political scientist at the University of Pennsylvania. “So they will expose to the other side, but it could be to make fun of it, or to know what they're saying to better argue against it, or just to yell at the television set."

A click, in other words, is not necessarily an endorsement, or even a sign of an open mind.

A version of this article appears in print on May 8, 2015, on page A1 of the New York edition with the headline: Facebook Use Polarizing? Site Begs to Differ .

The New York Times | http://nyti.ms/1ysbiMl

Edited by David Leonhardt

Follow Us:

Get the Upshot in your Inbox

The Upshot

RISK FACTOR

Income Inequality: It’s Also Bad for Your Health

MARCH 30, 2015

Margot Sanger-Katz

We know that living in a poor community makes you less likely to live a long life. New evidence suggests that living in a community with high income inequality also seems to be bad for your health.

A study from researchers at the University of Wisconsin Population Health Institute examined a series of risk factors that help explain the health (or sickness) of counties in the United States. In addition to the suspects you might expect — a high smoking rate, a lot of violent crime — the researchers found that people in unequal communities were more likely to die before the age of 75 than people in more equal communities, even if the average incomes were the same.

“It’s not just the level of income in a community that matters — it’s also how income is distributed,” said Bridget Catlin, the co-director of the project, called the County Health Rankings and Roadmaps.

Many factors besides inequality affect health, of course. How much people smoke, whether they are obese, and the safety of air and water, among many other factors, make a difference. But the effect of inequality was statistically significant, equivalent to a difference of about 11 days of life between high- and low-inequality places. The differences were small, but for every increment that a community became more unequal, the proportion of residents dying before the age of 75 went up.

(An aside: The entire County Health Rankings and Roadmap project is full of interesting maps, data and observations about the differences in risk factors and health in various counties. It is well worth a browse.)

The research on inequality at the county level is new, but existing literature suggests there are relationships between income inequality and life expectancy among countries in the world. “Inequality effects, over and above average income, are pretty well established,” said S.V. Subramanian, a professor of population health and geography at Harvard, who has studied the phenomenon. We know that inequality tends to concentrate income in fewer hands, creating more low-income households — and people in low-income households don’t live as long. But what causes the drop in life expectancy is debatable.

One theory is that while money does tend to buy better health, it makes a bigger difference for people low on the income scale than those at the top. That means that having fewer very poor people in a community will improve average health more than having fewer very rich people will diminish it.

But another, more sociological theory, has to do with the communities themselves. The researchers think that places where wealthy residents can essentially buy their way out of social services may have less cohesion and investment in things like education and public health that we know affect life span. There is also literature suggesting that it’s stressful to live among people who are wealthier than you. That stress may translate into mental health problems or cardiac disease for lower-income residents of unequal places.

The researchers measured inequality by comparing the number of people in a given place who earned above the 80th percentile in the county with the number of people earning less than the 20th percentile. Then they measured life expectancy using a custom measurement they developed — it counts the “potential life years lost” in each community by measuring all those who died before the age of 75, and the age at which they died. So someone who died at age 70 would have five years of potential life lost. Then they adjusted the numbers according to how old people were in the county, so counties with more old people wouldn’t look sicker than counties that were younger. The study looked at only the average life span and not that of higher-income versus lower-income residents.

For every one-point increase in the ratio between high and low earners in a county, there were about five years lost for every 1,000 people. That's about the same difference they observed when a community's smoking rate increased by 4 percent or its obesity rate rose by 3 percent. Researchers said that inequality effect persisted even when they compared communities of similar average income and racial composition.

Here's what that means in some real counties. The researchers compared the adjoining Park and Fremont Counties in Wyoming. Both have relatively small populations and are predominantly white. Both include parts of large national parks. But the Fremont inequality ratio is 4.6, compared with Park's 3.6. And, in Fremont County, there are 13 years of potential life lost for every 1,000 residents, compared with only 7.5 in Park.

We happen to be in the midst of a big experiment in redistribution of spending on health care, through the Affordable Care Act. The law, particularly in states that have opted to expand Medicaid, is now providing substantial income, in the form of health insurance, to Americans at the lower end of the income spectrum. Whether those new resources will ameliorate some of the effects of inequality will be something the researchers will be tracking in the coming years.

The Upshot provides news, analysis and graphics about politics, policy and everyday life. Follow us on Facebook and Twitter. Sign up for our weekly newsletter.

Media Contact:

Bethanne Fox or Toni Williams
Burness
(301) 652-1558
bfox@burness.com
twilliams@burness.com

Melissa Blair
Robert Wood Johnson Foundation
(609) 627-5937
media@rwjf.org

EMBARGOED FOR RELEASE UNTIL
Wednesday, March 25, 2015 at 12:01am ET

County Health Rankings Show Declines in Premature Death Rates
But Where You Live Matters to Your Health

Princeton, N.J. and Madison, Wis.—The 2015 *County Health Rankings* released today, show that premature deaths are dropping, with 60 percent of the nation's counties seeing declines. For instance, in the District of Columbia premature death rates have plummeted by nearly one-third based on data from 2004-2006 and 2010-2012. This marks the highest drop in the country for counties with populations of 65,000 or more. But for many counties these rates are not improving—forty percent of counties are not making progress in reducing premature deaths.

A rich resource of local-level data, the *Rankings* are an easy-to-use snapshot comparing the health of nearly every county in the nation. A collaboration between the Robert Wood Johnson Foundation (RWJF) and the University of Wisconsin Population Health Institute, the *Rankings* allow each state to see how its counties compare on 30 factors that impact health, including education, transportation, housing, violent crime, jobs, diet and exercise. The *Rankings* are available at www.countyhealthrankings.org.

This year's *Rankings* show that almost one out of four children in the U.S. lives in poverty. Child poverty rates are more than twice as high in the unhealthiest counties in each state than in the healthiest counties. The report also looks at distribution in income and the links between income levels and health.

"The *County Health Rankings* have helped galvanize communities across the nation to improve health," said Risa Lavizzo-Mourey, MD, RWJF president and CEO. "Solutions and innovation are coming from places as diverse as rural Williamson, West Virginia in the heart of Appalachia to urban New Orleans;

they are engaging business, public health, education, parents, and young people to build a Culture of Health."

Beyond poverty and income, the 2015 *County Health Rankings* Key Findings Report highlights two other key social and economic factors that drive health: violent crime and employment. These findings show that:

- **Violent Crime Rates are Highest in the South:** Violent crime rates, which affect health, well-being, and stress levels, are highest in the Southwest, Southeast, and Mississippi Delta regions.

- **Having a Job Influences Health:** Unemployment rates are 1.5 times higher in the least healthy counties in each state as they are in the healthiest counties. During the recession, counties in the West, Southeast, and rust belt region of the U.S. were hit hardest by growing unemployment. Many, but not all, of these counties have seen their unemployment rates drop since the recession ended in 2010.

This year's *Rankings* data also shines a light on the characteristics of healthy and unhealthy counties. The healthiest counties in each state have higher college attendance, fewer preventable hospital stays, and better access to parks and gyms. The least healthy counties in each state have more smokers, more teen births, and more alcohol related car crashes.

"In the six years since the *County Health Rankings* began, we've seen them serve as a rallying point for change. Communities are using the *Rankings* to inform their priorities as they work to build a Culture of Health," said Bridget Catlin, PhD, MHSA, co-director of the *County Health Rankings.*

The *County Health Rankings & Roadmaps* program offers data, tools, and resources to help communities throughout their journey to build a Culture of Health. Also part of the program is the *RWJF Culture of Health Prize* which honors communities that are working together to build a healthier, more vibrant community.

###

About the Robert Wood Johnson Foundation
For more than 40 years the Robert Wood Johnson Foundation has worked to improve the health and health care of all Americans. We are striving to build a national Culture of Health that will enable all to live longer, healthier lives now and for generations to come. For more information, visit *www.rwjf.org*. Follow the Foundation on Twitter at *www.rwjf.org/twitter* or on Facebook at *www.rwjf.org/facebook*.

About the University of Wisconsin Population Health Institute
The University of Wisconsin Population Health Institute advances health and well-being for all by developing and evaluating interventions and promoting evidence-based approaches to policy and practice at the local, state, and national levels. The Institute works across the full spectrum of factors that contribute to health. A focal point for health and health care dialogue within the University of

Wisconsin-Madison and beyond, and a convener of stakeholders, the Institute promotes an exchange of expertise between those in academia and those in the policy and practice arena. The Institute leads the work on the *County Health Rankings & Roadmaps* and manages the *RWJF Culture of Health Prize.* For more information, visit *http://uwphi.pophealth.wisc.edu*.

5. 5.

Sections Search

[Climate](#) |It's Not Your Imagination. Summers Are Getting Hotter.

- Share
- Tweet
- Email
- More
- Save

Account Log In 170 0 Settings

Close search

Site Search Navigation

Search NYTimes.com

Clear this text input Go

https://nyti.ms/2u3NOkX

Site Navigation

- Home Page
- World
- U.S.
- Politic
- N.Y.
- Business
- Business
- Opinion
- Opinion
- Tech
- Scienc
- Health
- Sports
- Sports

- Arts
- Arts
- Books
- Style
- Style
- Food
- Food
- Travel
- Magazine
- T Magazine
- Real Estate
- Obitua
- Video
- The U
- Reade
- Conferences

- Crossword
- Times Insider
- Newsletters
- The L work

- Multi
- Photog
- Podcasts

- NYT Store
- NYT Wine Club
- nytEducation
- Times Journeys
- Meal Kits

[We Went to the Moon. Why Can't We Solve Climate Change?](#)

10\.

[E.P.A. Won't Ban Chlorpyrifos, Pesticide Tied to Children's Health Problems](#)

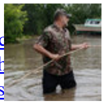
15\.

[Heat Waves in the Age of Climate Change: Longer, More Frequent and More Dangerous](#)

20\.

[Fewer Inspections for Aging Nuclear Plants, Regulators Propose](#)

25\.



[‘Toxic Stew’ Stirred Up by Disasters Poses Long-Term Danger, New Findings Show](#)

30\.

[New York Awards Offshore Wind Contracts in Bid to Reduce Emissions](#)

35\.

[House Passes Intelligence Bill That Would Expand Secrecy Around Operatives](#)

40\.

[Climate Fwd:](#)

[One Thing You Can Do: Know Your Tree Facts](#)

45\.

- Subscribe
- Manage Account
- Today's Paper
- Tools & Services
- Jobs
- Class
- Corre
- More

5.

We Went to the Moon. Why Can't We Solve Climate Change?

Site Mobile Navigation

Advertisement

10.

170

Climate

- Share

E.P.A. Won't Ban Chlorpyrifos, Pesticide Tied to Children's Health Problems

15.

It's Not Your Imagination. Summers Are Getting Hotter.

By NADJA POPOVICH and ADAM PEARCE JULY 28, 2017

Heat Waves in the Age of Climate Change: Longer, More Frequent and More Dangerous

Summer temperatures
in the Northern Hemisphere

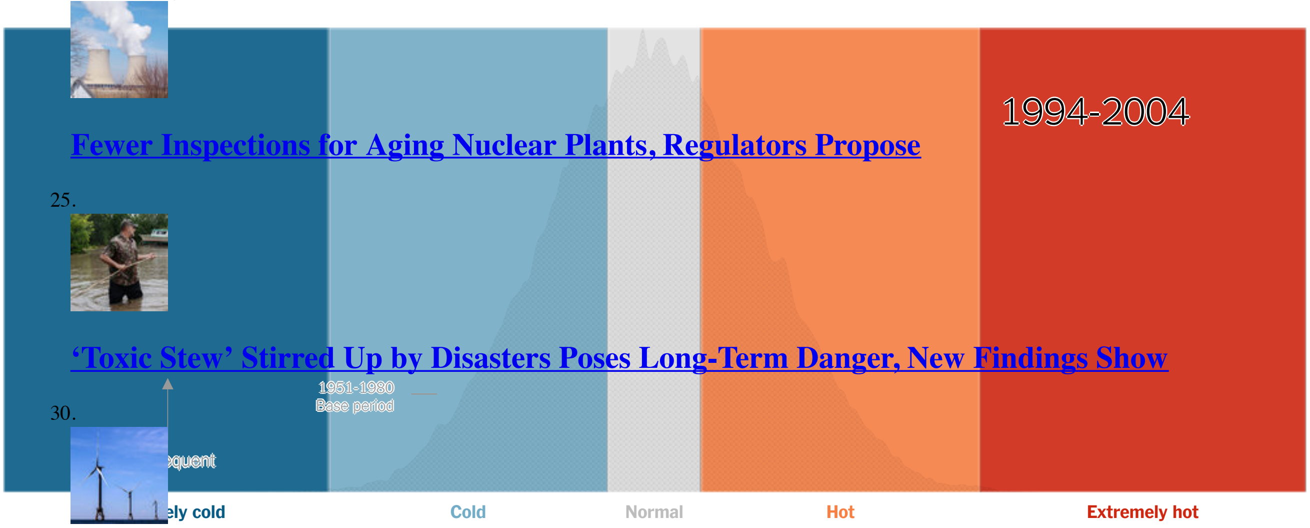
20.

Fewer Inspections for Aging Nuclear Plants, Regulators Propose

25.

'Toxic Stew' Stirred Up by Disasters Poses Long-Term Danger, New Findings Show

30.



New York Awards Offshore Wind Contracts in Bid to Reduce Emissions

35.

Extraordinarily hot summers — the kind that were virtually unheard-of in the 1950s — have become commonplace.

This year's scorching summer events, like heat waves rolling through southern Europe and temperatures nearing 130 degrees Fahrenheit in Pakistan, are part of this broader trend.

House Passes Intelligence Bill That Would Expand Secrecy Around Operatives

40.

The chart above, based on data from James Hansen, a retired NASA climate scientist and professor at Columbia University, shows how summer temperatures have shifted toward more extreme heat over the past several decades.

Climate Fwd:

One Thing You Can Do: Know Your Tree Facts

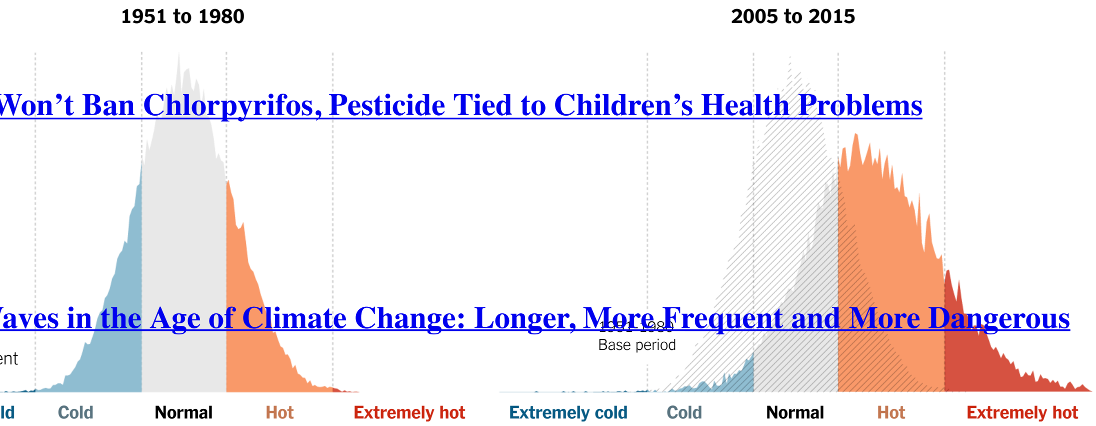
45.

To create the bell curves, Dr. Hansen and two colleagues compared actual summer temperatures for each decade since the 1980s to a fixed baseline average. During the base period, 1951 to 1980, about a third of local summer temperatures across the Northern

Hemisphere were in what they called a “near average” or normal range. A third were considered cold; a third were hot.

Since then, summer temperatures have shifted drastically, the researchers found. Between 2005 and 2015, two-thirds of values were in the hot category, and nearly 15 percent were in a new category: extremely hot.

Practically, that means most summers are now either hot or extremely hot compared with the mid-20th century.



The big increase in summer temperatures under the dark red category of extreme heat is “right in line” with what scientists expect to see as the climate warms overall, said Todd Sanford, director of research at Climate Central, a nonprofit science and news organization.

For each time period above, the distribution of summer temperatures forms what is known as a bell curve because most measurements fall near the average, forming the bump – or bell – in the middle. More extreme temperatures, which happen less frequently, fall in the wings, with heat waves on the right and cold-snaps on the left.

As the curve’s average – the top of the peak – shifts rightward over time, more temperatures in more places end up in the hot and extremely hot categories and fewer end up in the cold category.

Dr. Hansen’s curves also flatten out, which some have suggested is an indication of greater temperature variability. But other climate scientists, including Zeke Hausfather, an energy systems analyst at the University of California, Berkeley, have pointed out that this effect is mainly a reflection that some parts of the world are warming faster than others. There is no evidence that temperatures are becoming more variable in most parts of the world after warming has been accounted for.

Dr. Hansen’s data “really highlight that changes in the average, while they may seem modest, have big implications for the extremes. And that’s what’s going to affect society and ecosystems,” Dr. Sanford said. The findings reveal what has happened so far, and also provide “a glimpse to what’s in our

5. 5.

We Went to the Moon. Why Can’t We Solve Climate Change?

10.

E.P.A. Won’t Ban Chlorpyrifos, Pesticide Tied to Children’s Health Problems

15.

Heat Waves in the Age of Climate Change: Longer, More Frequent and More Dangerous

20.

Fewer Inspections for Aging Nuclear Plants, Regulators Propose

25.

‘Toxic Stew’ Stirred Up by Disasters Poses Long-Term Danger, New Findings Show

30.

New York Awards Offshore Wind Contracts in Bid to Reduce Emissions

35.

House Passes Intelligence Bill That Would Expand Secrecy Around Operatives

40.

Climate Fwd:

One Thing You Can Do: Know Your Tree Facts

45.

future.”

Source: Columbia University Earth Institute. Data via Makiko Sato and James Hansen. Based on: Hansen et al., 2012 (and discussion); 2016 update.

5. 5.

Temperatur s are determined by the normal distribution, so that a third of temperatures fall in each of the main three categories: hot, cold and normal for 1951 to 198 temperatures for each subsequent 11-year period are compared to the 1951 to 1980 baseline.

- Emai
- Share
- Tweet
- More

We Went to the Moon. Why Can’t We Solve Climate Change?

More on NYTimes.com

10.

E.P.A. Won’t Ban Chlorpyrifos, Pesticide Tied to Children’s Health Problems

15.

Tru savows ‘Send Her Back’ Chant as G.O.P. Frets Over Ugly Phrase

Jul. 18, 2019

Heat Waves in the Age of Climate Change: Longer, More Frequent and More Dangerous

Fact Check

Fewer Inspections for Aging Nuclear Plants, Regulators Propose

Examining Trump’s Claims About Representative Ilhan Omar

25.

Jul. 1

‘Toxic Stew’ Stirred Up by Disasters Poses Long-Term Danger, New Findings Show

30.

An emocratic Governors Urge 2020 Field Not to Veer Too Far Left

Jul. 19, 2019

New York Awards Offshore Wind Contracts in Bid to Reduce Emissions

35.

Iran Denies That U.S. Downed Its Drone in Strait of Hormuz

House Passes Intelligence Bill That Would Expand Secrecy Around Operatives

Jul. 19, 2019

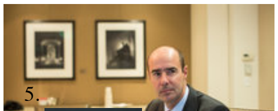
40.

Climate Fwd:

One Thing You Can Do: Know Your Tree Facts

45.

5.

- 

[Trump to Nominate Eugene Scalia for Labor Secretary Job]

Jul. 18, 2019

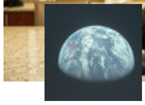
Site Inf on Navigation

- © 2019 The New York Times Company
- Home
- Search
- Accessibility concerns? Email us at accessibility@nytimes.com. We would love to hear from you.
- Cont
- Work
- Adve
- Your s
- Privacy
- Terms of Service
- Terms of Sale

Site Information Navigation

- Site
- Help
- Site Feedback
- Subscriptions

Go to the previous story
Go to the next story

5. 

We Went to the Moon. Why Can't We Solve Climate Change?

10.

E.P.A. Won't Ban Chlorpyrifos, Pesticide Tied to Children's Health Problems

15.

Heat Waves in the Age of Climate Change: Longer, More Frequent and More Dangerous

20.

Fewer Inspections for Aging Nuclear Plants, Regulators Propose

25.

'Toxic Stew' Stirred Up by Disasters Poses Long-Term Danger, New Findings Show

30.

New York Awards Offshore Wind Contracts in Bid to Reduce Emissions

35.

House Passes Intelligence Bill That Would Expand Secrecy Around Operatives

40.

Climate Fwd:

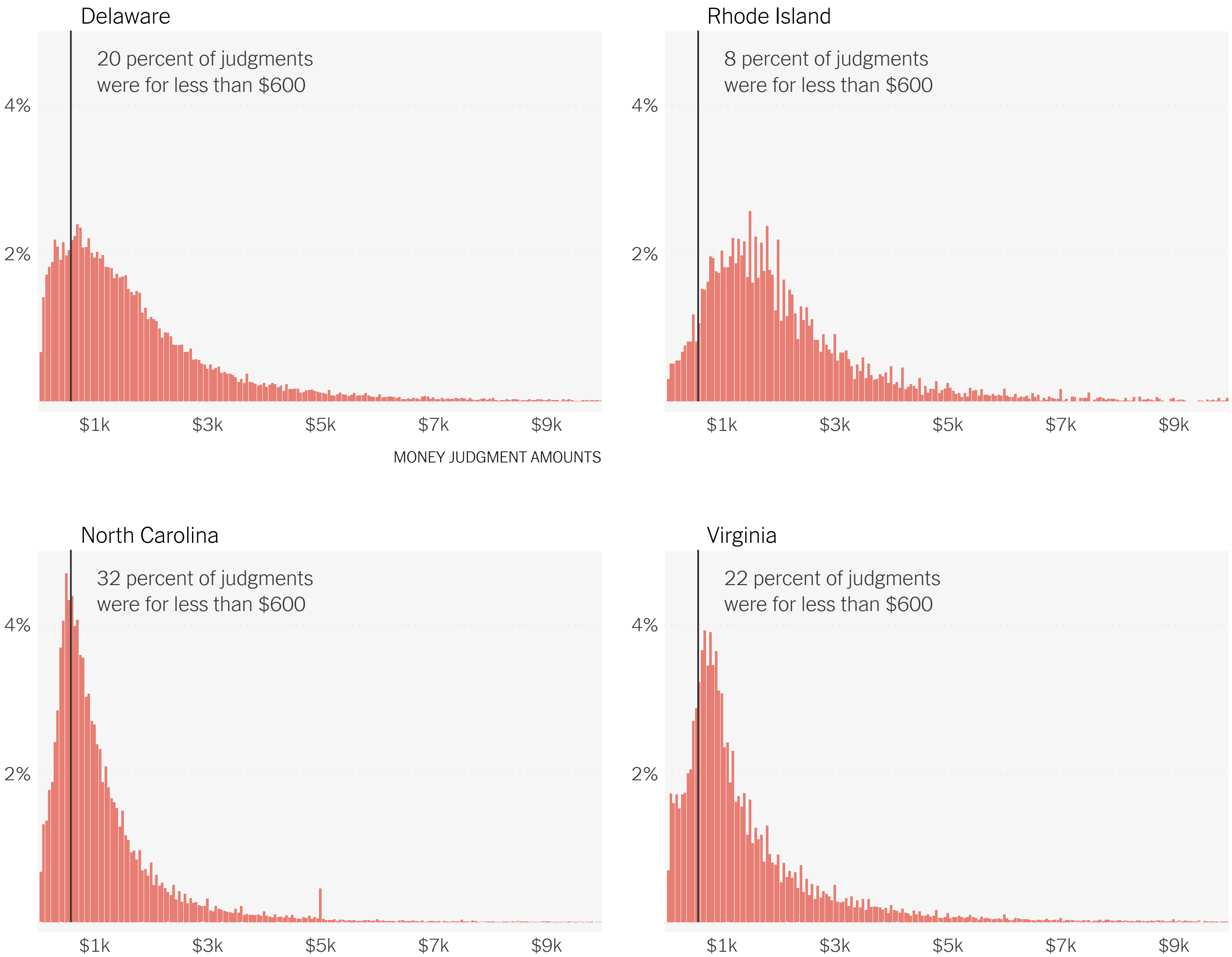
One Thing You Can Do: Know Your Tree Facts

45.

The New York Times

Many Renters Who Face Eviction Owe Less Than $600

Can Washington do something to help them? A growing number of politicians think so.



Only eviction filings that resulted in a money judgment are shown here. Data were collected by LexisNexis Risk Solutions and compiled and verified by the Eviction Lab at Princeton University. State data excludes counties where records couldn't be compared to a second source. Data from 2014 to 2016. • Quoctrung Bui/The New York Times

By Emily Badger

Dec. 12, 2019, 5:00 a.m. ET

Among the millions of recent eviction cases researchers have begun to compile across the country, there are a startling number of modest sums. There are dozens of families in Texas evicted with money judgments — unpaid rent, late fees, court costs — totaling $516. There are multiple families in Cumberland County, N.C., who owed all of $301. There is a household in Providence, R.I., whose 2016 court record shows a debt of just $127.

Such relatively small sums suggest that, for all of the intractable problems of poverty and affordable housing driving the nation's eviction crisis, a little intervention could help many people. And politicians in Washington increasingly have such ideas in mind: court translators, more legal aid, mediation — even emergency rent assistance.

One bill, to be introduced in the Senate on Thursday by a Democrat, Michael Bennet of Colorado, and a Republican, Rob Portman of Ohio, would create a federal grant program to fund local emergency aid for tenants at risk of eviction. The bill, which would also establish a national database tracking eviction cases, is the latest in a series of federal proposals aimed at a problem that touches high-cost coastal cities and smaller towns alike.

Several Democratic senators — Maggie Hassan of New Hampshire, Tim Kaine of Virginia and Chris Van Hollen of Maryland — introduced a bill this fall that would create federal grants for landlord-tenant mediation programs and translators. In the House, Alexandria Ocasio-Cortez has introduced a bill that would fund legal aid in states and cities that establish a right to counsel for tenants that is akin to a new mandate in New York City.

And in the Democratic primary, an anti-eviction agenda is now practically a required element of candidates' housing plans. Bernie Sanders supports a national "just cause" standard, limiting the grounds on which a landlord can evict a tenant. Cory Booker wants to prevent consumer reporting agencies from listing eviction cases won by the tenant. Amy Klobuchar wants to create new kinds of savings accounts that renters could tap in an emergency.

Such strategies most likely would not address the structural problems of sluggish wage growth and a scarcity of low-cost housing that underlie the eviction crisis. But they imply that even if eviction is a necessary remedy for landlords, perhaps there could be less of it.

"Sometimes you hear this response from property owners who say, 'What do I do if a tenant is behind five, six, seven months?' And that's a really important question," said Matthew Desmond, a sociologist at Princeton whose eviction research has influenced politicians. In data his Eviction Lab has analyzed from 22 states, that situation of tenants deep in debt is rare. It's far more common, the lab has found, that tenants owe the equivalent of less than a month's rent.

"That suggests a shallower end of the risk pool that you could lop off," Mr. Desmond said.

Failure to pay rent is by far the most common cause of eviction. Across the 22 states where the Eviction Lab has data covering at least some counties, the median money judgment for cases between 2014 and 2016 was $1,253. Because that total includes other costs accumulated during the court process, most tenants initially faced eviction for the failure to pay a smaller rent sum (which is often not documented in eviction records).

Nearly half of the money judgments the lab has analyzed in Virginia were for less than the median rent in the census tract where the eviction occurred. In North Carolina, that's true in more than 40 percent of cases. Across the 22 states, about a third of money judgments were for less than the local median rent.

"That third of the people that can't come up with a month's rent, finding a way to stabilize their housing situation will actually be less costly than the remedy of throwing them out in to the street — for everybody," said Mr. Bennet, who is also running for president. "Today, the externalities of that third of people are invisible to society."

The public doesn't see the cost of homeless shelters, or the cost to landlords of having to find new tenants, or the costs when newly homeless children change schools and classrooms are disrupted, said Mr. Bennet, who was previously the superintendent of the Denver Public Schools.

Many of these proposals are effectively trying to slow down the eviction process, or to create more alternatives to it.

"The court system has been set up so that the choice for the landlord is really eviction or nothing," Ms. Hassan said.

Her bill would also charge the Department of Housing and Urban Development with studying insurance models that landlords or tenants could buy into to guard against the possibility of a missed rent payment.

"Most families in poverty that are renting are spending half their income on rent, with no margin for error," said Mike Koprowski, who leads a network of advocacy groups through the National Low-Income Housing Coalition that has pushed for emergency assistance grants. "There's no margin for the broken-down car, the unreimbursed medical bill, the hours cut at your job."

Landlord associations respond that such precariousness is the real problem. Wages have stagnated for the poor, and the supply of housing affordable to the poorest renters has dwindled. Between 1990 and 2017, the national stock of rental housing grew by 10.9 million units, according to the Harvard Joint Center for Housing Studies. Over that same time, the number of units renting for less than $600 a month in inflation-adjusted dollars fell by 4 million. All net growth in rental housing in America, in other words, has been for higher-income tenants.

“We regard evictions generally as a symptom of a larger problem that we have, which is a lack of housing that’s affordable,” said Greg Brown, the senior vice president of government affairs for the National Apartment Association. “The question we have to ask ourselves is, do changes in the process address that core issue, or do they lengthen a process of eviction and really end up in the same place?”

One recent study linking eviction records in Cook County, Ill., with credit report and payday loans data suggests that policy interventions to the court process itself may be too late to help many poor families. The study found that in the years leading up to an eviction filing, tenants who would ultimately wind up in court had mounting and substantially higher debts, compared with random tenants in the same neighborhoods.

“The signs of that disruption and financial distress appear two to four years ahead of the eviction filings,” said Winnie van Dijk, one of the study’s authors. For policymakers who want to help these families, she said, “it’s not as simple as avoiding a court order for eviction, unfortunately.”

The Bennet-Portman bill envisions some outcomes where a tenant might lose housing they can’t afford but still land in a better place. The bill would also create a grant program to support community courts that might, for instance, be tied directly to local providers of social services. The Cleveland Municipal Housing Court currently operates a similar model, where the most vulnerable tenants facing eviction are flagged for social services. In 2018, about 17 percent of eviction cases that came before the court were referred to case workers who tried to find programs like mental health support and homeless services for tenants.

Sherrae Landrum, 74, was summoned to the Cleveland court last December for accumulating clutter that her landlord objected to. She ultimately lost her home, but a social worker the court connected her to helped her find a temporary homeless shelter and then a permanent home. The social worker drove her to doctor’s appointments and accompanied her to pay the deposit on her new apartment. She helped enroll Ms. Landrum in a home health aide service, a meal-delivery program, and a class to manage hoarding tendencies.

Eviction court can present, at least, an opportunity to connect tenants with agencies that might help halt cycles of poverty, said Casey Albitz, who runs the court's social service referral program.

"It was a blessing," Ms. Landrum said of her case. She has more resources and support now than she ever did before. "They did me a favor."

Emily Badger writes about cities and urban policy for The Upshot from the Washington bureau. She's particularly interested in housing, transportation and inequality — and how they're all connected. She joined The Times in 2016 from The Washington Post. @emilymbadger

READ 5 COMMENTS

The Washington Post

Wonkblog • Analysis

Not only are Americans becoming less happy — we're experiencing more pain too

By Christopher Ingraham December 6 at 6:00 AM

Americans "are in greater pain than citizens of other countries" and have been growing steadily more miserable for decades, according to a new working paper by David G. Blanchflower of Dartmouth College and Andrew Oswald of the University of Warwick.

For their paper, Blanchflower and Oswald investigate claims about happiness made by the Brookings Institution's Carol Graham in her recent book, "Happiness For All?". In the book, Graham draws primarily on Gallup data to argue American happiness is faltering as a rational response to growing inequality.

Among Graham's most striking finding is, as she puts it, "markers of well and ill-being, ranging from life satisfaction to stress, are more unequally shared across the rich and the poor in the U.S. than they are in Latin America, a region long known for high levels of inequality." Low-income Americans are particularly skeptical that hard work will improve their economic situation.

Blanchflower and Oswald wanted to see if other data sources corroborated Graham's findings. They first turn to the General Social Survey (GSS), a nationally-representative survey administered every several years and used frequently in social science work. The GSS data shows, unambiguously, that Americans' evaluations of their own happiness has been falling in recent years.

The decline is plainly visible across multiple demographic groups. Declines have been steepest among Americans with the least education, and the happiness gap between the most-educated and least-educated Americans has nearly doubled since 1972.

Blanchflower and Oswald note the GSS data shows similar trends for Americans' feelings about their finances — everyone is less optimistic about money relative to 1972, but optimism has dropped particularly sharply among the least-educated. They call these disparities a form of "psychological inequality," which is both a reflection of actual monetary inequality and a driver of it — after all, it's difficult to improve your financial situation if you don't believe your financial situation can be improved.

As another marker of psychological distress, Blanchflower and Oswald look at cross-country data on the experience of pain. In 2011, the International Social Survey Programme asked respondents in over 30 nations how often they had experienced bodily aches and pains in the past month.

Americans were the most likely to report frequent pain, with 34 percent saying they experienced it "often" or "very often." The average across all countries surveyed was just 20 percent.

"As the US is one of the richest countries in the world, and in principle might be expected to have one of the most comfortable lifestyles in the world, it seems strange — to put it at its mildest — that the nation should report such a lot of pain," Blanchflower and Oswald write.

Aware that some of this could be attributable to question translation issues or cultural differences (for instance, Americans may just be more predisposed to complain about pain than members of other nations), the authors ran the numbers controlling for age, gender, marital status, labor force status and education. The United States remained an outlier even when these factors were accounted for.

The nation's relatively stingy social safety net may be one factor contributing to this exceptionalism. Many Americans still lack access to health care, which is available universally in most other wealthy nations. The expense of health care, even for those who have insurance, could mean Americans experiencing aches and pains are more likely to tough it out and forego treatment, relative to people in other countries.

In the United States, health issues remain a major contributor to financial insecurity, meaning they likely contribute to some of the declining happiness and financial pessimism seen in the other research surveyed by Blanchflower and Oswald.

All told, the data underscore how country-level material wealth can be a poor indicator for the well being of its inhabitants.

41 Comments

Christopher Ingraham writes about politics, drug policy and all things data. He previously worked at the Brookings Institution and the Pew Research Center.

The New York Times | http://nyti.ms/1IFIWOo

The Opinion Pages | **CAMPAIGN STOPS**

Our Insane Addiction to Polls

Frank Bruni JAN. 23, 2016

REMEMBER the poll last week that had Bernie Sanders ahead of Hillary Clinton in New Hampshire by three points?

No, you're thinking. I've got it wrong. Sanders was up by 27.

That's true, if you're talking about the figures that CNN and WMUR released on Tuesday. I'm talking about the ones that Gravis Marketing and One America News Network released on *Wednesday*.

There were three polls of New Hampshire voters over just two days last week, according to the archive maintained by Real Clear Politics. There were three polls of Iowa voters on Thursday alone. One had Clinton up by eight, while another had Sanders up by that same margin. One had Donald Trump up by 11. Another had Ted Cruz up by two.

Over a monthlong period ending Thursday night — a monthlong period, mind you, that included the Christmas and New Year's break — there were 11 polls in Iowa, 10 in New Hampshire and nine nationally. There were polls focused on 10 different states.

And their findings were often treated as breathless news. On Wednesday evening, I visited the home page of the Politico website — I'm using Politico as a random example — and spotted four stories that were essentially about poll results.

My television was on: The CNN anchor Erin Burnett began her show with a report on the latest poll.

I'd say that we're in a period of polling bloat, but bloat is too wan a word. Where polling and the media's attention to it are concerned, we're gorging ourselves into a state of morbid obesity.

"I've never seen anything like it," said Ralph Reed, a longtime Republican strategist who thought that things were bad in 2008 and 2012 and realizes now that those were days of temperance and innocence, to be pined for and perhaps never savored again.

But it's not the crazy bounty of polls that fascinates him (and me) most. It's something else.

"There seems to be an inverse relationship between the preponderance of polling and the *reliability* of polling," Reed said, nailing one of the most illogical, paradoxical dynamics of the 2016 election so far.

We're leaning harder than ever on polling precisely when that makes the least sense. We're wallowing in polls even as they come to wildly different conclusions that should give us serious pause.

Good polls are arguably more difficult than ever to do, for reasons I'll go into later. And from 2012 forward, there have been prominent examples of how poorly they sometimes predict outcomes. They botched the most recent parliamentary elections in Britain and Israel. They botched the 2014 midterms in the United States, grossly miscalculating the margins in various congressional and gubernatorial races.

Last year, Reed noted, the Real Clear Politics average of polls in the Kentucky governor's race had the Democrat, Jack Conway, ahead of the Republican, Matt Bevin, by five points.

Bevin won by nearly nine.

"It was 14 points off," Reed marveled. "But everybody shrugs and moves on down the highway."

There are explanations for those shrugs, and they speak to the quirks and flaws of political journalism in a wired, revved-up world.

There are consequences, too. An obsession with polls and a quickness to weave narratives around them bolster certain candidates and retard others, and could well affect the outcome of this presidential election.

If Donald Trump wins the Republican nomination — or, heaven forbid, the White House — it will be partly because we in the media justified saturation coverage of him by pointing to polls, which in turn legitimized *his* fixation on them as proof that he's up to the job: He must be, because plenty of people apparently picture him in it.

"He *caresses* his polls numbers," said David Axelrod, one of the chief architects of Obama's 2008 and 2012 presidential campaigns, adding that Trump's first order of business when he steps up to a microphone is "a recitation of poll numbers. He's like a Lothario recounting his exploits every time he starts a speech."

If Jeb Bush's candidacy comes to naught, his underwhelming poll numbers — and how they were used to cast him instantly as an underachiever — will have been a factor. And if Clinton fails to win the nomination, the media's embrace of certain polls among an ever-changing riot of them will have played at least some small role.

"If she's three points behind in New Hampshire, it's a close race," Axelrod said. "But if she's 27 points behind, her campaign's in free fall. That's a sexier story and the one that's chosen. It becomes the meme. It becomes the prism through which everyone filters their coverage. It skews how people view everything that a candidate does: Is it conviction or desperation?"

Desperation makes a better story. So the media dwells on the most pessimistic projections, ensuring that polling, no matter how divorced from reality, shapes it.

Polls determined which Republican candidates participated in which debates, although just a couple of percentage points — the margin of error, really — separated a few of the prime-time debaters from the early birds.

Some of these polls were national ones, which have minimal relevance to the decisive, trajectory-setting contests in Iowa, New Hampshire and South Carolina.

"A national poll is absolutely meaningless," said Stuart Stevens, the chief strategist for Romney's 2012 campaign. "One of every nine Americans lives in California. So one of every nine voters in that poll is going to be in California. When's the last time anybody read a story about the Republican primary in California?"

And if you dig below the surface of these national polls — or of polls in the states with the first contests — you find a crucial detail that we in the media blithely gloss over: Many, if not most, voters haven't made up their minds. In last week's CNN/WMUR poll of New Hampshire voters, for example, about one in three Republicans said that they had definitely decided on a candidate.

Part of what I find so jarring about the media's insatiable appetite for polls right now is that it defies our past resolutions to go on a desperately needed diet. For all of my 30 years as a journalist, I've listened to reporters, editors and producers bemoan the "horse-race coverage" of campaigns and exhort one another to be better come the next election and concentrate instead on issues, records, biographies, substance.

Suddenly the apologies and exhortations are gone. We're worse. Every candidate's a thoroughbred, every day the Kentucky Derby and just about every other story on many newscasts and news sites an assessment of his or her odds. We're resigned to treating campaigns as sport. It represents the surest path to a large enough audience to keep a beleaguered industry economically viable.

Some dispatches are overt in their reliance on polls, others less so, using polls as the prompts or context for discussions of stalled campaign momentum or new campaign strategies. The Times is guilty. I'm guilty. Polls are the seemingly irresistible argot of political coverage; to purge them from your vocabulary is to speak in an unrecognizable tongue.

BESIDES which, we need the content. In a fast-metabolism world of constantly monitored smartphones and routinely refreshed browsers, news organizations are under greater pressure to produce fresh nuggets of information even as their budgets and resources for reporting decline. So we readily trumpet and analyze polls, a practice that has even given birth and currency to a whole strain of news that's survey-based, with practitioners who burrow ever deeper into numbers.

"You have a lot of people who don't ever go out on the campaign trail and actually talk to voters," Stevens said. "We now call those people data journalists. They don't have to report. It's a wonderful category."

Our demand for polls guarantees a robust supply of them: Churning these surveys out is great guaranteed publicity for news organizations, research companies and academic institutions. How many Americans are aware of Quinnipiac University in Hamden, Conn., or Monmouth University in West Long Branch, N.J., only because they commission and put their stamps on polls?

But not all of the surveys that the media cites are created equal. Gary Langer, the founder of Langer Research Associates and a former director of polling for ABC News, compared the variable quality of polls to the variable quality of what you eat. "There are well-crafted, delicious meals," he said. "Then there's fast food. And then there's listeria."

Listeria is common. People's reliance on cellphones has complicated polling:

Federal law forbids automated calls to such phones, so pollsters must bear the considerable expense of dialing them by hand or, alternately, make do without them, which can lead to imperfect results.

Beyond that, there's the question of whether the kind of people who consent to polls are true weather vanes.

"The entire industry rests on the idea that the people you get are representative of the people you don't get," said Jon Cohen, who supervised polling for The Washington Post from 2006 to 2013 and is now the vice president of survey research at SurveyMonkey. "I think that's an increasingly questionable premise and one that I keep in mind every time I design a survey."

"People don't enjoy or even tolerate those conversations the way they used to," he added. "Taking a survey is a participatory act, and different kinds of people participate in different things."

Can we extrapolate from the kinds of people who have been professing adoration of Trump — or, for that matter, of Sanders — into the general population? Apart from that, can we trust the constancy of their affections? Will they actually cast votes?

Over the next few weeks, we'll find out, and I promise you this: There will be some surprises. If they're significant enough, polling could be as big a loser in the 2016 primaries as any candidate is. And we in the media will be forced to apologize anew for our poll-mad behavior. Sadly, that doesn't mean we'll change it.

I invite you to follow me on Twitter at twitter.com/frankbruni and join me on Facebook.

Follow The New York Times Opinion section on Facebook and Twitter, and sign up for the Opinion Today newsletter.

A version of this op-ed appears in print on January 24, 2016, on page SR1 of the New York edition with the headline: Our Insane Addiction to Polls.

The Washington Post

Speaking of Science

Researchers replicate just 13 of 21 social science experiments published in top journals

By Joel Achenbach

, Reporter

August 27 at 12:17 PM

The "reproducibility crisis" in science is erupting again. A research project attempted to replicate 21 social science experiments published between 2010 and 2015 in the prestigious journals Science and Nature. Only 13 replication attempts succeeded. The other eight were duds, with no observed effects consistent with the original findings.

The failures do not necessarily mean the original results were erroneous, as the authors of this latest replication effort note. There could have been gremlins of some type in the second try. But the authors also noted that even in the replications that succeeded, the observed effect was on average only about 75 percent as large as the first time around.

The researchers conclude that there is a systematic bias in published findings, "partly due to false positives and partly due to the overestimated effect sizes of true positives."

The two-year replication project, published Monday in the journal Nature Human Behaviour, is likely to roil research institutions and scientific journals that in recent years have grappled with reproducibility issues. The ability to replicate a finding is fundamental to experimental science. This latest project provides a reminder that the publication of a finding in a peer-reviewed journal does not make it true.

Scientists are under attack from ideologues, special interests and conspiracy theorists who reject the evidence-based consensus in such areas as evolution, climate change, the safety of vaccines and cancer treatment. The replication crisis is different; it is largely an in-house problem with experimental design and statistical analysis.

Refreshingly, other scientists have a pretty good detector for which studies are likely to stand the test of time. In this latest effort, the researchers asked more than 200 peers to predict which studies would replicate and to what extent the effect sizes would be duplicated. The prediction market got it remarkably right. The study's authors suggest that scientific journals could tap into the "wisdom of crowds" when deciding how to treat submitted papers with novel results.

"I would have expected results to be more reproducible in these journals," said John Ioannidis, a professor of medicine at Stanford. He was not involved in this new research but is closely associated with the issue of reproducibility because of his authorship of an influential and extraordinarily provocative 2005 article with the headline "Why Most Published Research Findings Are False."

Simine Vazire, a University of California at Davis psychologist who is also active in the reproducibility movement, said the new project's replication success — 10 out of 17 experiments published in Science and 3 out of 4 published in Nature — "is not okay." She said, "There's no reason why the most prestigious journals shouldn't demand pretty strong evidence," and added that these experiments would not have been difficult to attempt to replicate before publication.

One of the studies that didn't replicate attempted to study whether self-reported religiosity would change among test subjects who had first been asked to look at an image of the famous Auguste Rodin sculpture "The Thinker." The study found that people became less religious after exposure to that image.

"Our study in hindsight was outright silly," said Will Gervais, an associate professor of psychology at the University of Kentucky. Gervais said that his original study oversold a "random flip in the data," although other parts of his paper did replicate.

Another experiment, conducted in Boston in 2008 and published in Science in 2010, divided passersby into "heavy" and "light" groups and gave them either a heavy clipboard or a light clipboard containing the résumé of a job applicant. The original experiment found that people holding the heavier clipboard were more likely to rate applicants as suitable for the job. The replication found no such effect. (The replication protocol deviated slightly, in that it was conducted in Charlottesville and not Boston, and passersby were given $5 for their time rather than candy.)

The advocates for greater reproducibility believe that publication pressures create an environment ripe for false positives. Scientists need to publish, and journal editors are eager to publish novel, interesting findings.

Brian Nosek, the leader of this latest reproducibility effort, is executive director of the Center for Open Science, a nonprofit that promotes transparency and reproducibility in research. In an interview with The Washington Post, he acknowledged that the focus on false positives comes at a time when science is already under attack from special interests. But he said, “I think the benefits far, far outweigh the risks.”

He went on: “The reason to trust science is because science doesn't trust itself. We are constantly questioning the basis of our claims and the methods we use to test those claims. That’s why science is so credible.”

Nosek and his allies have drawn heat for their efforts. A major report led by Nosek and published in 2015 in Science found that only about 40 percent of 100 psychology experiments could be replicated (the precise percentage depended on how one defined a successful replication). But that report incited sharp criticism from Harvard psychologist Dan Gilbert and three other researchers, who in a letter to the journal argued that many of the replication experiments didn’t follow the original protocols.

Gilbert and his colleagues argued that, in fact, the results of the Nosek-led project were consistent with psychology experiments being largely replicable.

A statement issued by the journal Science pointed out that all the experiments scrutinized in this latest effort were published before a decision several years ago by Science, Nature and other journals to adopt new guidelines designed to increase reproducibility, in part by greater sharing of data. “Our editorial standards have tightened,” said the statement from Science.

Science deputy editor emeritus Barbara Jasny said in an interview that the failure to replicate studies does not mean that the original experiments were faulty, because “there are differences in protocol, there are differences in study samples.” She noted that the journal Science serves an interdisciplinary audience.

“We do judge on more than just technical competence. We look for papers that may have applications in different fields. We look for papers that are important advances in their own field,” she said.

She said it's important for graduate schools to have uniform methods for teaching students how to design experiments and analyze statistics and advocated more funding for replication studies.

“You can say, 'Oh, this is terrible, it didn’t replicate.' Or you could say: 'This is the way science works. It evolves. People do more studies,' " Jasny said. “Not every paper is going to be perfect when it comes out.”

Read more:

The new scientific revolution: Reproducibility at last

Researchers struggle to replicate 5 cancer experiments

Joel Achenbach
Joel Achenbach covers science and politics for the National desk. He has been a staff writer for The Post since 1990.

Study: A third of U.S. adults don't get enough sleep

By Morgan Manella, Special to CNN

Updated 5:06 PM ET, Thu February 18, 2016

8 photos: Tips for better sleep

Setting an alarm might be the only thing that helps you get up in the morning, but try setting one at night

WATCH LIVE

with greater risks of obesity, diabetes, heart disease, mental illness

Researchers found more than a third of adults reported sleeping less than seven hours in a 24-hour period

states are the most sleep deprived.

The CDC study analyzed data from the 2014 Behavioral Risk Factor Surveillance System to determine whether adults are getting enough sleep. The survey respondents included 444,306 people in all 50 states and the District of Columbia. Researchers found that more than one-third of the adults reported sleeping less than seven hours in a 24-hour period.

Research has shown lack of sleep is associated greater risk of obesity, diabetes, heart disease, mental illness and other chronic conditions. The Centers for Disease Control and Prevention has even called inadequate sleep a public health problem.

Adults 18 to 60 years should be sleeping at least seven hours a night, according to the Academy of Sleep Medicine and the Sleep Research Society, which are sleep-related professional associations.

RELATED: The great American sleep recession

"People just aren't putting sleep on the top of their priority list," said study author Anne Wheaton, PhD, an epidemiologist at the CDC. "They know they should eat right, get exercise, quit smoking, but sleep just isn't at the top of their board. And maybe they aren't aware of the impact sleep can have on your health. It doesn't just make you sleepy, but it can also affect your health and safety."

The study results suggest the need for public awareness and education about sleep health and workplace policies that ensure healthy amounts of sleep for shift workers, according to the study. Health care providers should also discuss the significance of healthy sleep duration with patients and identify why they aren't sleeping enough.

This study was the first to look at sleep hours on a state level, said Wheaton, which allowed them to map which states got more sleep than others.

States in the Southeast and along the Appalachian Mountains reported the least amount of sleep, according to the study. The state with the lowest reported amount of sleep was Hawaii, and the states with the highest reported amount of sleep were South Dakota, Colorado and Minnesota.

State- and county-level data is important because it helps public health departments "see where the problem is most severe," said Wheaton.

For the past decade, about one-third of adults have consistently reported not getting enough sleep, according to Wheaton. She emphasized the importance of establishing good sleep habits, such as going to bed and waking up at the same time each morning; having a good sleep environment, where the bedroom is dark and at a good temperature; removing electronics from your bedroom; avoiding big meals, caffeine, and alcohol before bed; and exercising regularly.

If you're following these guidelines and are still having sleep issues, Wheaton suggests speaking with a physician to see if there is something else that needs to be done.

"It's a public health problem," said Wheaton. "The reason we are trying to draw attention to it is that first it affects such a large proportion of the population and second that it's tied to so many health conditions that are such a big issue."

How your smell can reveal if you're sick

Here's an incentive that really makes people exercise more

18-year-old arrested for pretending to be a doctor, police say

Teen charged with pretending to be a doctor: I never said I was an M.D.

Flash

The Washington Post

Food

Those french fries could kill you, a new study says. But don’t panic!

By Tim Carman June 16 at 11:32 AM

Hey, you, the dude reading this story on your phone over a pile of french fries: Back slowly away from the crispy spuds. They’re out to get you.

That’s the apparent takeaway of a study published this month by the American Journal of Clinical Nutrition. It analyzed the potato consumption of 4,440 American participants, aged 45 to 79 years, over an eight-year period. Researchers used questionnaires to determine each person’s spud-eating habits, including both fried and unfried products, and then used the data to trace links between potato consumption and mortality.

“No study existed about this possible association!” emailed Nicola Veronese, a scientist with the National Research Council in Padova, Italy. Veronese was the lead author and one of a dozen researchers who took part in the study. “There were some studies, re: potato consumption and cardiovascular disease and mortality, but we did not find any paper re: potatoes and mortality!”

Exactly 236 people died during the course of the study. After adjusting for a variety of factors — education, race, income, alcohol consumption and exercise, among other things — the researchers concluded that people who eat french fries more than twice a week are doomed. Doomed!

Okay, they didn’t actually say that. What they did say was that folks who ate “fried potatoes” two or more times a week “were at an increased risk of mortality.” And not the kind of minuscule increase that’s easy to brush off for those firmly committed to their death sticks. The researchers concluded that frequent fried potato eaters more than doubled their risk of premature death.

The ray of hope for tuber lovers? “The consumption of unfried potatoes was not associated with an increased mortality risk,” the study noted. No word if those unfried potatoes were drenched with butter, slathered with sour cream and sprinkled with pre-shredded cheddar.

Everyone, of course, cried fryer-oil tears over the news.

But the truly aggrieved party was the National Potato Council, based in Washington. John Keeling, the organization's chief executive, released a statement saying the "study has significant methodological flaws, which have led to misinterpretations of the data." Among the council's complaints: The participants were taken from a study on osteoarthritis, which meant the subjects either had osteoarthritis of the knee or were at high risk for it. This population, the council argues, "cannot be generalized to other populations."

The council also noted that the participants were asked to fill out a single questionnaire in the "year preceding the start of the study … No other attempt was made to record the participants' dietary patterns in the entire intervening eight years of the study."

"Based on these data, it is very much of a stretch to brand fried potatoes, or any other form of potato, as unhealthy," Keeling said in his statement. "The food consumption reported in the study may not have reflected usage over the course of the lifetime, further illustrating the danger of branding potatoes (or any other food item) as being unhealthy or healthy in the context of this study."

Keeling went on to promote the nutritional value of potatoes, which, of course, is his job.

Veronese did not dispute some of Keeling's charges, agreeing that the research subjects were taken from a study on osteoarthritis and that the one-time questionnaire does have "some limitations." But Veronese said such one-off questionnaires are "common" to long-term studies. What's more, the researcher added, osteoarthritis subjects share similar characteristics with the general population in the United States. "Our findings," Veronese emailed, "would be similar in other populations, but other studies are needed of course."

Marion Nestle, professor of nutrition, food studies and public health at New York University, wasn't so alarmed by the study's results.

"First, this is an association," Nestle emailed. "Fried potatoes are associated with somewhat higher mortality, but this does not mean that they cause death. People who eat a lot of fried potatoes might have other unhealthy lifestyle practices — they might have worse diets in general, not exercise, smoke more or drink more."

"Second," Nestle added, "the association is not strictly dose-related. At lower levels of intake, the association is not statistically significant. The most significant associations are at the highest levels of intake of fried potatoes — three times a week or more. The moral here is moderation. If you love french fries, make them a once-in-a-while treat."

You're welcome, Internet.

More from Food:

Blue catfish are destroying the Chesapeake Bay. Congress isn't helping.

Should white chefs sell burritos? A Portland food cart's revealing controversy.

The surprising number of American adults who think chocolate milk comes from brown cows

Tim Carman serves as the full-time writer for the Post's Food section and as the $20 Diner for the Weekend section, a double duty that requires he ingest more calories than a draft horse. Follow @timcarman

Study: Acetaminophen reduces not only pain, but pleasure, too

By Don Melvin, CNN

Updated 9:28 AM ET, Wed April 15, 2015

BREAKING NEWS

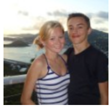
Jurors find former Patriots player Aaron Hernandez guilty of murder. Watch the sentencing live on CNNgo.

Photos: A guide to (legal) pain relief 8 photos

Acetaminophen – This compound can ease minor muscle, back, tooth and joint pain and reduce fever. Sold under brand names such as Tylenol, Liquiprin and Panadol, it works by regulating the part of your brain that controls your body's temperature and inhibits the synthesis of prostaglandin in the central nervous system. A new study has found that it could reduce pleasure as well. Too much of the drug can cause rashes, liver failure and even death. Here's a guide to some of the most commonly used pain relief medications:

1 of 8 Hide Caption

The latest from CNN Health



Real 'Fault in Our Stars' couple reunited by hope

Story highlights

- Subjects taking acetaminophen reacted less strongly to both pleasant and unpleasant photos
- Each week, 52 million Americans use the pain reliever
- Unknown whether other pain products produce the same effect

(CNN)—Feeling so happy you just can't stand it? You might want to pop some acetaminophen.

A new study has found that acetaminophen, the main ingredient in Tylenol, most forms of Midol and more than 600 other medicines, reduces not only pain but pleasure, as well.

The authors of the study, which was published this week in Psychological Science, say that it was already known that acetaminophen blunted psychological pain. But their new research led them to the conclusion that it also blunted joy -- in other words, that it narrowed the range of feelings experienced.

"This means that using Tylenol or similar products might have broader consequences than previously thought," said Geoffrey Durso, a doctoral student in social psychology at Ohio State University and the lead author of the study. "Rather than just being a pain reliever, acetaminophen can be seen as an all-purpose emotion reliever."

Subjects evaluated pleasant, unpleasant photos

How much is too much Tylenol?
02:42

PLAY VIDEO ↘

The researchers tested their thesis by showing 82 college students 40 photographs -- some of highly pleasant images, such as children with kittens, and some of highly unpleasant images, such as children who were malnourished.

Half of the participants in the study were given "an acute dose" of acetaminophen -- 1,000 milligrams -- and the other half were given a placebo with the same appearance. The subjects were then asked to rate the photos according to how unpleasant or pleasant they were.

Those who took the acetaminophen rated all the photos less extremely than those who took the placebo.

"In other words, positive photos were not seen as positively under the influence of acetaminophen and negative photos were not seen as negatively," the authors reported.

Drug did not alter sense of magnitude in general

The researchers followed up by testing a group of 85 people to see whether this change in judgment applied just to emotions or whether the drug blunted people's evaluation of magnitude in general.

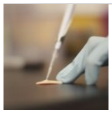
This group showed the same blunting of emotional reactions. But acetaminophen did not affect how much blue they saw in each photo.

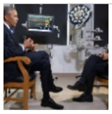
But people who participated in the study did not appear to know they were acting differently, said Baldwin Way, an assistant professor of psychology who was another of the study's authors.

"Most people probably aren't aware of how their emotions may be impacted when they take

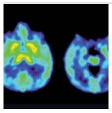
'Fearless' Ebola nurse trains at Emory University

Are human head transplants coming soon?


60 years after a vaccine, this could eradicate polio


President Obama links climate change to public health

Two doctors fight for their own choice of how to die

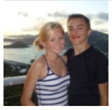

Could veterans have concussion-related CTE?

Focus on Health

You can stress-eat these superfoods

Your brain on multitasking

Why your commute is bad for you


Real 'Fault in Our Stars' couple reunited by hope

Stave off soreness like a pro athlete

Could this prevent dementia?

More Top Stories

Rita Wilson reveals cancer fight

Mars may have puddles

Teacher says he paid for

acetaminophen," Way said.

Each week, about 23% of American adults -- or 52 million people -- use a medicine containing acetaminophen, according to the nonprofit Consumer Healthcare Products Association.

The authors said it was not known whether other pain relievers, such as ibuprofen and aspirin, have the same effect. But have no fear -- they plan to study that question, as well.

Follow @Don_Melvin

More from CNN

'People blatantly lied about Darren Wilson' ▶

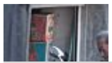
Killer co-pilot's chilling words to ex-girlfriend ▶

Travelers beware! Pickpocket artist shares secrets ▶

Viral tornado photo a sign from God? ▶

This is where Africa's multimillionaires live

Andy Murray weds Kim Sears: 'Royal wedding of Scotland'


sex with 12K women

What cop said after S.C. shooting

SpaceX rocket lands, falls over

Aaron Hernandez verdict: Guilty of murder

More from CNN

Police release video from alleged scene of gang rape

Defector: Why North Koreans Hate America

Aaron Hernandez trial: Jury has reached a verdict

Dash cam video shows police car hit suspect

Admit it: You love Tax Day! (Opinion)

Why voters won't fall for Rubio

More from Health

The post-recall food safety check

Woman sues university over weight-loss suggestion

Could veterans have concussion-related CTE?

More from Don Melvin

Three weeks of Saudi strikes in Yemen, no peace in sight

Boko Haram kidnapping of 200 Nigerian schoolgirls, a year later

Healthgrades

Treating insomnia may cure depression

8 common complications of COPD

Signs of MS? Find a local doctor

7 essential medical tests for diabetics

How glucose meters work

New York City, NY 58°

Search CNN

NEWS
U.S.
WORLD
POLITICS
TECH
HEALTH
ENTERTAINMENT
LIVING
TRAVEL
MONEY
SPORTS

VIDEO
CNNGO
LATEST NEWS
MUST WATCH VIDEOS
DIGITAL STUDIOS

TV
CNNGO
SCHEDULE
CNN FILMS
SHOWS A-Z
FACES OF CNN WORLDWIDE

OPINIONS
POLITICAL OP-EDS
SOCIAL COMMENTARY
IREPORT

MORE...
PHOTOS
LONGFORM
INVESTIGATIONS
CNN PROFILES A-Z
CNN LEADERSHIP

CNN U.S. Edition

Terms of service | Privacy guidelines | AdChoices | Advertise with us | About us | Work for us | Help | Transcripts | License Footage | CNN Newsource

The New York Times

Watch 4 Decades of Inequality Drive American Cities Apart

The biggest metropolitan areas are now the most unequal.

By Emily Badger and Kevin Quealy

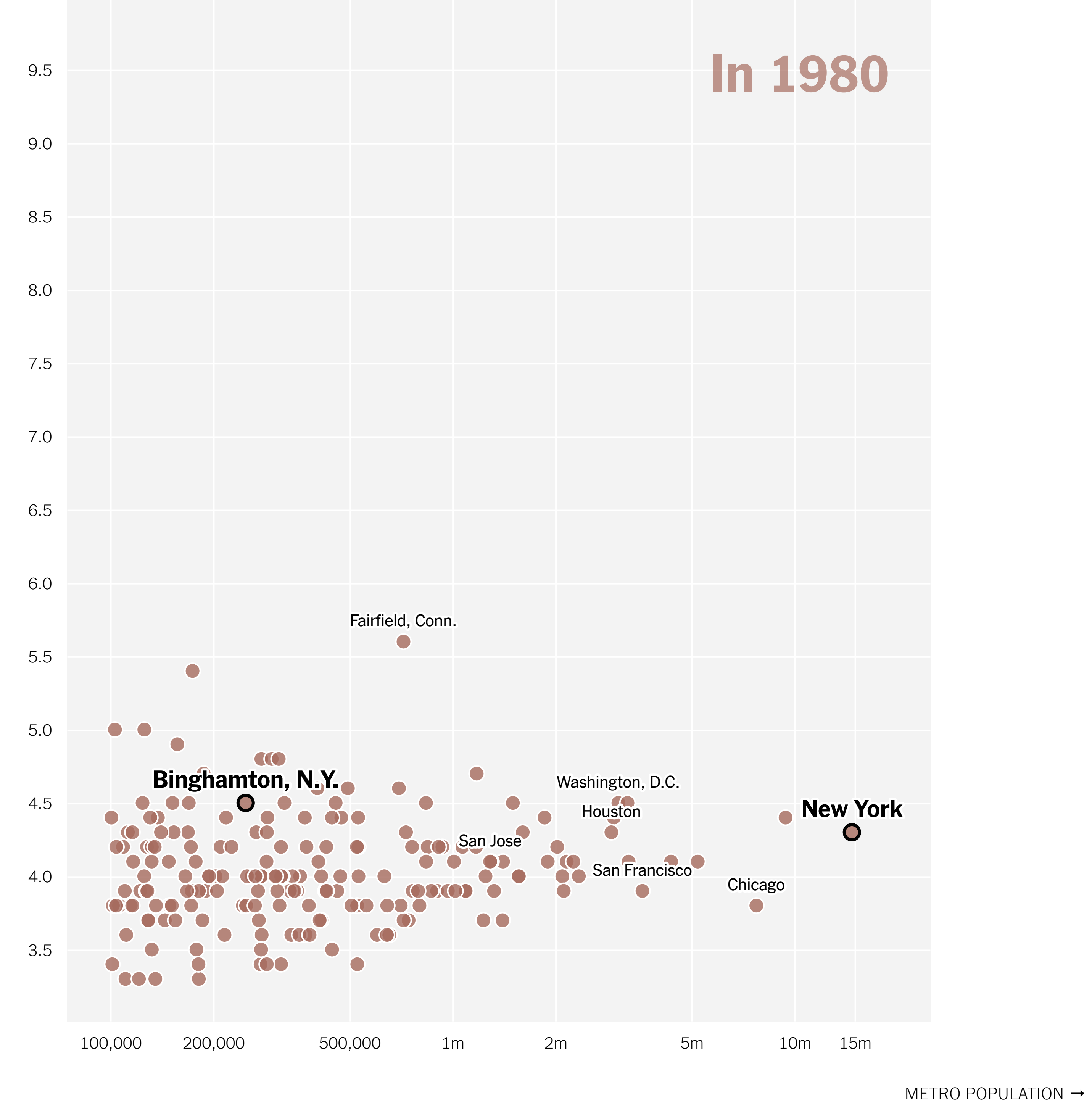
Dec. 2, 2019, 5:00 a.m. ET

In 1980, highly paid workers in Binghamton, N.Y., earned about four and a half times what low-wage workers there did. The gap between them, in a region full of I.B.M. executives and manufacturing jobs, was about the same as the gap between the workers near the top and the bottom in metro New York.

Since then, the two regions have diverged. I.B.M. shed jobs in Binghamton. Other manufacturing disappeared, too. High-paying work in the new knowledge economy concentrated in New York, and so did well-educated workers. As a result, by one measure, wage inequality today is much higher in New York than it is in Binghamton.

Ratio of 90th-percentile wages to 10th-percentile wages in 195 metro areas

↑ MORE INEQUALITY



Dots represent metro areas; names are shortened for clarity. ▪ Analysis of census and American Community Survey data by Jaison Abel and Richard Deitz, Federal Reserve Bank of New York

What has happened over the last four decades is only partly a story of New York's rise as a global hub and Binghamton's struggles. Economic inequality has been rising everywhere in the United States. But it has been rising much more in the booming places that promise hefty incomes to engineers, lawyers and innovators. And those places today are also the largest metros in the country: New York, Los Angeles, San Francisco, San Jose, Houston, Washington.

This chart, using data from a recent analysis by Jaison Abel and Richard Deitz of the New York Fed, captures several dynamics that have remade the U.S. economy since 1980. Thriving and stagnant places are pulling apart from each other. And within the most prosperous regions, inequality is widening to new extremes. That this inequality now so clearly correlates with city size — the largest metros are the most unequal — also shows how changes in the economy are both rewarding and rattling what we have come to think of as "superstar cities."

In these places, inequality and economic growth now go hand in hand.

Back in 1980, Binghamton's wage inequality made the region among the most unequal in the country, according to the Fed analysis. It ranked 20th of the 195 metros shown here as measured by comparing the wages of workers at the 90th percentile with those at the 10th percentile of the local wage distribution, a measure that captures the breadth of disparities in the local economy without focusing solely on the very top. In 1980, New York City was slightly less unequal, ranking 44th by this measure.

Forty years ago, none of the country's 10 largest metros were among the 20 most unequal. By 2015, San Francisco, New York, Houston, Los Angeles, Dallas and Washington had jumped onto that list, pulled there by the skyrocketing wages of high-skilled workers. Binghamton over the same period had become one of the least unequal metros, in part because many I.B.M. executives and well-paid manufacturing workers had vanished from its economy.

In effect, something we often think of as undesirable (high inequality) has been a signal of something positive in big cities (a strong economy). And in Binghamton, relatively low inequality has been a signal of a weak economy. (The Fairfield-Bridgeport, Conn., metro stands out in either era because the deep poverty of its urban core is surrounded by particularly rich suburbs.)

These patterns are hard to reconcile with appeals today for reducing inequality, both within big cities and across the country. What are Americans supposed to make of the fact that more high-paying jobs by definition widen inequality? Should New Yorkers be O.K. with growing inequality in New York if it's driven by rising wages for high-skilled workers, and not falling wages for low-skilled ones?

"That's more of a political question," said Nathaniel Baum-Snow, an economist at the University of Toronto. "That's a question of what we decide our values should be as a society."

Tom VanHeuvelen, a sociologist at the University of Minnesota who has also researched these patterns, said: "It seems obvious to me that it doesn't need to be the way that it is right now. This isn't the only inevitable outcome we have when we think about the relationship between cities, affluence and inequality."

Economists say that the same forces that are driving economic growth in big cities are also responsible for inequality. And those forces have accumulated and reinforced each other since 1980.

High-skilled workers have been in increasing demand, and increasingly rewarded. In New York, the real wages for workers at the 10th percentile grew by about 15 percent between 1980 and 2015, according to the Fed researchers. For the median worker, they grew by about 40 percent. For workers at the 90th percentile, they nearly doubled.

That's partly because when highly skilled workers and their firms cluster in the same place today, they're all more productive, research shows. And in major cities, they're also tied directly into the global economy.

"If you're someone who has skills for the new economy, your skills turn out to be more valuable in bigger cities, in a way that wasn't true 30 to 40 years ago," Mr. Baum-Snow said.

It's no surprise, then, that high-skilled workers have been sorting into big, prosperous cities, compounding the advantages of these places (and draining less prosperous places of these workers).

At the same time, automation, globalization and the decline of manufacturing have decimated well-paying jobs that once required no more than a high school diploma. That has hollowed out both the middle class in big cities and the economic engine in smaller cities. The result is that changes in the economy have disproportionately rewarded some places and harmed others, pushing their trajectories apart.

Add one more dynamic to all of this: Inequality has been rising nationally since the 1980s. But because the Bay Area and New York regions already had more than their fair share of one-percenters (or 10 percenters) in 1980, the national growth in income inequality has been magnified in those places.

“We’ve had this pulling apart of the overall income distribution,” said Robert Manduca, a Ph.D. student in sociology and social policy at Harvard who has found that about half of the economic divergence between different parts of the country is explained by trends in national inequality. “That overall pulling apart has had very different effects in different places, based on which kinds of people were already living in those places.”

Mr. Manduca says national policies like reinvigorating antitrust laws would be most effective at reducing inequality (the consolidation of many industries has meant, among other things, that smaller cities that once had company headquarters have lost those jobs, sometimes to big cities).

It is hard to imagine local officials combating all these forces. Increases to the minimum wage are likely to be swamped — at least in this measure — by the gains of workers at the top. Policies that tax high earners more to fund housing or education for the poor would redistribute some of the uneven gains of the modern economy. But they would not alter the fact that this economy values an engineer so much more than a line cook.

“If you brought the bottom up, it would be a better world,” said Richard Florida, a professor at the University of Toronto who has written extensively about these trends. “But you’d still have a big rise in wage inequality.”