Data Visualization and Basic Statistical Testing

Kimberly Greco, MPH

View course video.

Boston Children's Hospital
Until every child is well℠

Course Overview

Course Objective

Provide a foundation in the basic statistical methods and principles necessary to understand, interpret, and communicate insights from data.

Course Structure

Lecture 1: Getting to Know Your Data: Types of Data and Descriptive Statistics

Lecture 2: Sampling Concepts and Comparing Two Means

Lecture 3: Linear Models and Correlation

Lecture 4: Comparing Proportions and Measures of Association

Lecture Outline

- ❑ **Types of Data** *(qualitative vs. quantitative)*
- ❑ **Summarizing Data**
  - *Graphical Methods*
  - *Numerical Summary Measures*
- ❑ **Confidence Intervals and P-Values**

The Research Process

Research Question

Hypotheses

Study Design

Data Collection

Data Analysis

Evidence-Based Information

The Research Process

Research Question

Hypotheses

Study Design

Data Collection

Data Analysis

Evidence-Based Information

A clear and focused research question will guide the research process. Every decision regarding study design, data collection, and data analysis should be connected to your research question.

The Research Process


Research Question


Hypotheses

Study Design


Data Collection


Data Analysis


Evidence-Based Information

A hypothesis is an attempt to answer your question with an explanation that can be tested. Hypothesis testing is central to the statistical methods we will cover in this course.

The Research Process

Research Question

Hypotheses

Study Design

Data Collection

Data Analysis

Evidence-Based Information

Study design, data collection methods, and analytic techniques are selected to address your research question.

The Research Process

Research Question

Hypotheses

Study Design

Data Collection

Data Analysis

Evidence-Based Information

Results are interpreted and effectively communicated through text, tables, and figures. Data has been successfully translated into actionable insights.

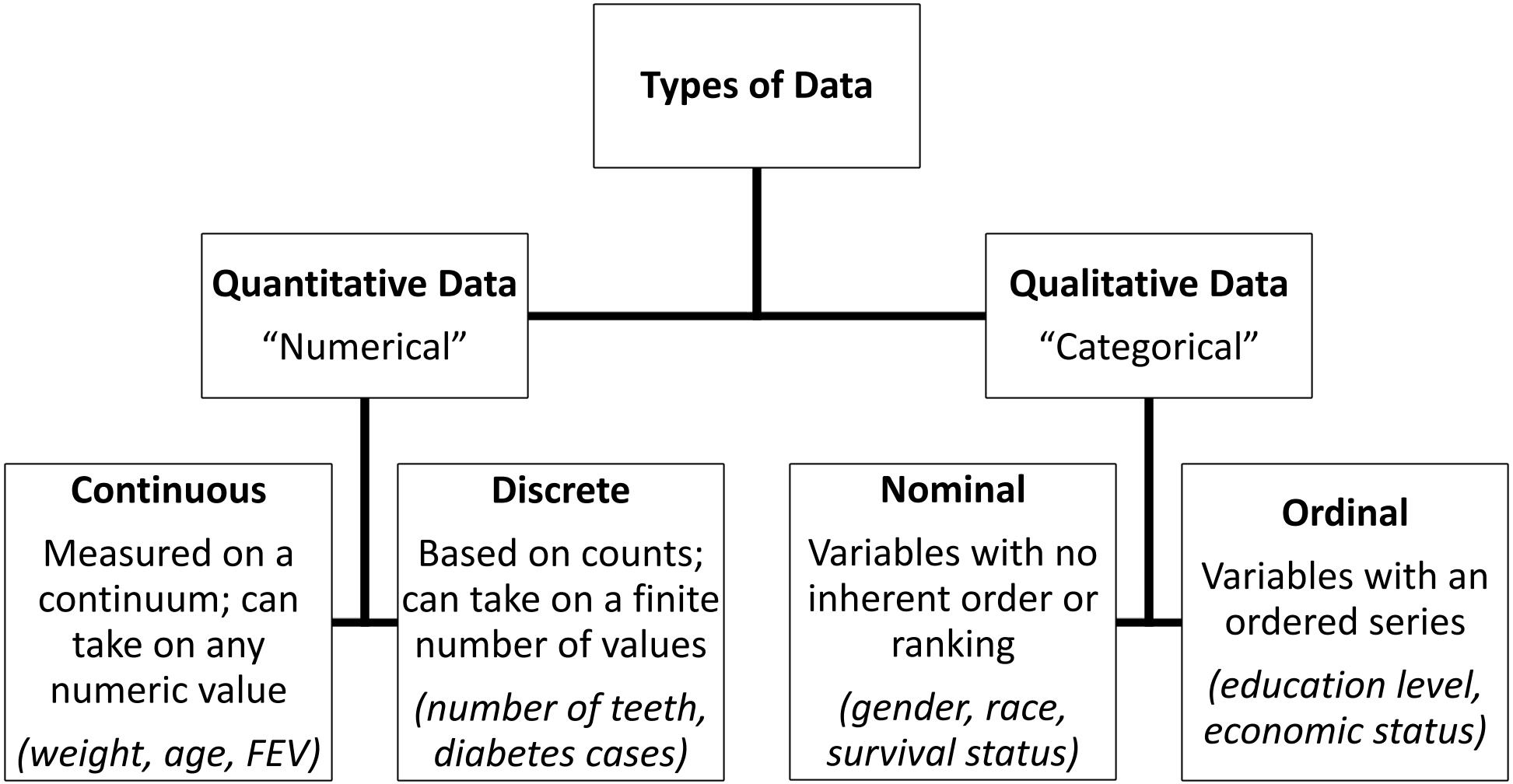
Types of Data

Before you begin any analysis, ask yourself:

What kind of data am I working with?

Data type determines…

- How you display the data in a table or graph
- How you summarize the data, graphically and numerically
- What statistical methods you use to analyze the data


Types of Data
Quantitative Data
"Numerical"
Qualitative Data
"Categorical"
Continuous
Measured on a continuum; can take on any numeric value
(weight, age, FEV)
Discrete
Based on counts; can take on a finite number of values
(number of teeth, diabetes cases)
Nominal
Variables with no inherent order or ranking
(gender, race, survival status)
Ordinal
Variables with an ordered series
(education level, economic status)

Categorical Data

Nominal Variables

Categories are not ordered. Two-level nominal variables are dichotomous.

Race	
1	White
2	Black
3	Asian
4	Other

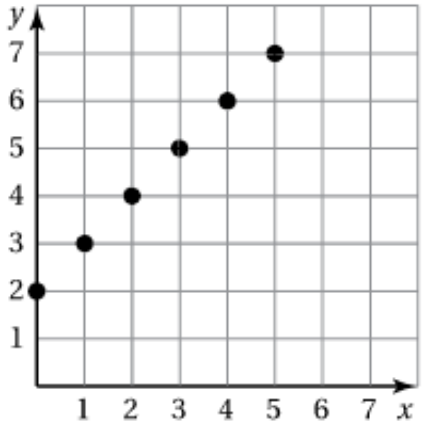
Hispanic	
0	No
1	Yes

Ordinal Variables

The order among the categories is important, but not the numerical values.

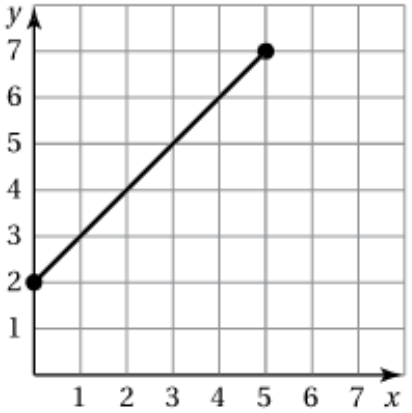
Education Level	
1	Less than high school
2	Graduated high school
3	Some college or associate
4	College degree or higher

Numerical Data



Discrete Variables

- *Order and magnitude are important*
- *Usually integers and counts*



Continuous Variables

- *Order and magnitude are still important*
- *Can take on any continuous value (not restricted to be integers)*

Grouping Continuous Data

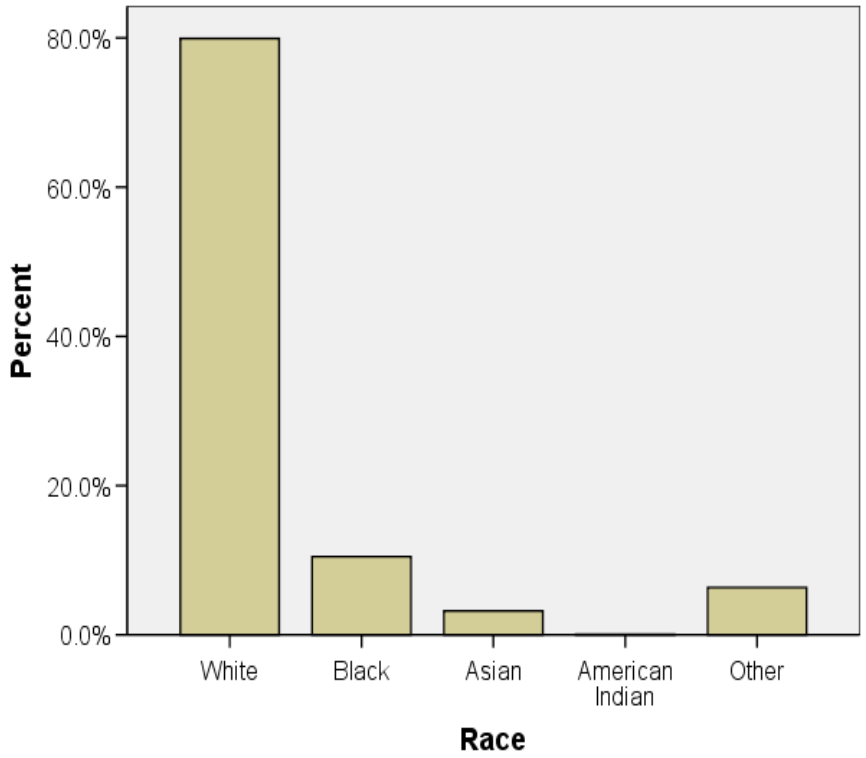
Continuous measurements can be transformed into ordinal or dichotomous ones.

Examples

- *BMI categorized into underweight, normal, overweight, or obese*
- *Symptom score categorized into mild, moderate, or severe*
- *Age categorized as younger vs. older*

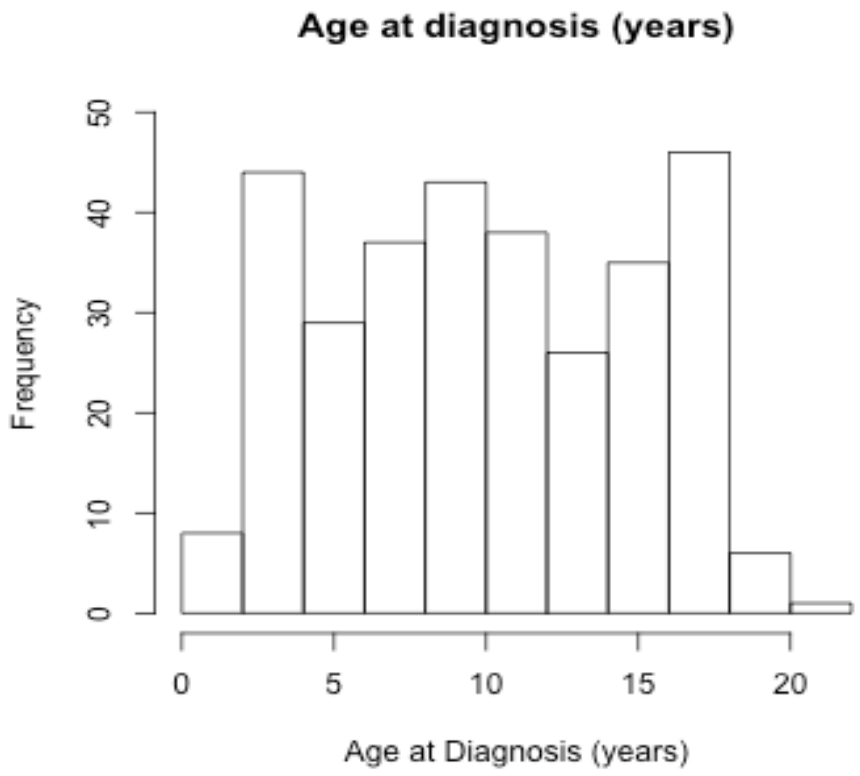
Note: Putting data into buckets leads to a loss of information, but it makes sense when clinically meaningful cut-points exist

Bar Plots



- Display the frequency distribution for **nominal** or **ordinal** data
  - **X-Axis** displays categories
  - **Y-Axis** displays N or % of cases in each category (should always start at zero)
- Bars should have equal width and be separate from each other so as not to imply continuity

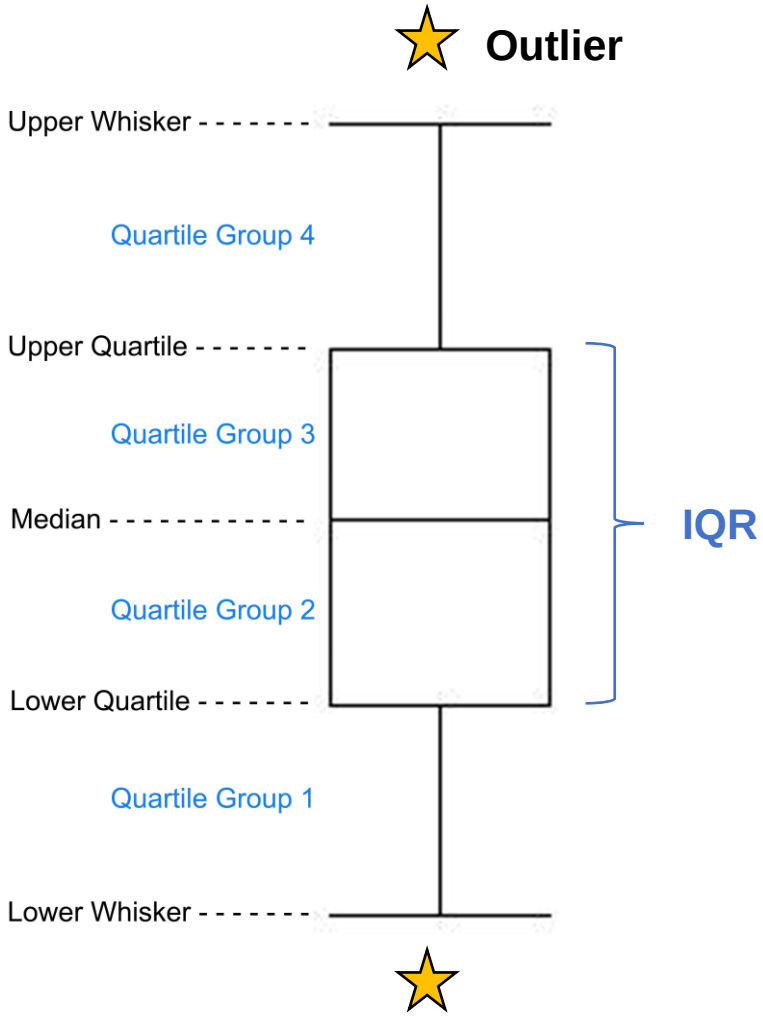
Histograms



- Display the frequency distribution for **discrete** or **continuous** data

- Similar to bar plots except…
  - **X-Axis** displays a range of values (instead of distinct categories)
  - **Y-Axis** displays N or % of cases in each interval/group of values
  - Bars are not separated from each other, since we assume "continuity" on the x-axis

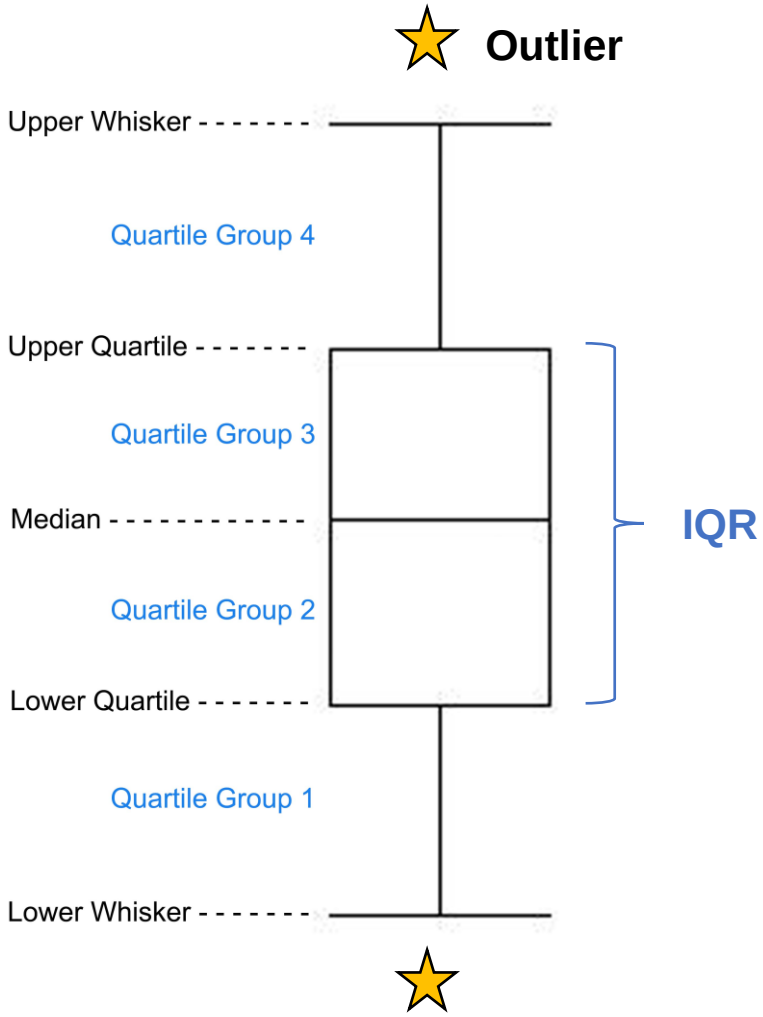
Box Plots



The "Box": Represents the inter-quartile range (IQR), or middle 50% of the data

- **Median**: Marks the mid-point of the data, where half the values are greater than or equal to this value and half are less than or equal to this value
- **Upper Quartile:** 75th percentile (75% of scores fall below the upper quartile)
- **Lower Quartile:** 25th percentile (25% of scores fall below the upper quartile)

Box Plots

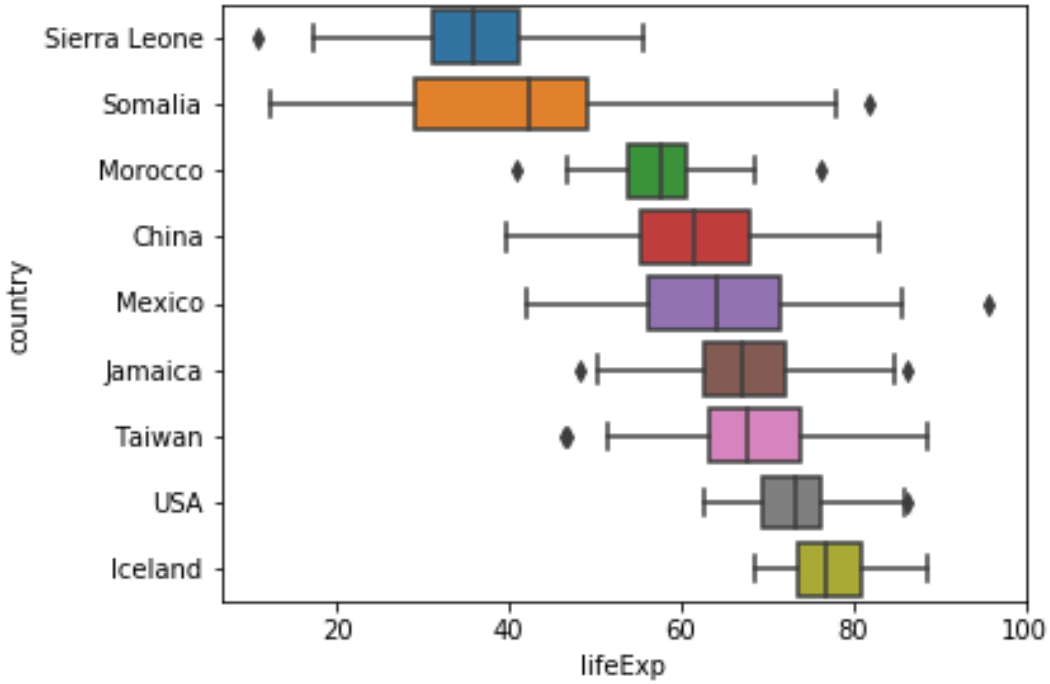


The "Whiskers": The definition of whiskers can vary. They may represent…

- The full extent of the data (maximum and minimum data points).
- The most extreme observations that are within 1.5*IQR.

Outliers are plotted outside the whiskers.

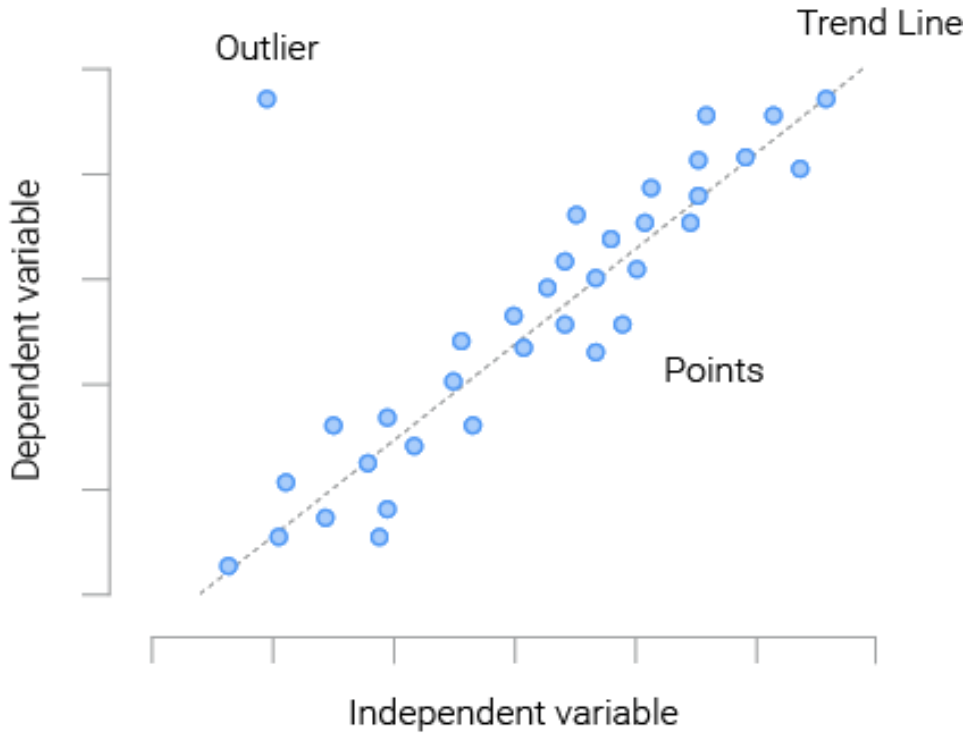
Box Plots



Box plots are useful for visualizing the distribution of variables.

- Presentation can be vertical or horizontal
- Length of the box indicates variation in percent agreement
  - For example, Morocco (green) and USA (gray) have less variation (shorter box → high % agreement) in life expectancy relative to Somalia (orange) and Mexico (purple)

Two-Way Scatter Plots



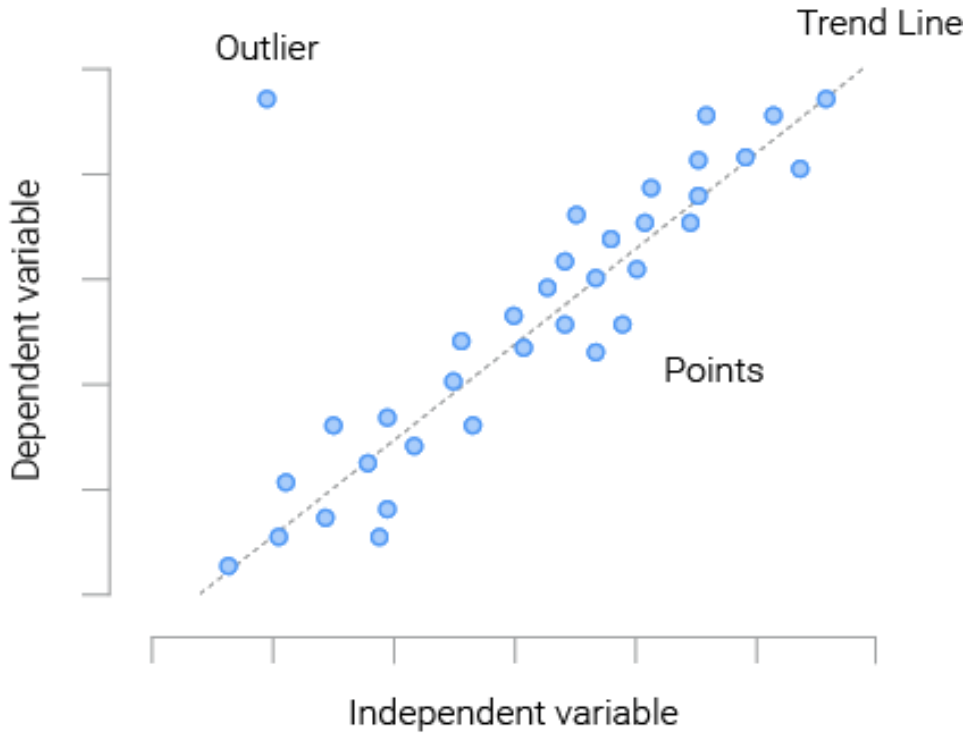
Scatter plots are used to show the relationship between two quantitative variables.

- Age vs. forced expiratory volume (FEV)
- Flow cytometry vs. image analysis (%)

Each point on graph represents a pair of values.

- x-axis is the scale for 1st variable (or independent variable)
- y-axis is the scale for 2nd variable (or dependent variable)

Two-Way Scatter Plot



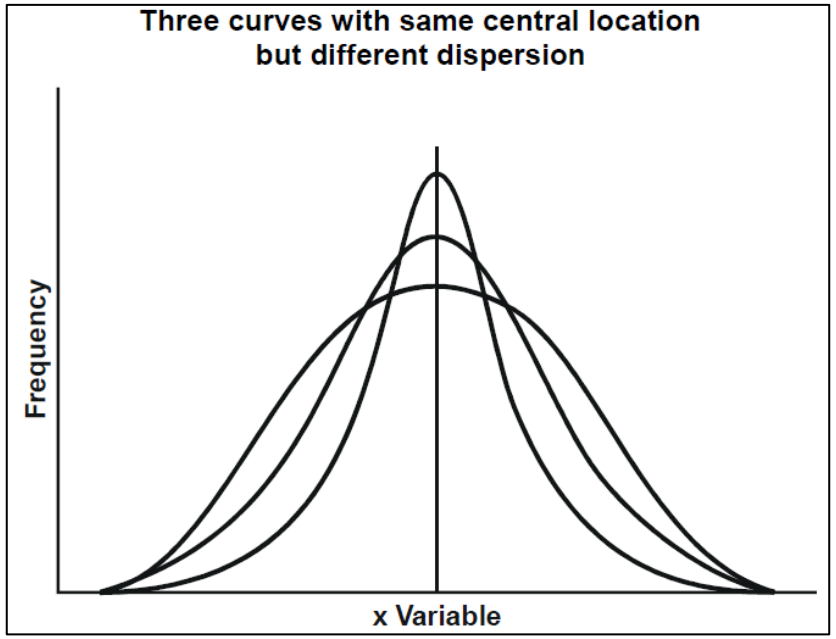
Strength of relationship is measured by how close points are to the x=y trend line.

- Close to line → strong correlation
- Further from line → weak correlation

Things to remember:

- If scale for 1st & 2nd variable are the same, then use same range on the x-axis and y-axis.
- If plotting 1st variable against more than one variable, keep axis of 1st variable consistent in all graphs.

Numerical Summary Measures



Measures of central tendency or location

- Mean, median, and mode

Measures of dispersion

- Quantify the variation among the values
- Range, interquartile range, standard deviation and variance

Numerical Summary Measures

Natural way to convey information

- "On average most people exercise 3-4 times per week"
- "Most people would say that we have had a hot summer"

Where is the "center" of the data?

- Mean → Average observation
- Median → Middle observation
- Mode → Most frequently occurring value

Measures of Central Tendency: **Mean**

- **Calculated by summing all values and dividing by the number of observations**
  - Takes the magnitude of each value into account
- **Most commonly used measure of central tendency**
  - Used for discrete and continuous data
  - In general not appropriate for categorical data
    - Exception - dichotomous data. Assume cases alive have value of 1 and cases died value of 0.
    - Mean = proportion of cases with value of 1 = proportion of cases alive

Measures of Central Tendency: **Mean**

The mean is influenced by extreme values (outliers)

Q: What to do in this case? Can we just delete these outliers?

A: No, we need measures that are not sensitive to outliers. One such measure is the median.

Measures of Central Tendency: **Median**

- **Defined as the 50th percentile**
  - Rank observations from smallest to largest
  - Half of the values are $\geq$ median and other half are $\leq$ median
  - Does not take the magnitude of each value into account, just the position
- **Equal to "center" of the ranked values**
  - If odd number of observations → middle value
  - If even number of observations → average of 2 middle values
- **Used for ordinal, discrete, and continuous data**

Measures of Central Tendency: **Mode**

- Defined as the value or observation that occurs most frequently
- Used for all types of data

Mode →

	Post-Treatment Condition	n	%
1	Much improved	9	13%
2	**Slightly improved**	**28**	**39%**
3	Stays the same	16	23%
4	Slightly worse	12	17%
5	Much worse	6	8%
	Total	71	100%

Example: Forced Expiratory Volume (FEV)

Order	FEV
1	2.15
2	2.25
3	2.30
4	2.60
5	2.68
6	2.75
7	2.82
8	2.85
9	3.00
10	3.38
11	3.50
12	4.02
13	4.05
Sum	**38.35**

FEV Data

- 13 subjects
- Values are ordered from smallest to largest

Mean FEV

- N = 13
- Sum = 38.35
- Mean = (38.35 / 13) = 2.95

Example: Forced Expiratory Volume (FEV)

Order	FEV
1	2.15
2	2.25
3	2.30
4	2.60
5	2.68
6	2.75
7	2.82
8	2.85
9	3.00
10	3.38
11	3.50
12	4.02
13	4.05
Sum	**38.35**

Median FEV

- Order the data (smallest → largest)
- With 13 observations, median is the value of order = (13+1) / 2 = 7
- Median = 2.82

What is the Mode?

- All FEV values are unique → no mode

What if we have an even number of observations?

Example: Forced Expiratory Volume (FEV)

Order	FEV
1	2.15
2	2.25
3	2.30
4	2.60
5	2.68
6	2.75
7	2.82
8	2.85
9	3.00
10	3.38
11	3.50
12	4.02

Median FEV

- Order the data
- With 12 observations, median is average of 6^{th} and 7^{th} values

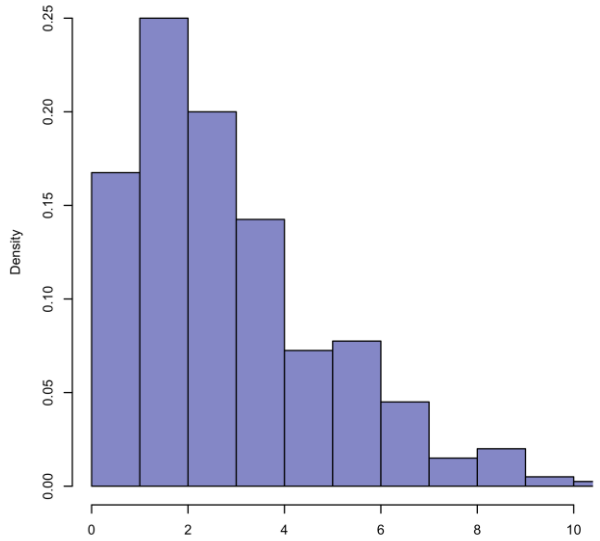
Median = (2.75 + 2.82) / 2

Median = 2.785

Best measure to report?

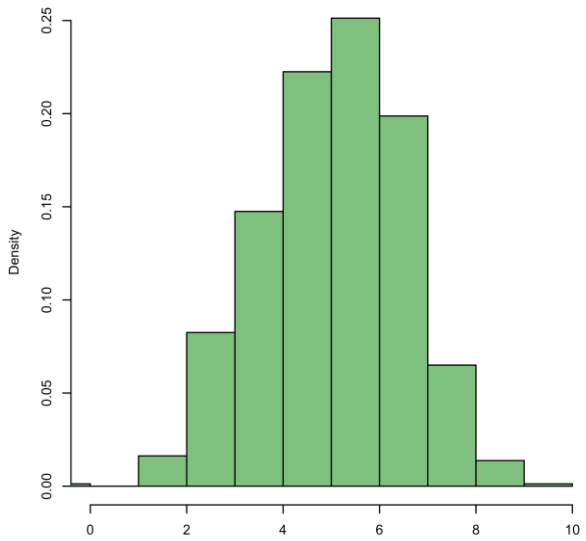
- **Depends on the type of data and the way in which the values are distributed**
  - If data is symmetric and unimodal (i.e., only 1 peak), then mean, median, and mode should all be roughly equal
  - If data is skewed, then the mean is sensitive to extreme values → the median might be more appropriate
- **What do I mean by skewed data?**

Symmetric vs. Skewed



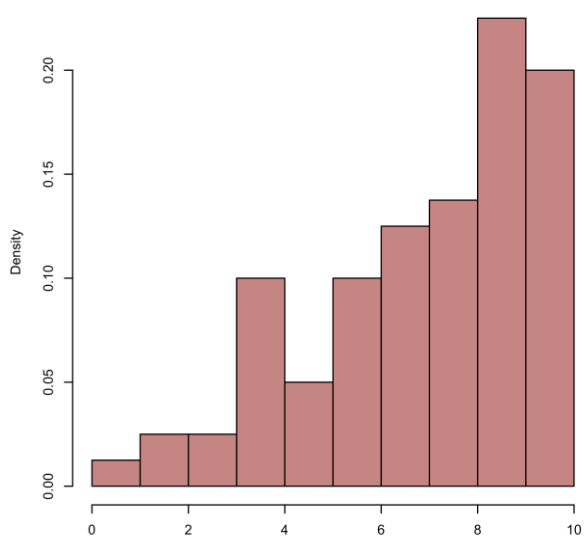
Skewed to right

- Mostly low values
- Median might be more appropriate



Symmetric

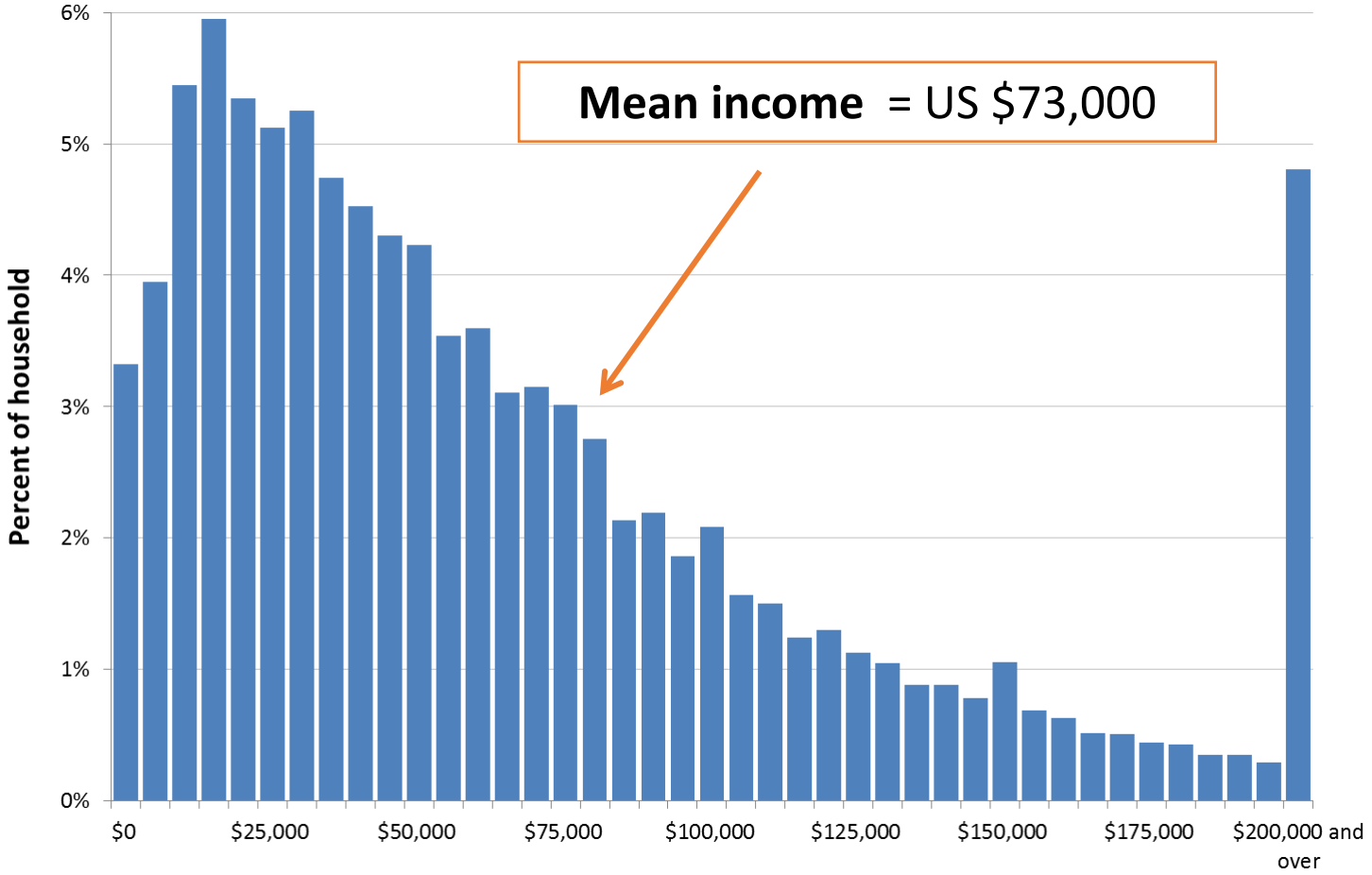
- Mean and median would be about the same



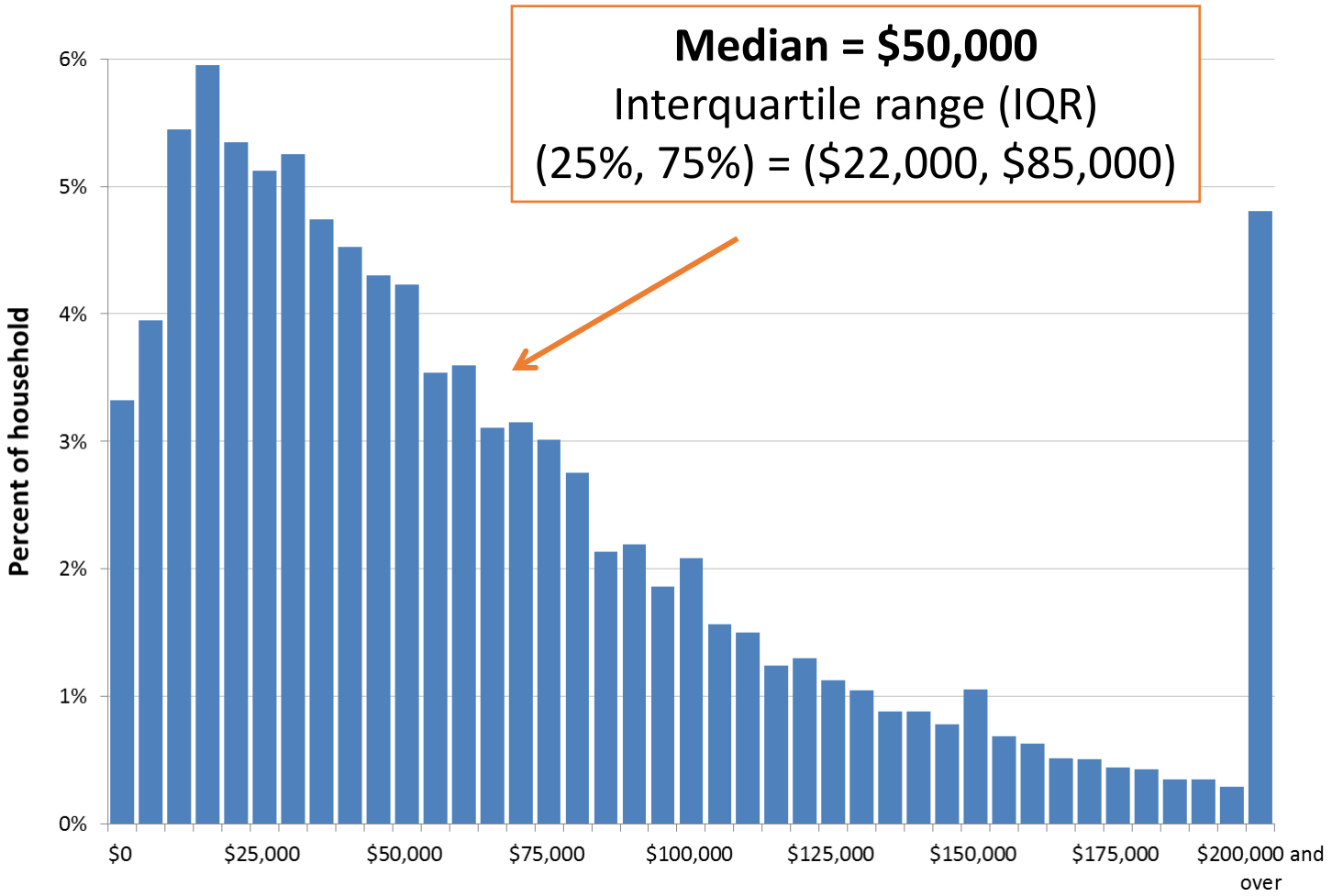
Skewed to left

- Mostly high values
- Median might be more appropriate

Example of skewed distribution: Household income



Example of skewed distribution: Household income



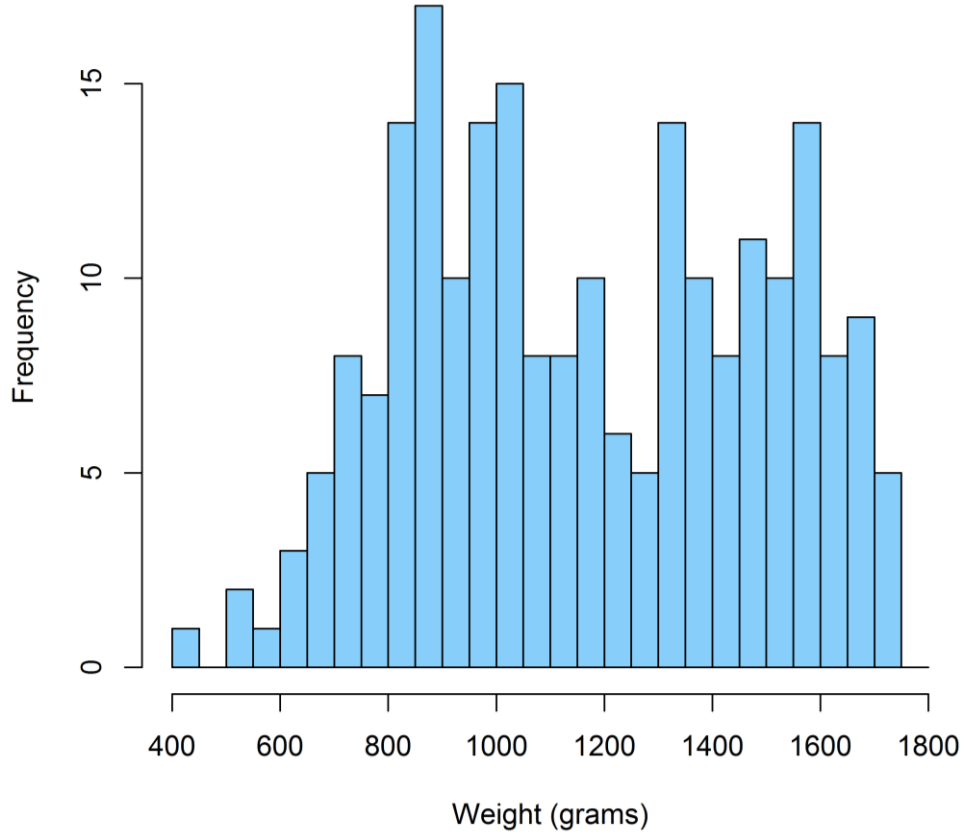
Important to use BOTH numerical measures & graphical methods

Example: Distribution of birth weight among infants (<1750 g)

Mean=1,173
Median=1,140

Mean ≈ Median, so we expect a symmetric distribution

What do you think about the histogram?

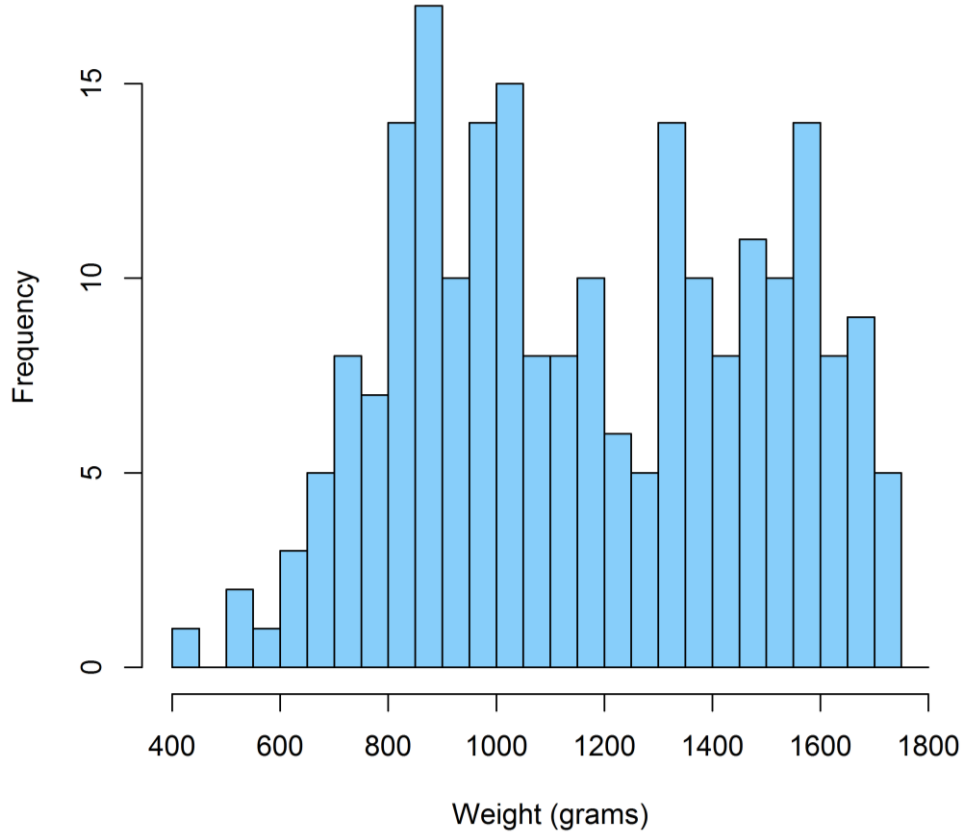


Important to use BOTH numerical measures & graphical methods

Example: Distribution of birth weight among infants (<1750 g)

Mean=1,173
Median=1,140

Mean ≈ Median, so we expect a symmetric distribution



What do you think about the histogram?

- Bimodal distribution
- Low and high weight appear to have different centers

Measures of Dispersion: **Range**

Range: Difference in smallest and largest values

- Easy to compute but not very useful because…
  - Uses only extreme data to calculate
  - Highly sensitive to very small and large values

An alternative…

Interquartile Range (IQR): Difference in 25th and 75th percentiles

- Includes middle 50% of the data
- Less sensitive to very small and large values

Measures of Dispersion: **Variance**

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Variance

- Indicates… average degree to which each point differs from the mean
- Calculated as… average of squared distance between observed value and mean value

Standard Deviation

- Indicates… how spread out a group of numbers is from the mean
- Calculated as… the square root of the variance
- Standard deviation is used because it has the same units of measurement as the mean

Example: Calculating Variance & SD

FEV	$(x_i - \bar{x})$
2.30	-0.65
2.15	-0.80
3.50	0.55
2.60	-0.35
2.75	-0.20
2.82	-0.13
4.05	1.10
2.25	-0.70
2.68	-0.27
3.00	0.05
4.02	1.07
2.85	-0.10
3.38	0.43
Total	

Step 1: Calculate difference between each observation and the mean.

N = 13

Mean = 2.95

Example: Calculating Variance & SD

FEV	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2.30	-0.65	0.423
2.15	-0.80	0.640
3.50	0.55	0.303
2.60	-0.35	0.123
2.75	-0.20	0.040
2.82	-0.13	0.169
4.05	1.10	1.210
2.25	-0.70	0.490
2.68	-0.27	0.073
3.00	0.05	0.003
4.02	1.07	1.145
2.85	-0.10	0.010
3.38	0.43	0.185
Total	0.0	4.66

Step 2: Square the difference between each observation and the mean. Sum.

Step 3: Divide by number of observations minus 1.

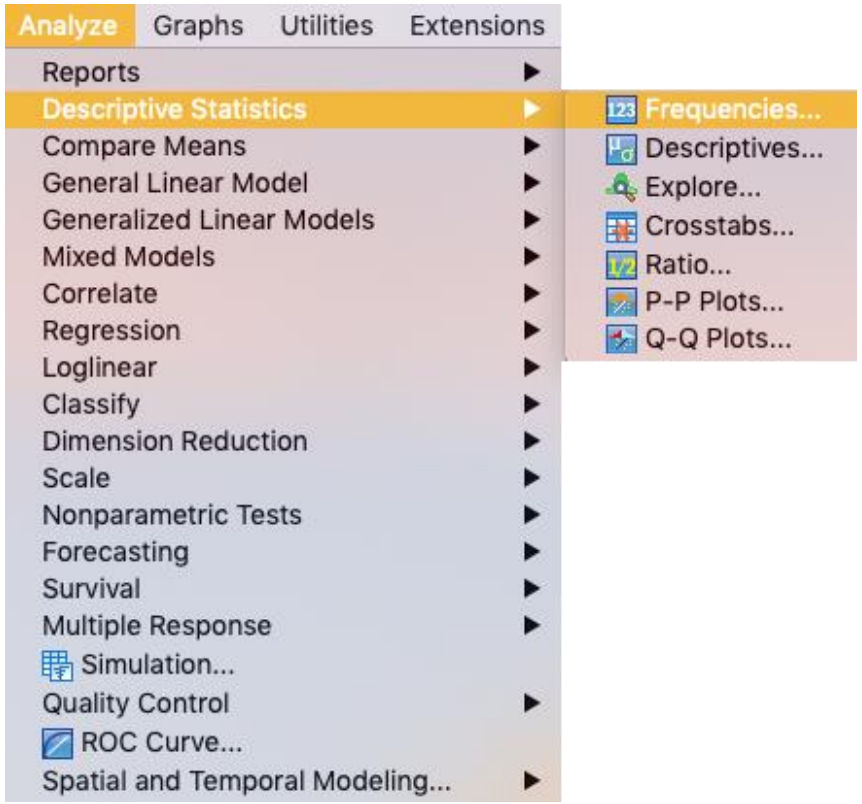
Variance = 4.66 / (13-1)
= 0.388

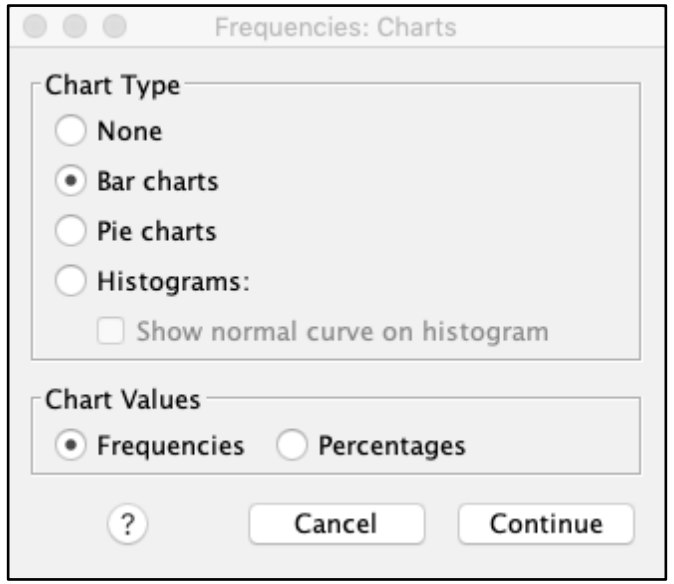
Step 4: Calculate standard deviation as square root of the variance.

Standard Deviation = $\sqrt{0.388}$
= 0.623

Summarizing Data in SPSS (Graphical)

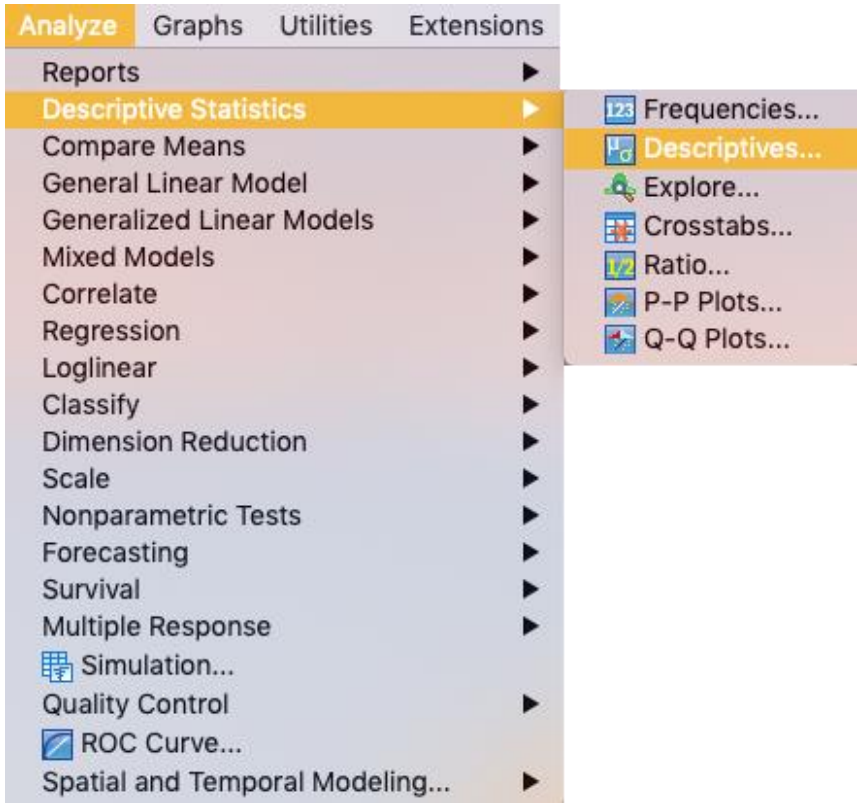
SPSS: Analyze > Descriptive Statistics > Frequencies

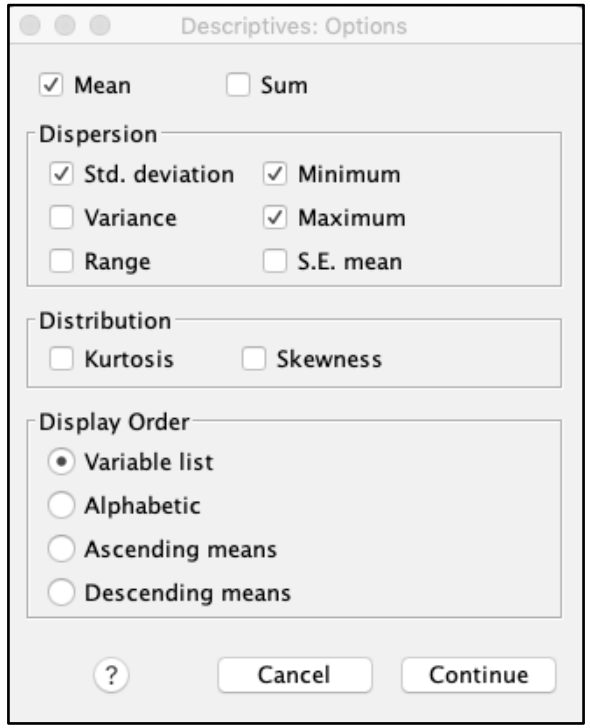




Summarizing Data in SPSS (Numerical)

SPSS: Analyze > Descriptive Statistics > Descriptives





Boston Children's Hospital
Until every child is well™

Presenting Numerical Summaries

Continuous Variables:
Mean ± SD
Median (IQR)

Categorical Variables:
N (%)

Variables	Treatment (N=628)	Control (N=372)
Age		
N (Nmiss)	628 (0)	372 (0)
Mean ± SD	41.6±7.6	48.8±7.4
Min–Max	29.0–59.0	29.0–61.0
Median (IQR)	40.5 (35.5–48.0)	50.0 (44.5–55.0)
Gender (%)		
Female	414 (65.9)	173 (46.5)
Male	214 (34.1)	199 (53.5)
Weight		
N (Nmiss)	628 (0)	370 (2)
Mean ± SD	148.8±26.4	157.3±28.3
Min–Max	87.0–243.0	71.0–250.0
Median (IQR)	146.0 (129.5–165.0)	155.0 (137.0–175.0)

N, number of non-missing values; Nmiss, number of missing values; SD, standard deviation; IQR, interquartile range.

Important:
Always present…

1. **The number of observations (N) in the sample**
2. **The number of observations missing for each variable**

Questions?

What's next?

Now that we have mean and standard deviation, how do we know if results are "important" or "significant"?

Two common quantities are reported:

- Confidence intervals
- P-values

Why do we calculate and report these?

- To draw "inferences" about a population mean from our sample

Calculating a Confidence Interval (CI)

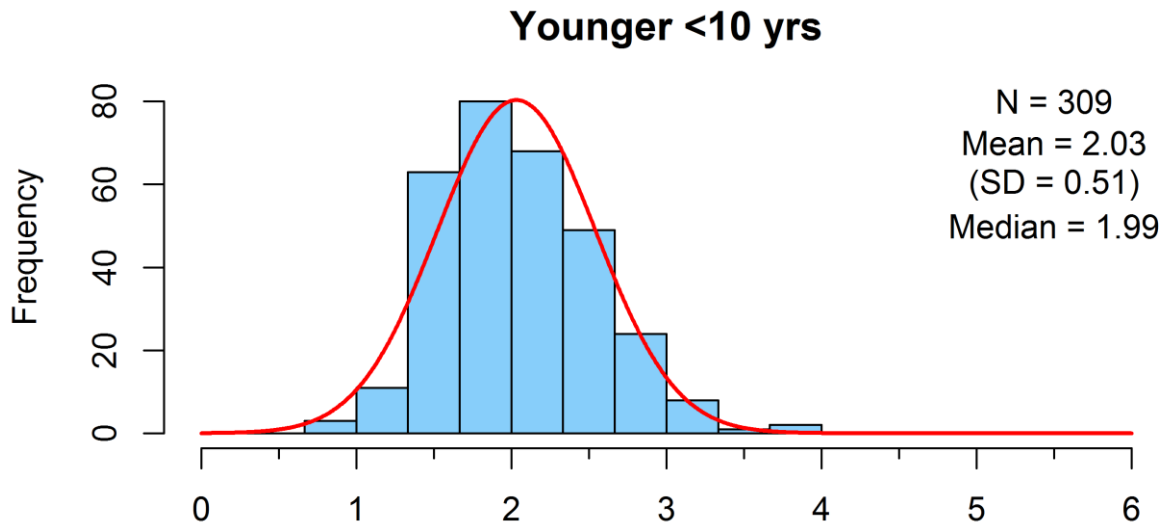
Confidence intervals indicate the precision of the sample mean with n subjects (reflects variability in sampling)

- Lower Limit = $mean - \left(Z_{critical} * \frac{SD}{\sqrt{n}}\right)$
- Upper Limit= $mean + \left(Z_{critical} * \frac{SD}{\sqrt{n}}\right)$

$Z_{critical}$ (also denoted $Z_{1-\alpha/2}$)

- Measures the number of standard errors to be added and subtracted from the mean in order to achieve a desired confidence level
- Equal to 1.96 for 95% confidence interval
- Obtained from a "reference" distribution (more on this next lecture…)

Example: Calculating a Confidence Interval

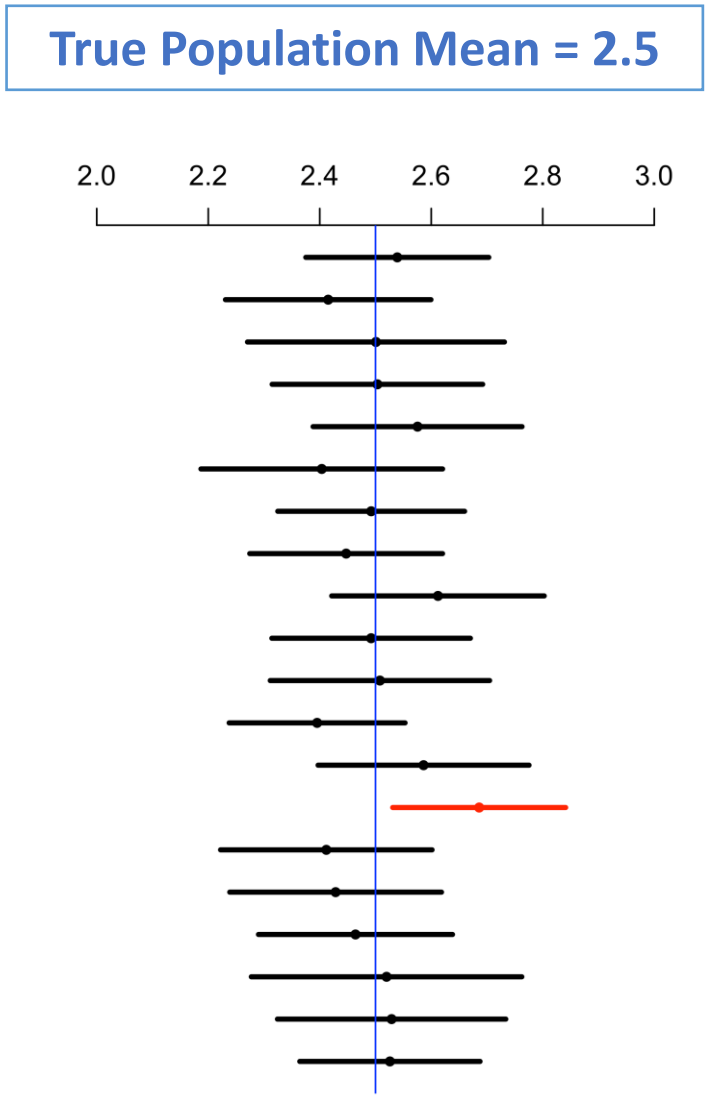


95% Lower Limit= $2.03 - 1.96 * 0.51/\sqrt{309} = 1.97$

95% Upper Limit= $2.03 + 1.96 * 0.51/\sqrt{309} = 2.09$

We'll come back to the interpretation…

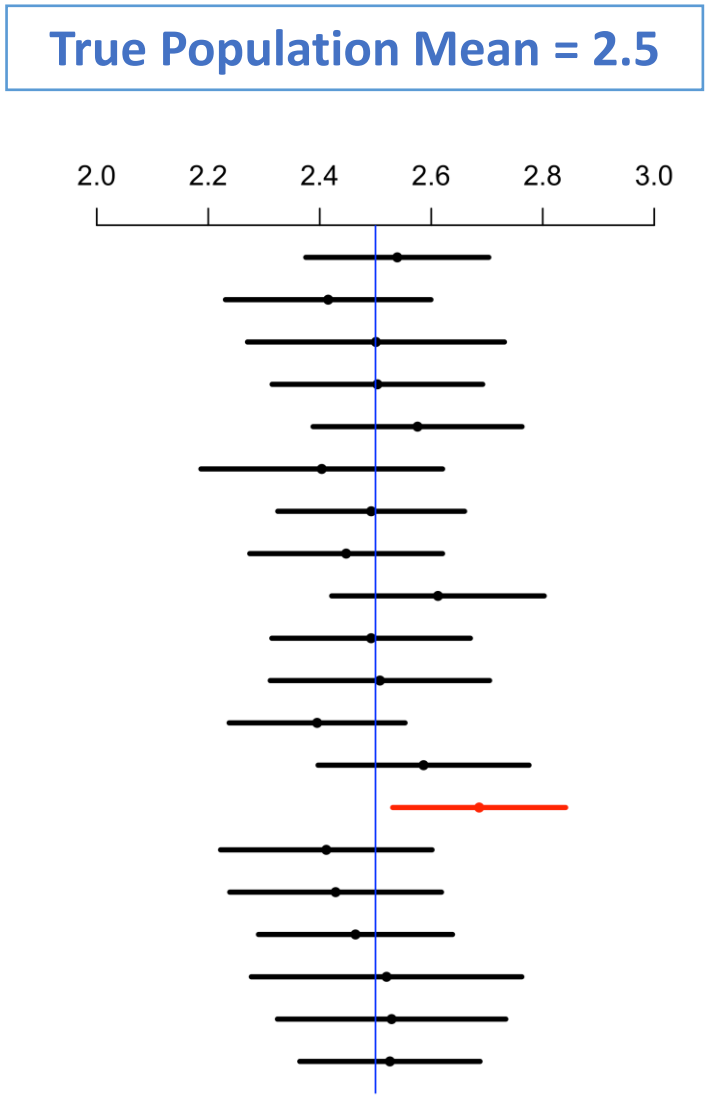
What does a confidence interval mean?



A fundamental task of biostatistics is to analyze samples in order to make inferences about the population from which the samples were drawn.

A confidence interval tells us how confident we can be that the results of a study reflect what we would expect to find if it were possible to survey the entire population.

What does a confidence interval mean?



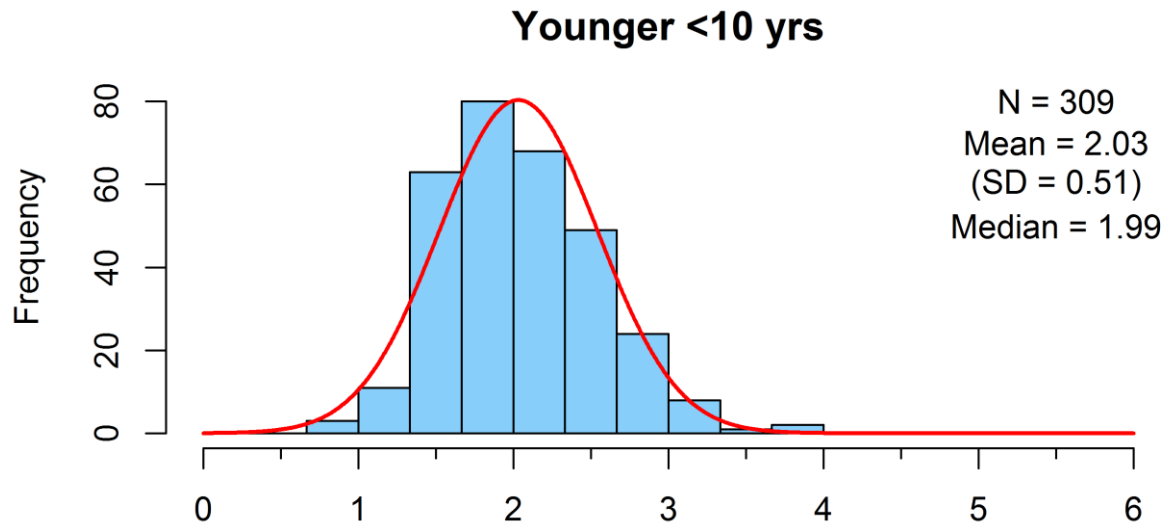
So, a 95% confidence interval is a range of values that we can be 95% certain contains the true mean of the population.

In practice…

- Repeat an experiment 20 times
- Calculate the 95% CIs for each experiment
- Expect 95% of the intervals to contain the true population mean ($\geq$ 19 of the intervals)

Boston Children's Hospital
Until every child is well™

Example: Calculating a Confidence Interval



Interpretation:
Based on our sample data, we are 95% confident that the true mean FEV in the population is between 1.97 and 2.09.

$$95\%\ \text{Lower Limit} = 2.03 - 1.96 * 0.51/\sqrt{309} = 1.97$$

$$95\%\ \text{Upper Limit} = 2.03 + 1.96 * 0.51/\sqrt{309} = 2.09$$

P-Values

- Another common quantity reported is the p-value
- Defined as the probability of observing a test statistic more extreme than the observed value by chance
- Values range between 0-1
  - Small p → result not likely due to chance
  - Large p → result is likely due to chance
- Tied to concept of hypothesis testing

P-Values & Hypothesis Testing

Hypothesis Testing

- Purpose: examine whether a difference (or an effect) is present or not by using statistical tests
- Decision:
  - "Yes, there is a significant difference…"
  - "No, there is not a significant difference …"
- P-values are tied to "significance" → $p < 0.05$ is a commonly used threshold for significance

Process of Hypothesis Testing

1

Define null (H_0) and alternative (H_1) hypotheses.

Example: Means of 2 groups: equal (H_0) vs. not equal (H_1)

2

Determine how different the observed data is from the null hypothesis →

calculate your test statistic

3

*Calculate the probability of observing a value ≥ **your test statistic** if the means are equal (i.e., H_0 is true) →*

p-value

4

Decide whether or not to reject the null hypothesis that the means are equal

Next Class

- To fully understand confidence intervals and p-values, we need to take a step back and briefly discuss a few concepts:
  - Sampling
  - Normal Distribution
  - Central Limit Theorem
- This will lead us into the comparison of two means

Questions?

kimberly.greco@childrens.harvard.edu