Certificate in Applied Biostatistics

Brian Healy, PhD

Harvard Catalyst Applied Biostatistics Certificate Program

Acknowledgments

- Many people to thank:
  - MGH CRP
  - MGH ECOR
  - Harvard Catalyst
  - Harvard Catalyst Post-Graduate Education Program

Lecture outline

- Course introduction
- STATA introduction
- Types of data
- Summary statistics
  - Location: mean, median
  - Spread: standard deviation, variance, range
- Graphics
  - Box plot, histogram
  - Scatterplot

Welcome!!!

- Certificate in Applied Biostatistics
- Main instructor: Brian Healy
- Additional instructors
  - Henry Feldman
  - E. John Orav
  - Garrett Fitzmaurice
  - Paul Catalano
- Course format: Lecture/Web learning

About me

- Assistant professor at HMS
- Research focus: Multiple sclerosis
- Most of my example are based on MS, but I would be happy to integrate other examples if they are of interest to people

About course

- Idea came from many people asking for more info
- Course objectives
  - How to think statistically
  - How to choose the correct statistical analysis
  - Focus on concepts and how to apply the concepts in a statistical package
  - When you need more help

What this course is not

- Formal math derivations of all statistical tests
  - Math is often necessary
  - I will present formulas, but usually I will not derive the formulas unless there is benefit from derivation
- Introduction to all statistical packages
- Analysis of your dataset

What I ask of you

- Have some faith
  - Biostatistics is a wonderful and fun topic that you will be able to learn
- Try the practicums
  - Best way to learn is to try it yourself
- Email me with questions
- Do not fall too far behind
  - Biostatistics builds on itself

Assessment

- To receive your certificate of participation, you will need to complete 80% of the homework assignments and 80% of the short questions.

Textbooks

- There is no textbook for this course since we are covering many different topics
- Potential books of interest
  - *Fundamentals of Biostatistics* by Rosner
  - *Applied Regression Analysis and Other Multivariable Methods* by Kleinbaum et al
  - *Applied Longitudinal Analysis* by Fitzmaurice, Laird and Ware
  - *Applied Survival Analysis* by Hosmer, Lemeshow and May
  - *Applied Logistic Regression* by Hosmer and Lemeshow

Course topics

- Five units
  - Basic techniques
  - Study design
  - Regression
  - Correlated outcomes
  - Survival analysis/Additional techniques
- Syllabus for more specifics

STATA

- Many common statistical packages (i.e. SPSS, SAS, STATA, R)
- Focus of this course will be STATA
  - Inexpensive
  - Comprehensive
  - User-friendly
- To order a copy of STATA, go to http://www.stata.com/order/new/edu/gradplans/gp-direct.html

Introduction to biostatistics

Outline for today

- Research study
- Types of data
- Measures of location
- Measures of spread
- Graphics
- Correlation

Goals of lecture

- At the end of this lecture, you will be able to:
  - Identify different types of data
  - Identify the ways to summarize and visualize each type of data
  - Interpret scatterplots

Big picture

- A research study involves careful planning prior to any data collection or analysis
- A study is only as good as the study design
- Prior to completing any data analysis, you should inspect your data using summary statistics and graphics

Research study

I. Study design

- Experimental question- What are you trying to learn? How will you prove this?
- Sample selection- Who are you going to study?

II. Data collection

- What should be collected?

III. Analysis of data

- Results- Was there any effect?
- Conclusions- What does this all mean? To whom do results apply?

How is statistics related to each stage?

I. Study design
- Experimental question- Define **outcome**, sources of variability, **and analysis plan**
- Sample selection- Sample size, type of sample

II. Data collection
- What to collect?

III. Analysis of data
- Results- **Hypothesis test**
- Conclusion- Significance of effect/generalizability

Population vs. sample

- When we complete a research study, we are usually collecting a sample of people or animals that come from a larger population
  - We usually hope that our sample is a representative and random sample from the population so that we can make an inference about the population based on our sample
- This concept is critical and we will come back to it many times (Lecture 3)

Description vs. inference

- The first step in any study is to describe the data that has been collected
- The goal of most studies is to also use the data to infer something about a larger group
  - We will describe how we do statistical inference later in the course
  - Difference between case report and research study

Data description

Example

- My research focus is data analysis of multiple sclerosis (MS)
- At the Partners MS Center, we have a longitudinal study of patients (CLIMB)
- A subgroup of this study are given questionnaires related to quality of life
  - Patient reported outcomes
  - Sample size is 252

Variables

- A variable is something that is measured in all of the people/ in our sample
- Common variables in experiments are:
  - Age
  - Disease grade-an ordinal scale
  - Presence/absence of disease
  - Expression a cytokine of interest (ex. IFN-γ)
  - Time to disease

Types of variables

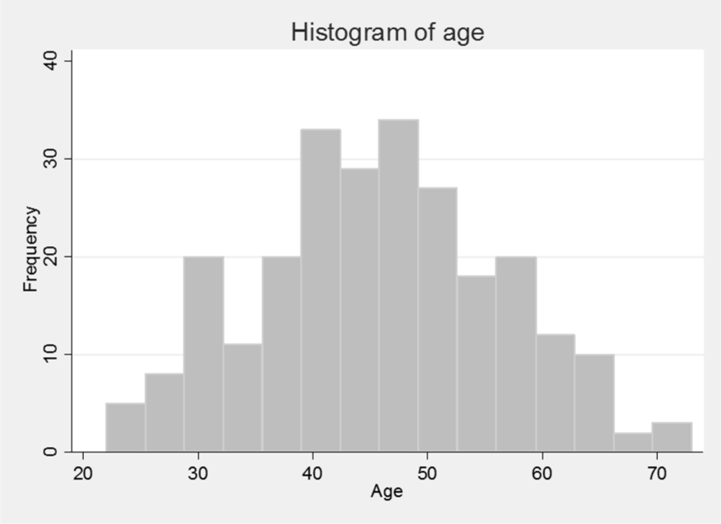
- Continuous variable: Age, expression level
- Dichotomous (binary) variable: Dead/alive, Wild type/mutant
- Nominal/categorical variable: Race, group membership
- Ordinal variable: Mild/Moderate/Severe, level of stat knowledge
- Count variable: Number of lesions
- Time to event variable: Time to death

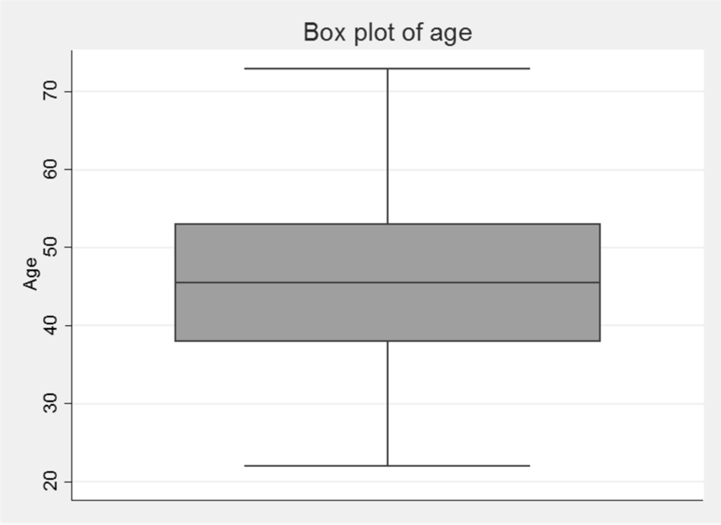
Ways to express data

- All data has a **distribution** that we would like to describe
- Numerical summaries
  - **Summary statistics**
  - Summarize a large group of numbers with one or two numbers
- Graphics
  - Many types of graphs help provide a better understanding of a dataset

Continuous variables

- Summary statistics
  - **Location**
    - **Mean**
    - **Median**
  - **Variability**
    - **Standard deviation**
- Graphs





Mean

- The most common measure of location is the sample mean or average
- To calculate the mean of a group of numbers, we add all of the numbers and divide by the total number of numbers
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The mean age in our sample is 45.7 years

Standard deviation

- The most common measure of spread is the sample **standard deviation**
- The sample standard deviation squared is the sample **variance**
- Each describes the deviation of points from the mean
- $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- The sample standard deviation is 10.5

Median

- The **median** is the middle number or 50th percentile
- To calculate the median, we list the number in order and we find the middle
- Median is sometimes preferred when there is a large skew in the data
- The median age in our sample is 45.5

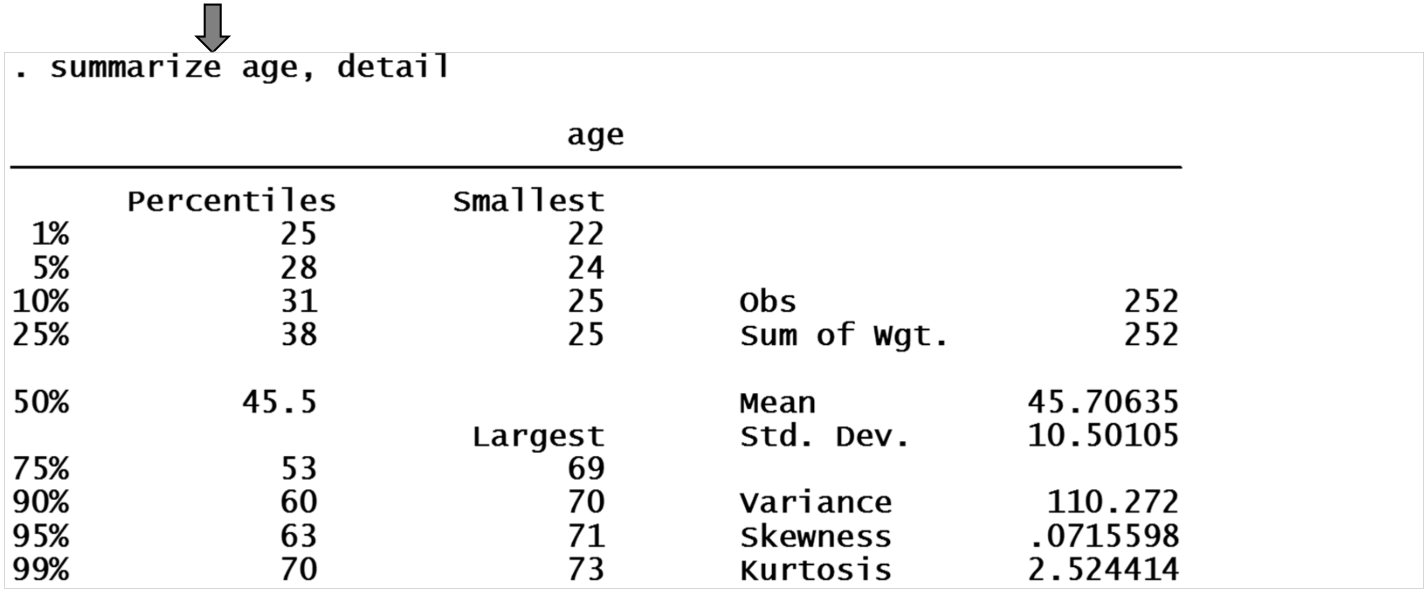
Interquartile range and range

- The interquartile range (IQR) is the distance between the 25th and 75th percentile
- The range is the distance between the minimum and maximum
- The IQR in our sample is 53-38=15
- The range in our sample is 73-22=51

STATA

- All of these statistics are easily calculated in STATA

STATA command



```
. summarize age, detail

                             age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           25             22
 5%           28             24
10%           31             25       Obs                 252
25%           38             25       Sum of Wgt.         252

50%         45.5                      Mean           45.70635
                        Largest       Std. Dev.      10.50105
75%           53             69
90%           60             70       Variance        110.272
95%           63             71       Skewness       .0715598
99%           70             73       Kurtosis       2.524414
```

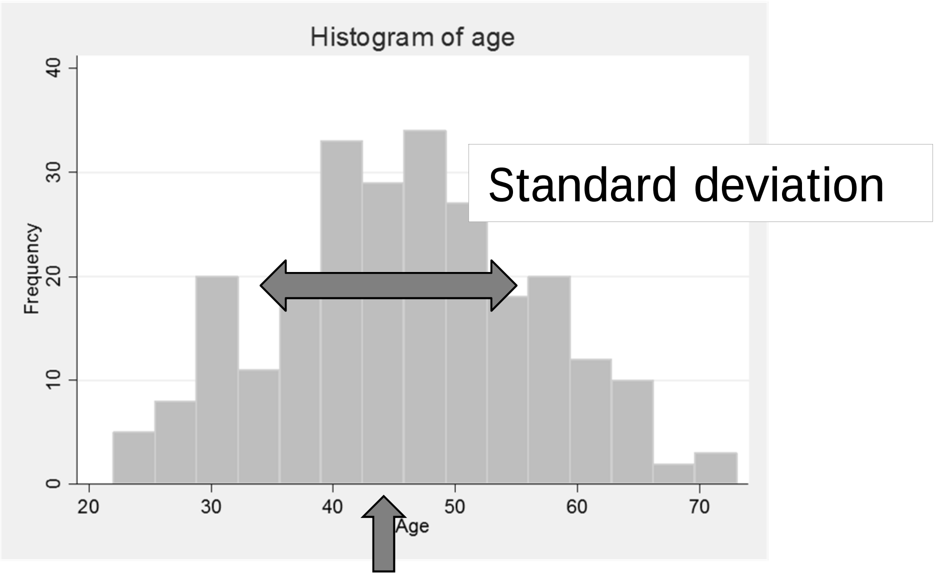
STATA menus: Statistics\Summaries, tables, and tests\Summary and descriptive statistics\Summary statistics

Histogram

- A histogram shows the distribution of continuous data by breaking the data into bins and showing the frequency in each bin
- From a histogram, you can usually get a reasonable understanding of:
  - Mean and standard deviation
  - **Symmetry/skewness**

Example

- On this histogram



Standard deviation

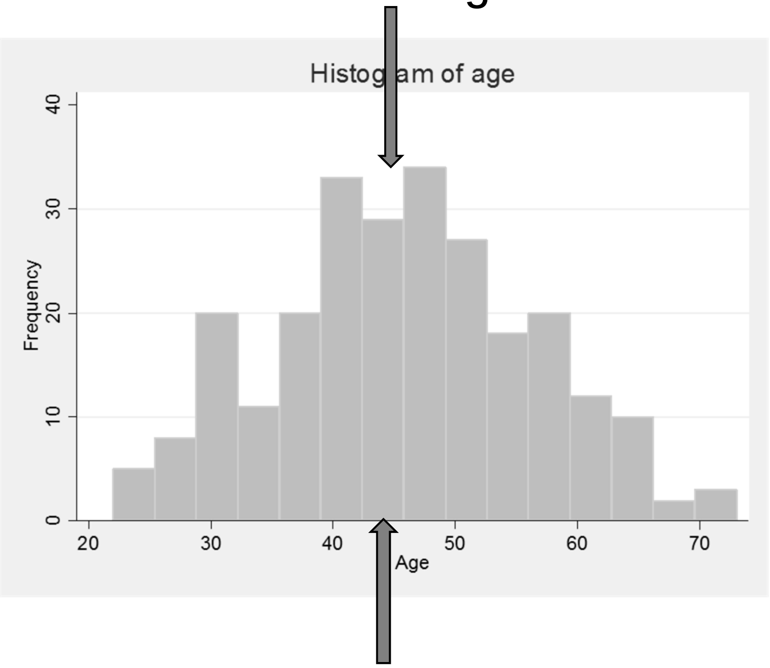
Sample mean=45.7

STATA command: histogram test_age, frequency xtitle(Age) title(Histogram of age)

Skewed example

Symmetric distribution

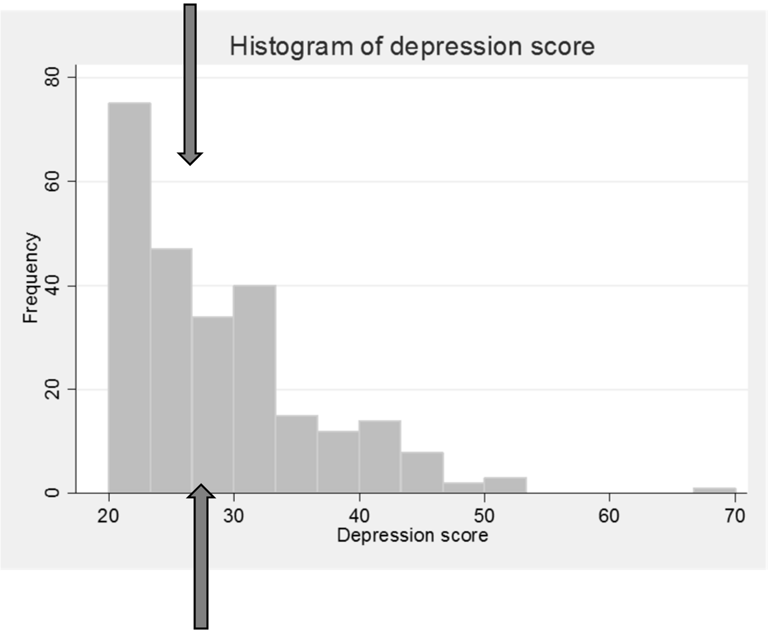
Median age=46



Mean age=45.7

Skewed distribution

Median score=27



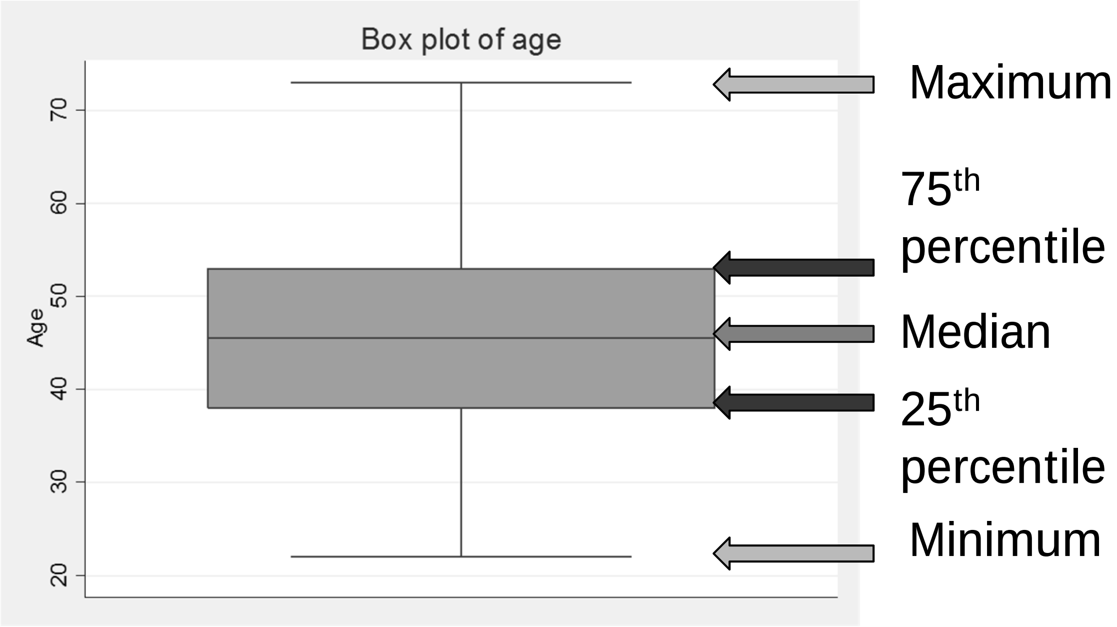
Mean score=28.7

Relationship between mean and median

- Symmetric distribution: mean and median are very close and provide similar information
- Skewed distribution/outliers: mean can be affected by high or low values more than median

Box plot

- A box plot also shows the distribution of continuous data
- Shows median and IQR



Maximum

75th percentile

Median

25th percentile

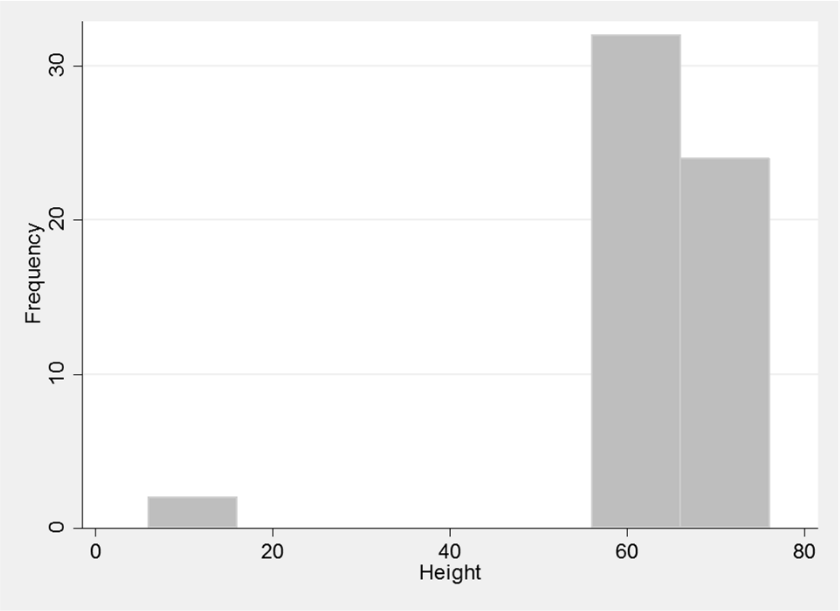
Minimum

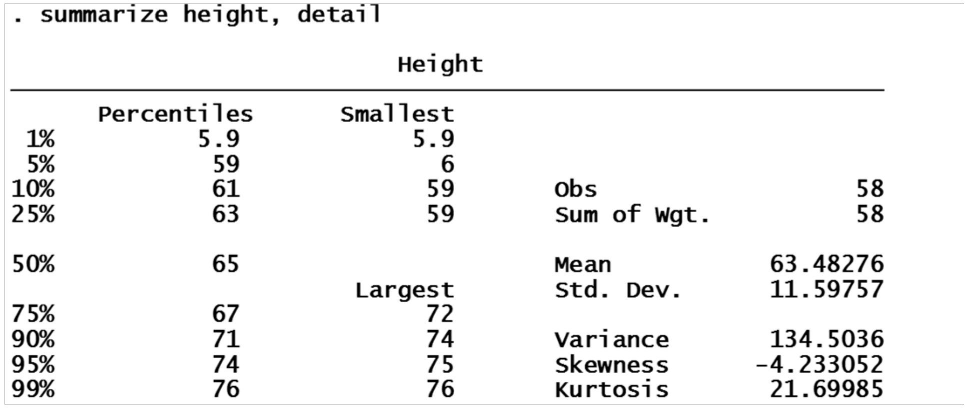
Aside

- One ancillary benefit of computing these summary statistics and graphics is errors in data entry can be more easily identified
- It is critical to ensure that your data are "clean" before you proceed with any analysis

Example

- Here is a histogram of the heights in inches in the previous class dataset
- What is wrong?





```
. summarize height, detail

                            Height
-------------------------------------------------------------
      Percentiles      Smallest
 1%          5.9            5.9
 5%           59              6
10%           61             59       Obs                  58
25%           63             59       Sum of Wgt.          58

50%           65                      Mean           63.48276
                        Largest       Std. Dev.      11.59757
75%           67             72
90%           71             74       Variance       134.5036
95%           74             75       Skewness      -4.233052
99%           76             76       Kurtosis       21.69985
```

Conclusions

- There are several summary statistics and graphs for continuous variables
- The best summary statistic to report is often related to the distribution of the data
- Always a good idea to investigate the summary statistics/graphs as a first step to ensure there are no problems in the dataset

More than just the mean

- Often we would like to know more than just the mean of a sample
  - We often would like to know the error associated with the estimated mean
  - Mean +/- Standard deviation
    - Table 1 of many papers
    - Data description
  - Mean +/- Standard error (Lecture 3)
    - 95% confidence interval
    - Inference

Example-SDMT

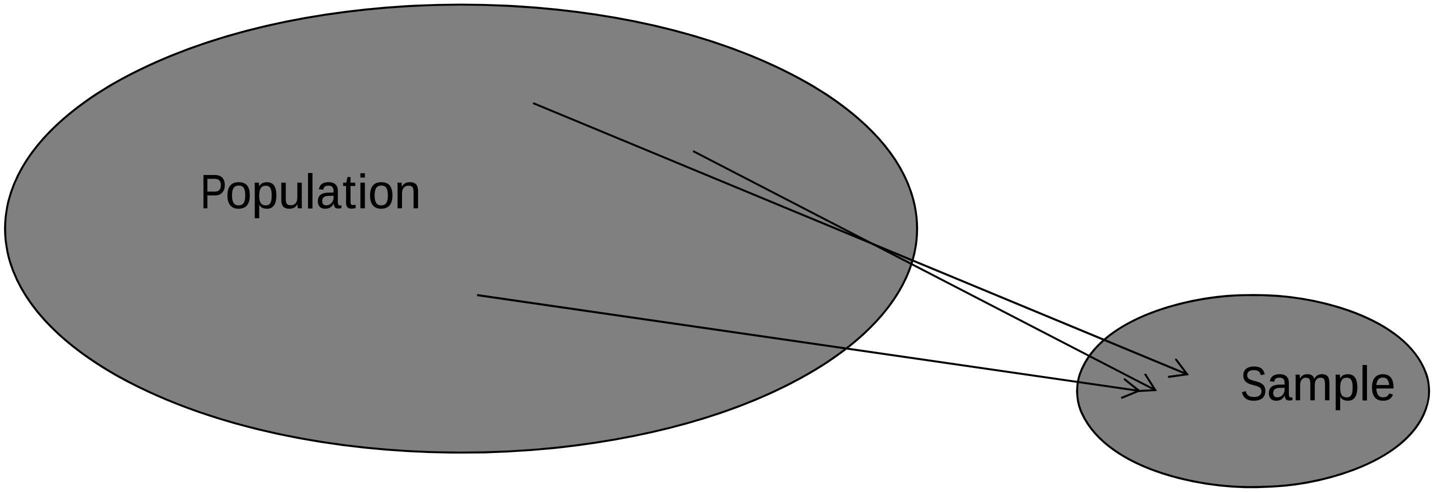
- One way to measure speed of information processing (a domain of cognitive functioning) is the symbol digit modalities test (SDMT)
- This measure has been shown to have desirable properties in RRMS patients so it will be our measure of cognitive functioning
- We would like to estimate the mean SDMT score in RRMS patients

How could we investigate this?

- To investigate cognitive functioning in RRMS patients, we could find all RRMS patients and measure the SDMT in all patients
  - This would mean that we tested the entire population so our sample mean would be the population mean
  - Impossible or at least very unrealistic
- Alternatively, we could take a sample

Sample

- We would like our sample to be
  - Representative of the population
  - Random



Estimation of population mean

- You have collected one sample of *n* people
- How would you estimate the population mean?
- Two types of estimates:
  - **Point estimate**: a single number estimate of the parameter of interest
  - **Interval estimate**: give an interval around the point estimate incorporating the variability

Sample mean

- The sample mean is a logical point estimate of the population mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

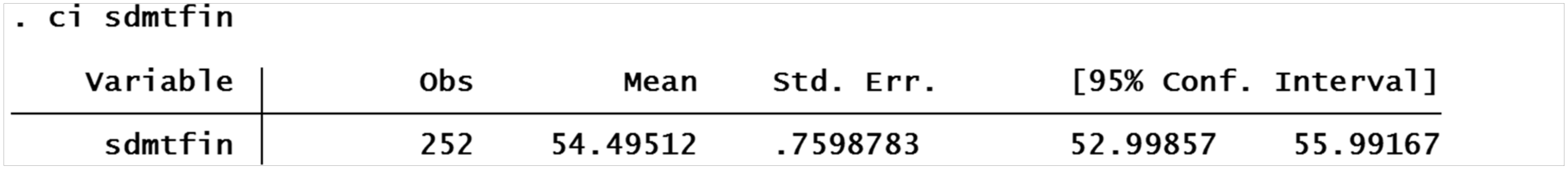
- Is this exactly the population mean?
- If you took a second sample, what would you expect about this sample mean compared to the original sample mean?

Sampling variability

- The reason for the differences between random samples is called sampling variability or chance
- By chance, sometimes our sample mean is high and other times our sample mean is low
- A lot of this course will focus on understand the amount of sampling variability
- 95% confidence interval will incorporate the uncertainty in the estimated mean (Lecture 3)

STATA output

- Here is the STATA output for a 95% confidence interval
  - – We will describe this in detail in Lecture 3



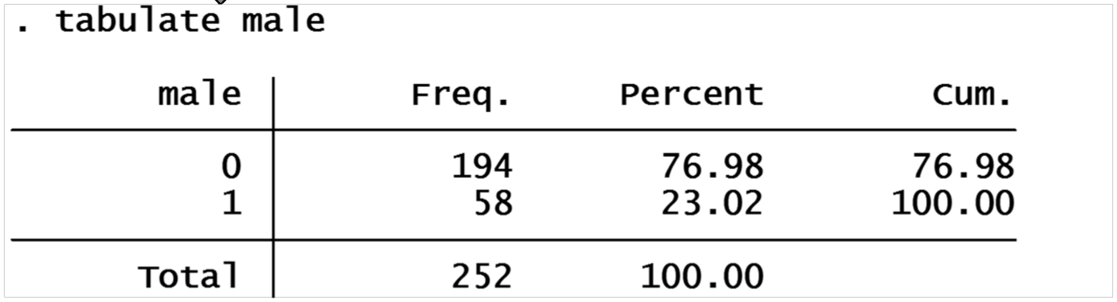
```
. ci sdmtfin

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     sdmtfin |        252    54.49512    .7598783        52.99857    55.99167
```

Dichotomous variables

- Dichotomous (binary) variables represent two categories
- The most common way to summarize a dichotomous variable is as a proportion
- For example, in our sample the proportion of men is 58/252
- $\hat{p}_{males} = \frac{58}{252} = 0.23$
- $\hat{p}_{females} = \frac{194}{252} = 0.77$

Proportion calculation in STATA

- To calculate a proportion in STATA, you can use many commands, including *tabulate*



```
. tabulate male

       male |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        194       76.98       76.98
          1 |         58       23.02      100.00
------------+-----------------------------------
      Total |        252      100.00
```

Is this the correct interpretation?

"We'll only do 72% of it, since it's been reported that 28% of all surgery is unnecessary."

Inference with proportions

- In the next two months, we will see proportions in the form of political polls
  - 51% of people support Barack Obama
  - 52% of people support Elizabeth Warren
- Each poll will be accompanied by a margin of error
  - This is a way of expressing the range of plausible values
  - 95% confidence interval

Aside

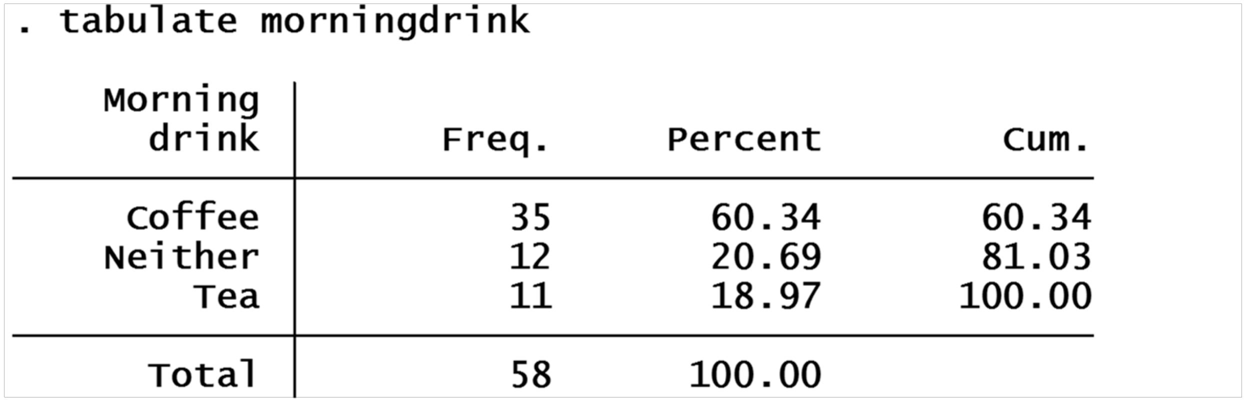
- Why is there so much variability from poll to poll?
- What causes the results of a poll to change?
  - **Chance**-A new sample of people will give a different answer than the previous
  - **Bias**-Has the population sampled changed?
    - The entire population vs. likely voters
    - Massachusetts vs. entire country
  - How you gather your sample affects your conclusions
  - Lectures 3-6

Nominal/categorical variable

- Nominal or categorical variables have multiple levels, but these levels are not ordered
- Dichotomous is a special case of nominal
- Examples:
  - Race
  - Hair color
  - Treatment group

Table

- One way to express a nominal/categorical variable is a table
- Using the previous class dataset, we can tabulate the number and proportion of people who like coffee, tea and neither



```
. tabulate morningdrink

    Morning |
      drink |      Freq.     Percent        Cum.
------------+-----------------------------------
     Coffee |         35       60.34       60.34
    Neither |         12       20.69       81.03
        Tea |         11       18.97      100.00
------------+-----------------------------------
      Total |         58      100.00
```

Ordinal variable

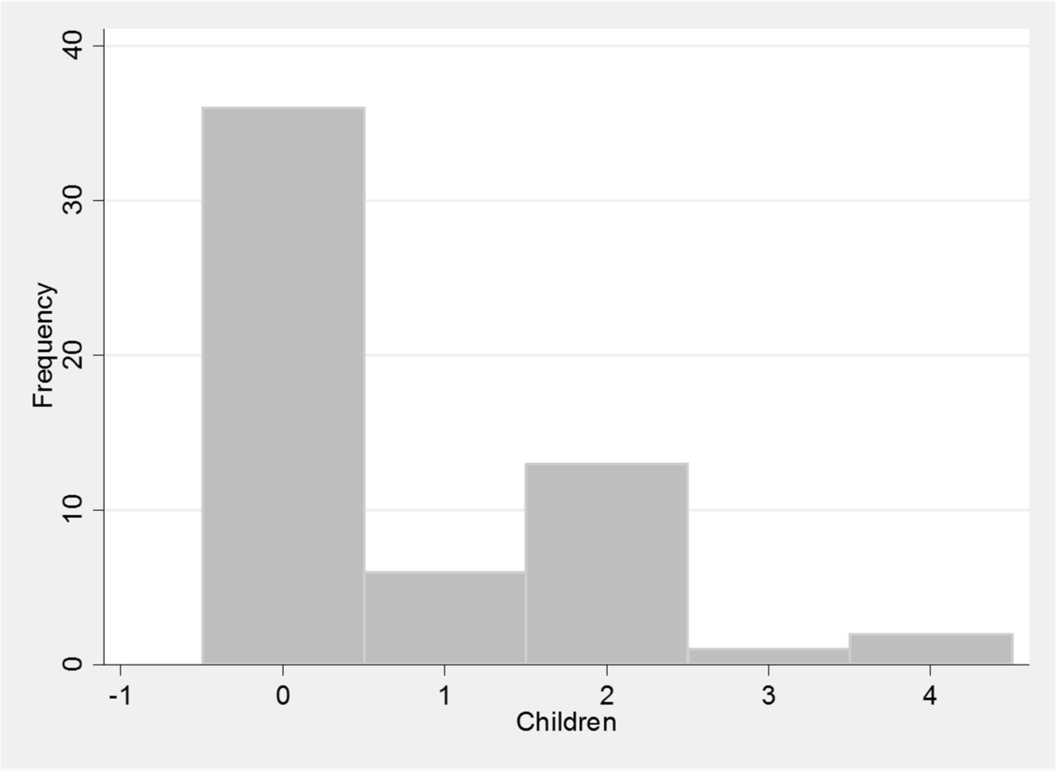
- Ordinal variables are the most challenging type of data in my opinion
- Two main types of ordinal data
  - Order but no magnitude
    - Mild, moderate, severe
  - Order and information about magnitude
    - Fatigue score
- Approaches for handling ordinal data based on number of categories

Count variables

- Count outcomes are very common, especially in epidemiology
- For count outcomes, the mean is usually an interesting statistic
  - In previous class dataset, the mean number of children is 0.74
  - Although this number is not a possible number of children, it is still interpretable

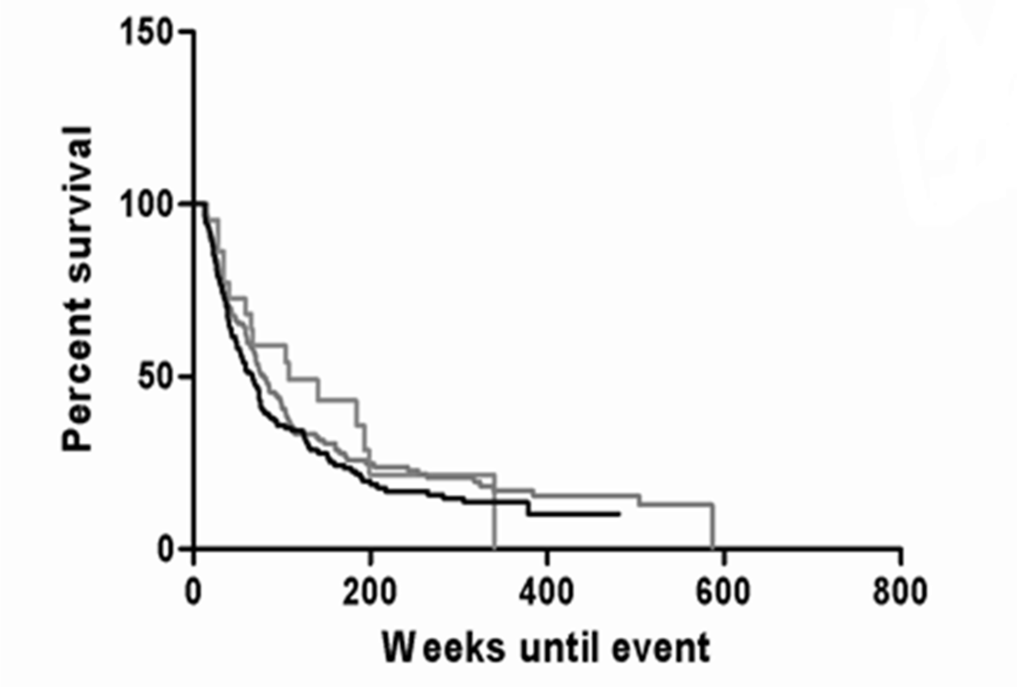
Graph

- A bar graph can be used to show the distribution of a count variable
- STATA uses *histogram* for this too



Time to event

- Survival time
  - – Median
- Graph
  - – Kaplan-Meier curve
  - – NEJM article



Conclusions

- Each type of data has specific summary statistics and graphics
- Identifying the types of data that you are dealing with will ensure that you use the best statistics/graphs

Description vs. comparison

- In many instances, description of the outcome variable is the focus
  - Estimate and **confidence interval**
- Description is not enough, rather comparison is of interest
- What do we need for comparison?
  - Second variable-usually called **explanatory variable or predictor**

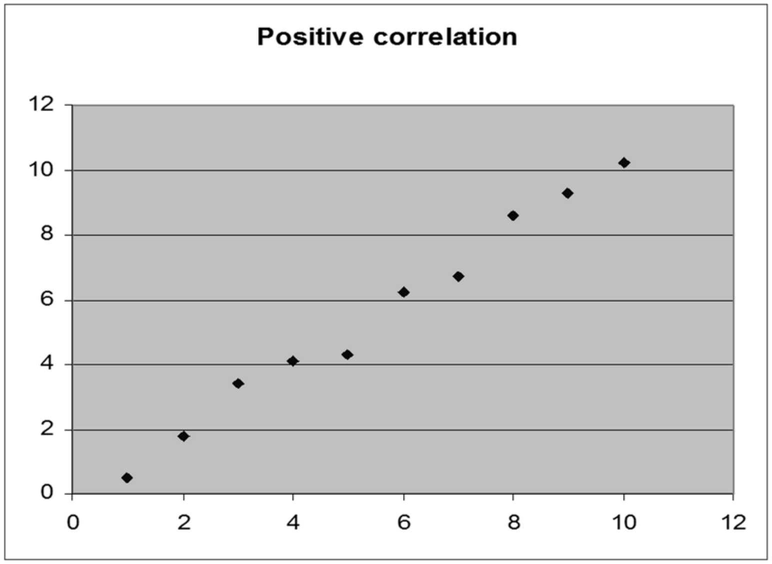
Contingency table

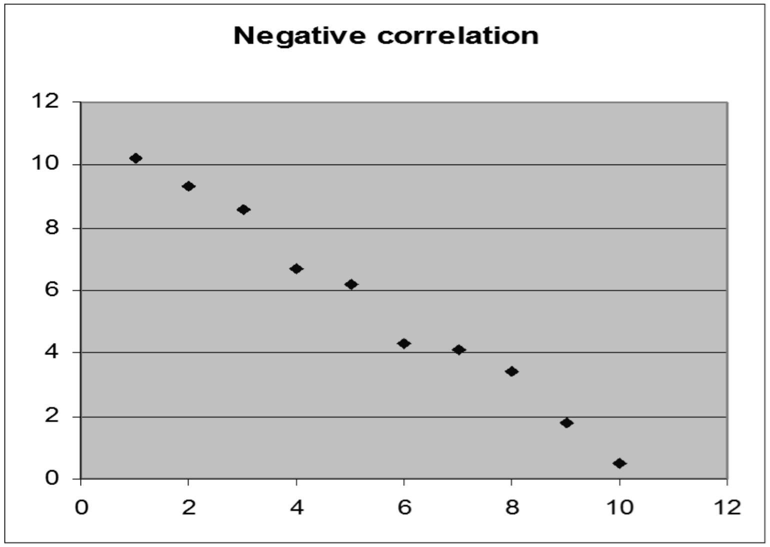
- When we are investigating the relationship between two dichotomous variables, we often construct a 2x2 contingency table
  - Is there an effect on gender on the likelihood of success of a treatment?
- We will discuss these in detail in Lecture 6

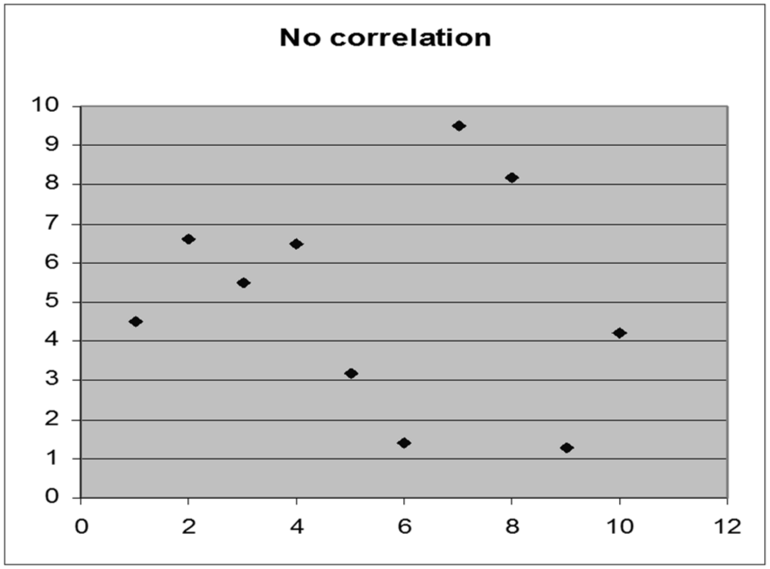
	Success	Failure	Total
Male	a	b	n_1
Female	c	d	n_2
Total	m_1	m_2	N

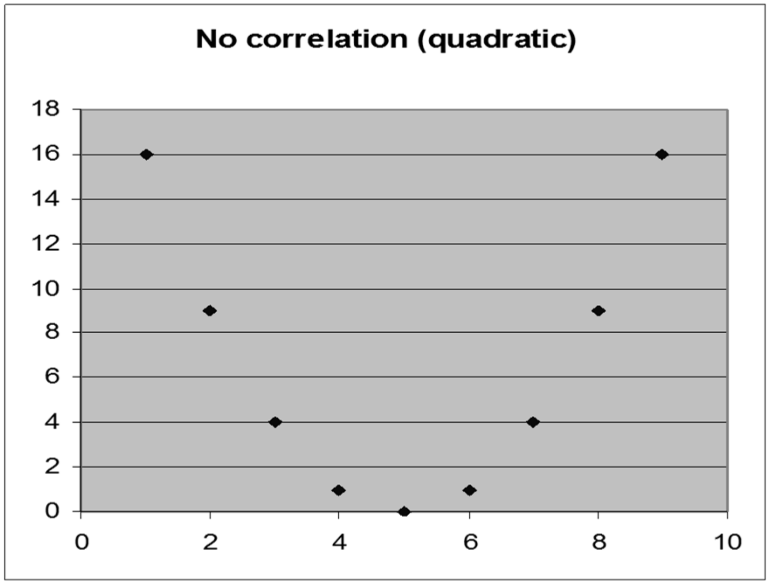
Correlation

- When we are investigating the relationship between two continuous variables, we often describe the correlation between two variables (Lecture 7)
- A positive correlation implies that as one variable goes up, the other variable also goes up (ex. Height and weight)
- A negative correlation implies that as one variable goes up, the other variable goes down (ex. Age and hearing)


Positive correlation
12
10
8
6
4
2
0
0
2
4
6
8
10
12

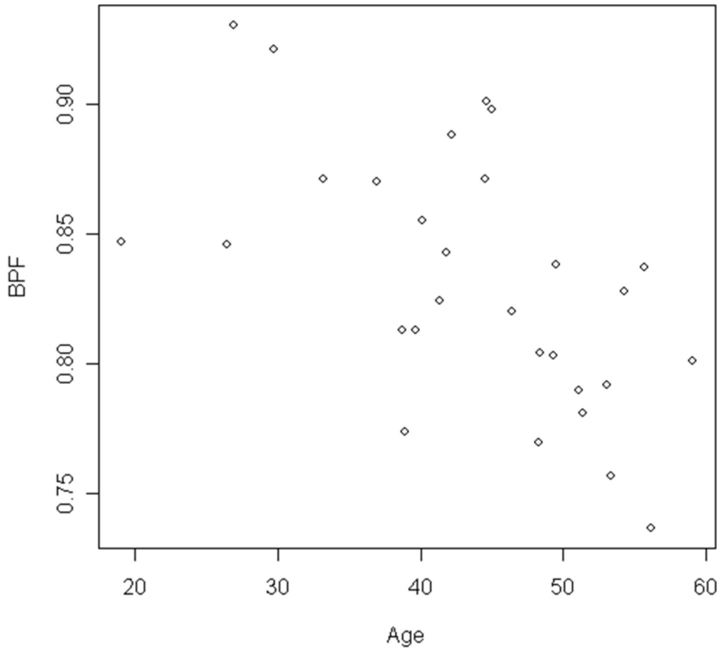

Negative correlation
12
10
8
6
4
2
0
0
2
4
6
8
10
12


No correlation
10
9
8
7
6
5
4
3
2
1
0
0
2
4
6
8
10
12

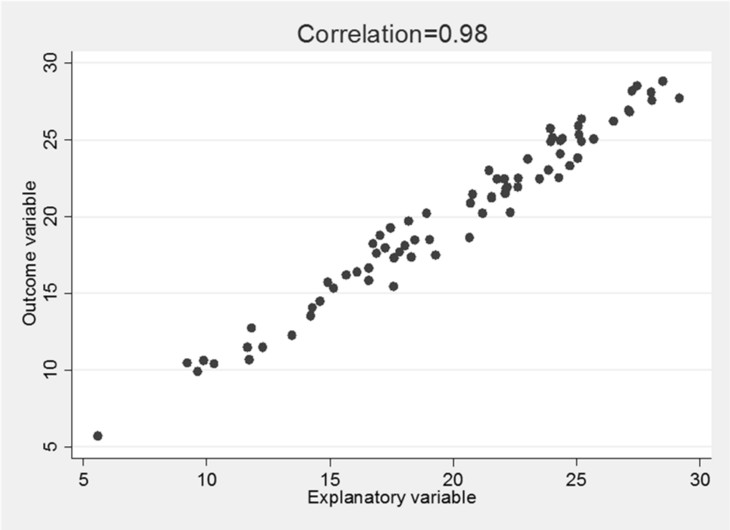

No correlation (quadratic)
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10

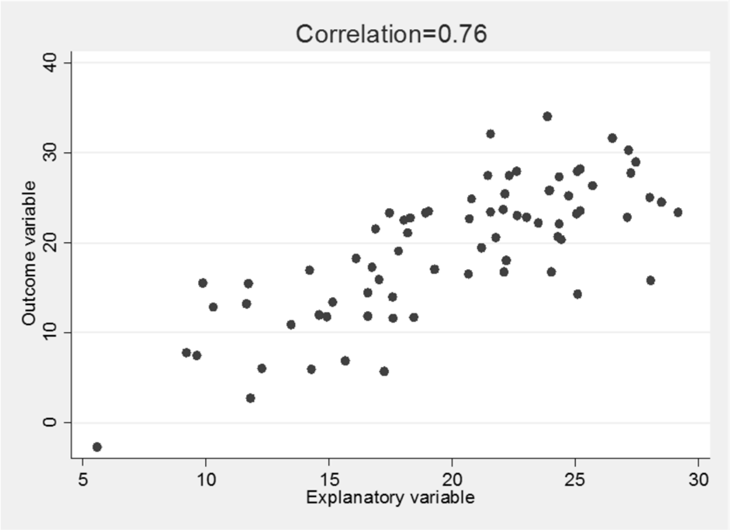
Scatterplot

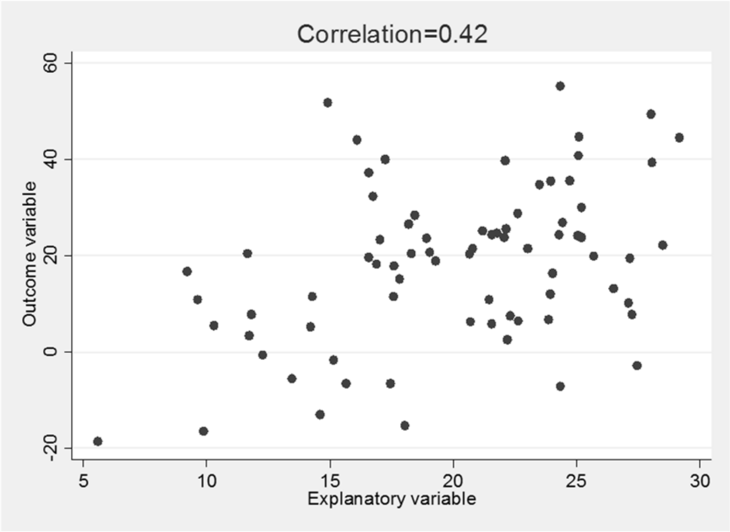
- Scatterplot is the best way to see the relationship between two continuous variables

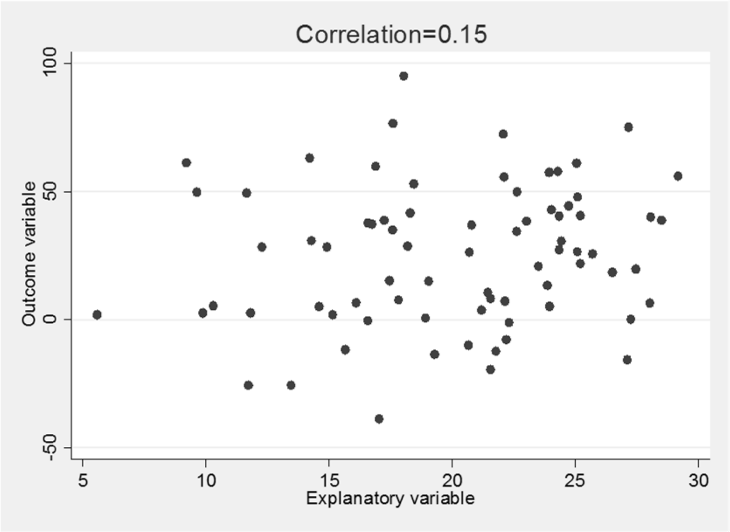


Level of correlation


Correlation=0.98


Correlation=0.76


Correlation=0.42


Correlation=0.15

Quick thoughts

- “Correlation does not equal causation”
- “Statistically significant correlation” and “important/meaningful correlation” are not the same
  - With a large sample, correlation coefficient of 0.1 can be statistically significant. At the same time, this weak correlation may be important clinically or it may not be. This is the one instance where “statistically significant” and “clinically meaningful” may not be the same

Things to come

- Basics of probability
  - How we make an inference
- Confidence intervals
- Types of analysis
  - One sample tests
  - t-test
  - ANOVA
  - Linear regression

Types of analysis-independent samples

Outcome	Explanatory	Analysis
Continuous	Dichotomous	t-test, Wilcoxon test
Continuous	Categorical	ANOVA, linear regression
Continuous	Continuous	Correlation, linear regression
Dichotomous	Dichotomous	Chi-square test, logistic regression
Dichotomous	Continuous	Logistic regression
Time to event	Dichotomous	Log-rank test

What we learned

- During this lecture we discussed
  - Different types of data
  - Ways to summarize and visualize each type of data
  - Interpretation of scatterplots