Statistics&Probability I

Tian Zheng
Department of Statistics
Data Science Institute
Columbia University

Introduction

STATISTICS & PROBABILITY

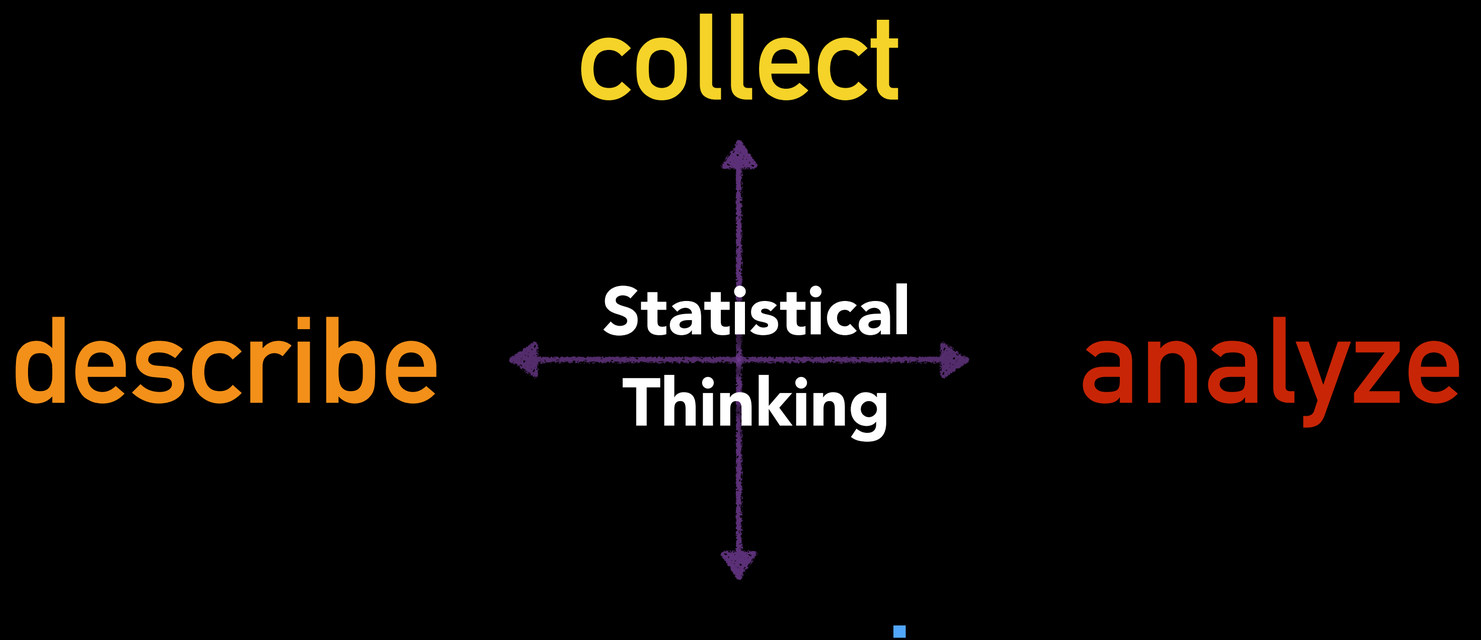
Variation

Chance

Randomness

STATISTICS & DATA

PROBABILITY THEORY & STATISTICAL METHODS



Statistical Thinking

Two Examples
from Today's News

Example 1
A sample survey

GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market

1025 randomly sampled Americans

TELEPHONE
SURVEY

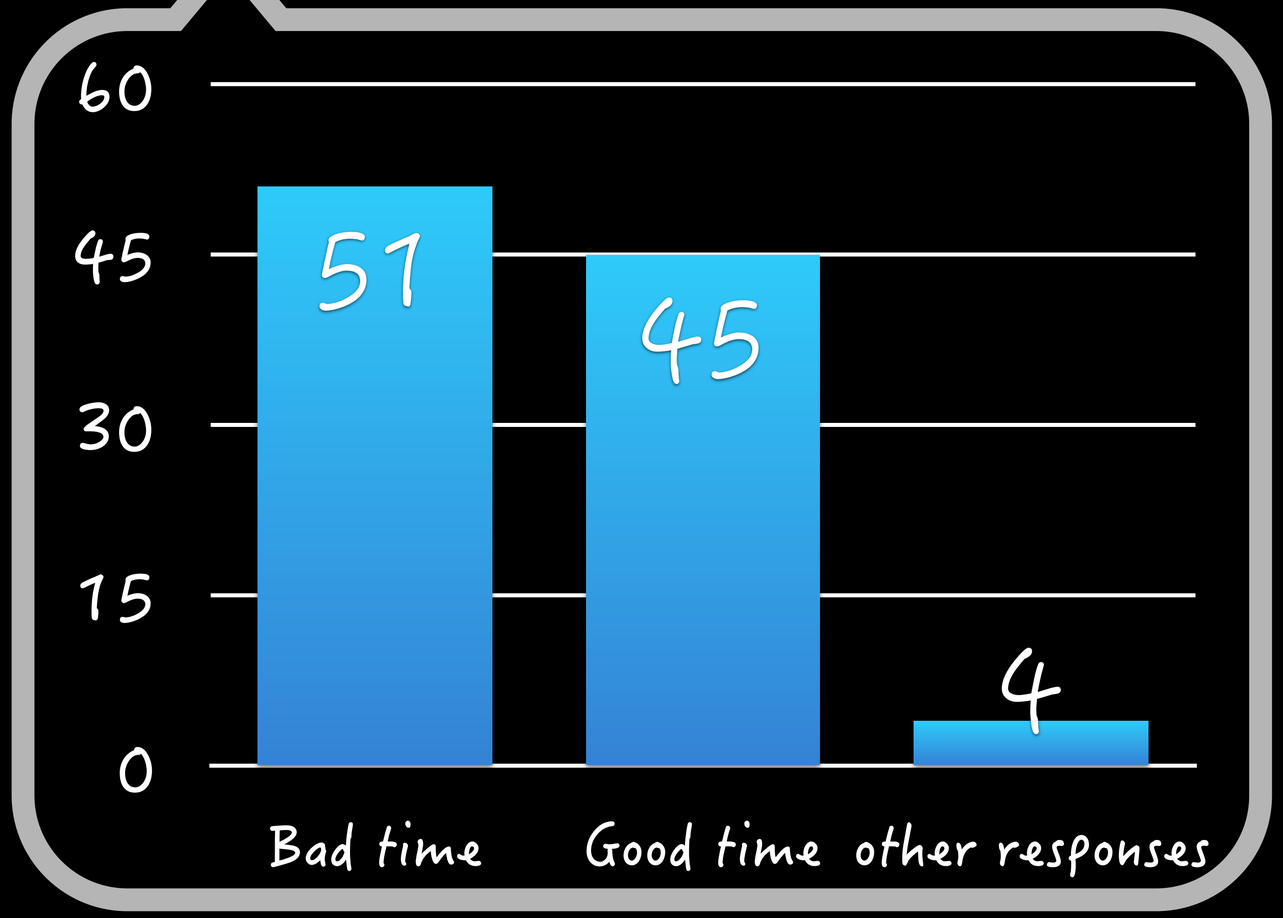
GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market

"Is it a GOOD TIME to FIND A JOB?"

GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market



GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market

Margin of error is
4% for the 95%
confidence interval

Margin of Error?

4%? 95%?

Confidence Interval?

STATISTICAL THINKING I

What was this study trying to find out?

Population of interest:
Americans

Information (variable) of interest:
Their perception on the job market

STATISTICAL THINKING I

STATISTICAL THINKING I

Why should we care about the opinion of the 1025 Americans who participated in this survey?

1025 Survey Participants

Representative?

300 Million Americans

STATISTICAL THINKING I

Statistics ***derives***

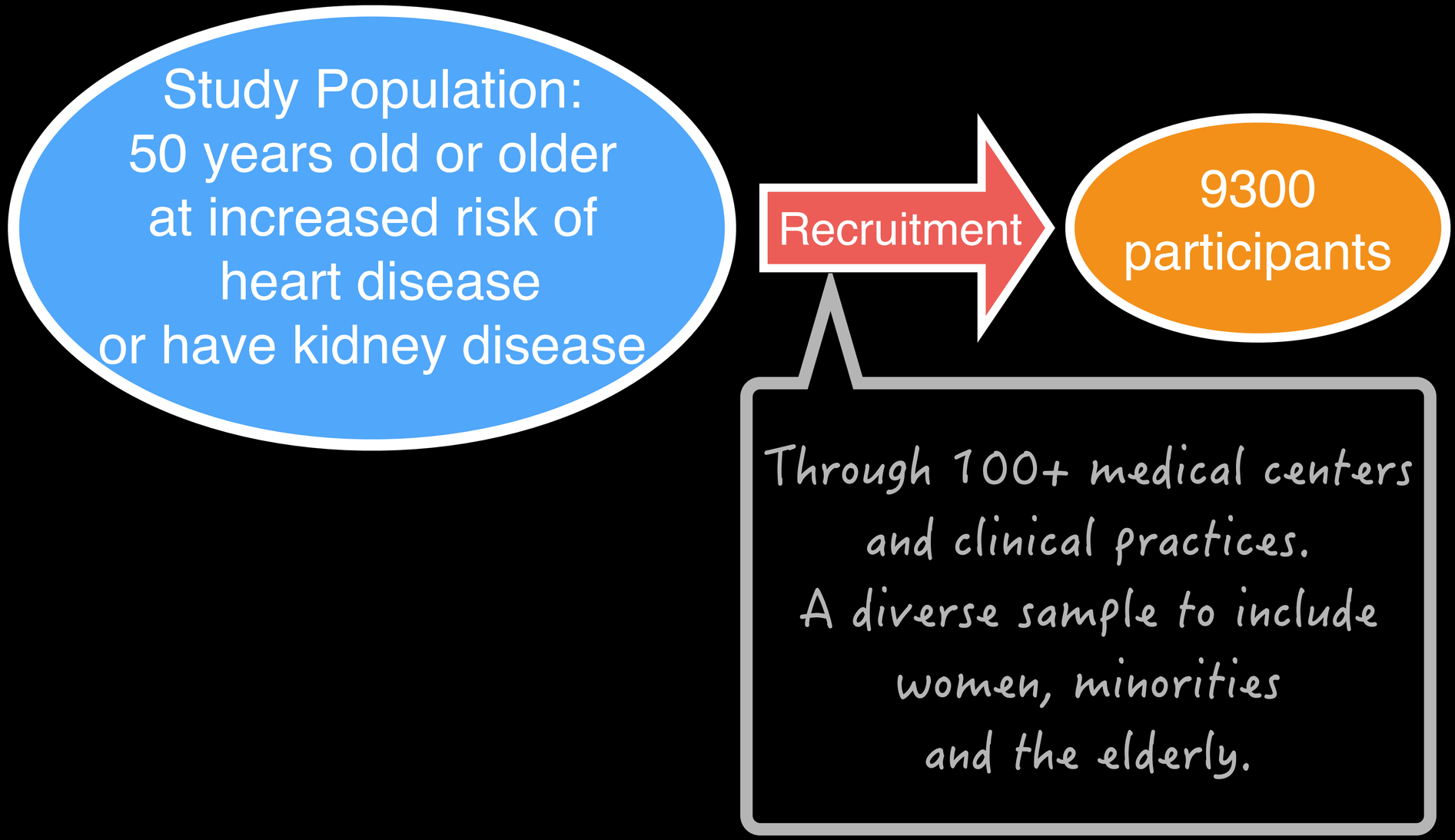
Knowledge

Sample  ***Population***

Example 2
A clinical trial

Example 2 - NIH News Release

Study participants



Study Design

Similar except for difference due to chance

9300 participants

Random assignment

Treatment group

New targeted level of blood pressure control

health results

Control group

Conventional targeted level of blood pressure control

health results

Difference

Due to chance?

Statistical thinking II

STATISTICAL THINKING 2

Statistics ***establishes***

Statistical Significance of observed signal

by studying randomness

Statistics&Probability I

Derive good answers from data

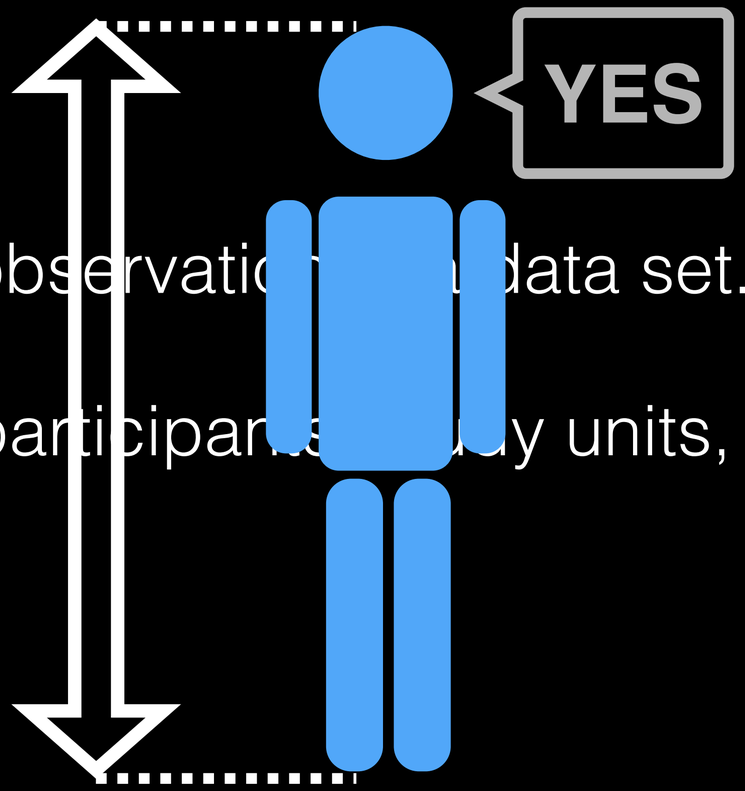
Data: numbers with context

- degrees — today's highest temperature
- years old — age of a cancer patient when the cancer was diagnosed
- pounds — weight of a 10 years old boy
- seconds — how long a study participant can hold the plank position

70

From Individuals to Statistics

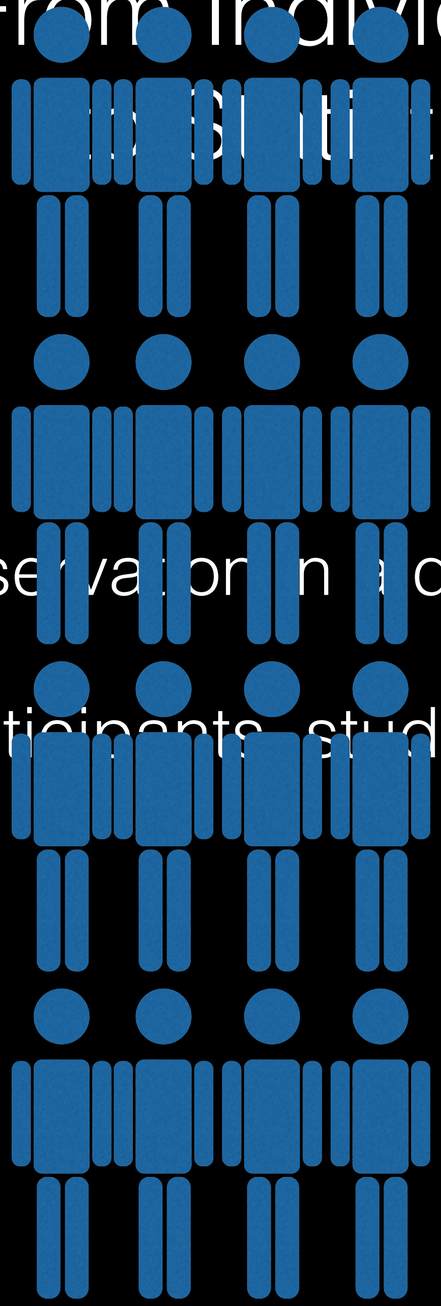
- Individuals: units of observation in a data set.
  - Also called study participants, study units, subjects, etc.



An individual in a study

- ***Gender:*** *Male*
- ***"Q:Is it a good time to find a job?":*** *"Yes"*
- ***Height:*** *70 inches*
- ***Education, income, social behaviors***

From Individuals to Statistics



- Individuals: units of observation in a dat
  - Also called study participants, study units, subjects, etc.

Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School
...			

From Individuals to Statistics

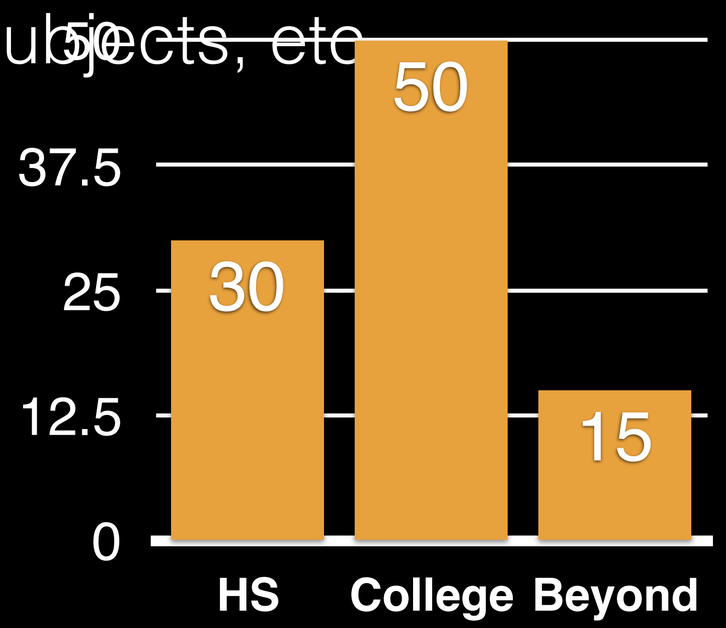
- **Categorical:** the values represent different categories for the individuals; do not have arithmetical meaning.
- Quantitative: the values represent numerical quantities that can be ordered and averaged.
- Ordinal: the values represent ordered categories; such as “how often do you exercise?”—*Everyday, frequently, sometimes, rarely, never.*

From Individuals to Statistics

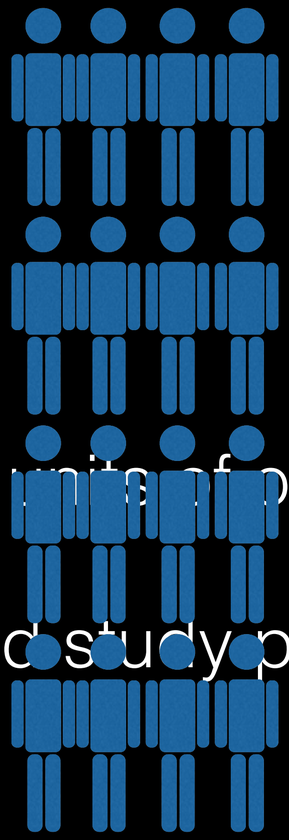
- Individua et.
  - Also called study participants, study units, subjects, etc

Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School
…			

High School	College	Beyond college
20	50	15

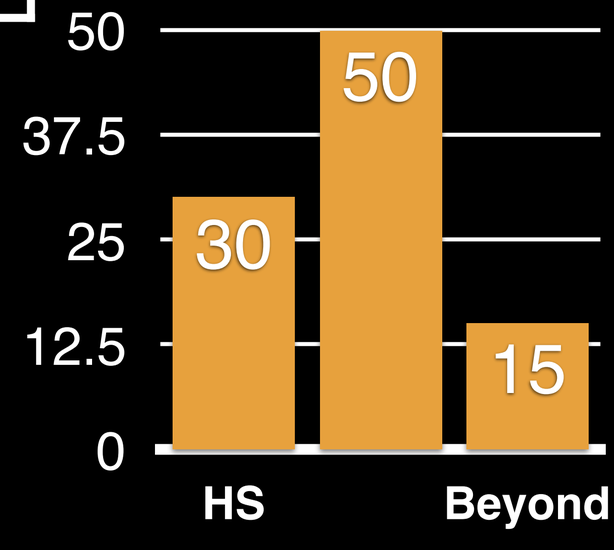


From Individuals to Statistics



- Individuals: units of observation in a data set.
  - Also called study participants, study units, subjects, etc.

Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School
...			



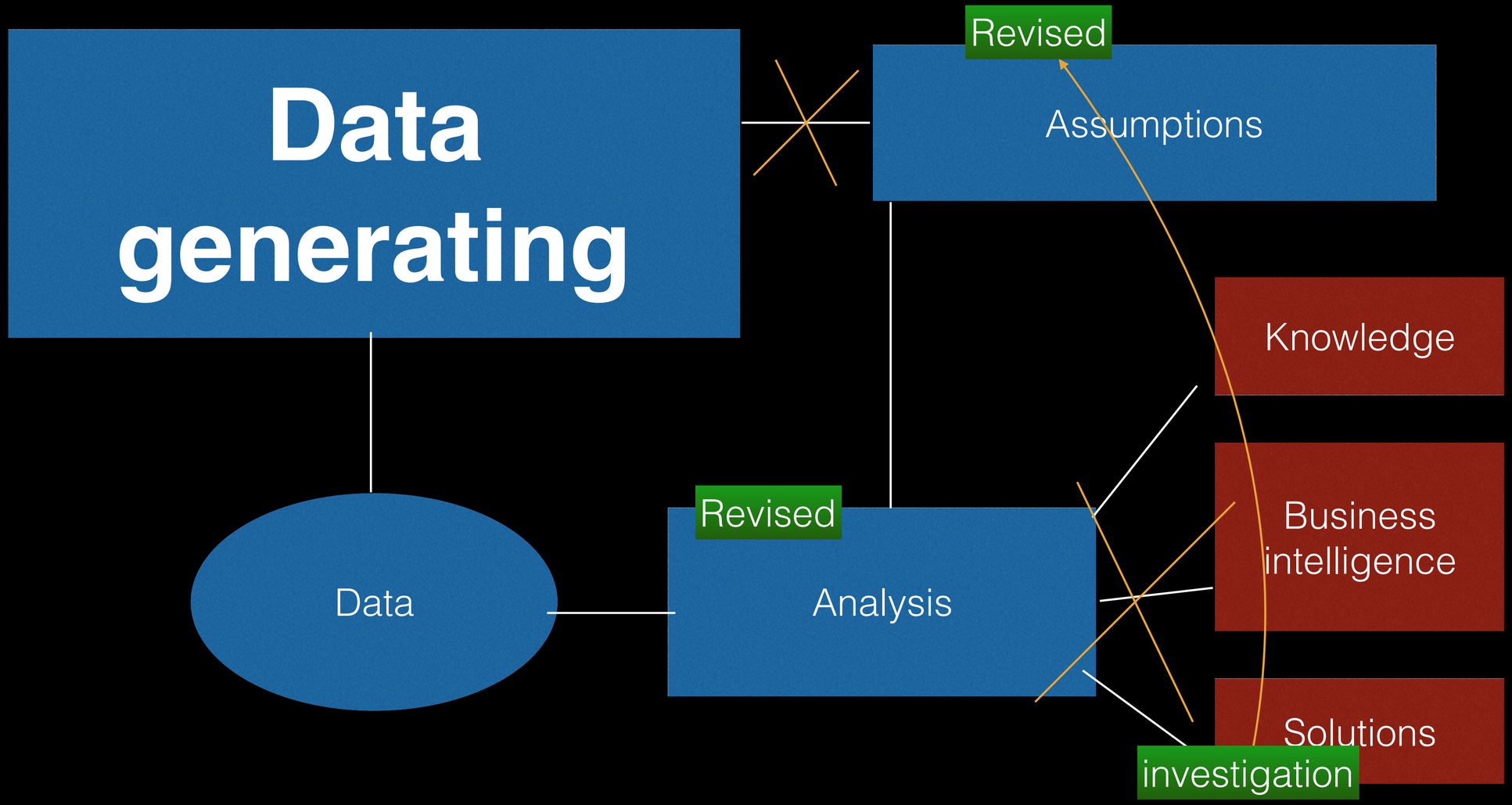
From Individuals to Statistics

- Individuals: units of observation in a data set.
  - Also called study participants, study units, subjects, etc.

Statistics are

- ★ summaries of numerical data
- ★ do NOT tell the whole story
- ★ useful and meaningful

From Data to answers



Garbage in, Garbage out?

- Validity of results depend on the validity of assumptions on the data generating process.
  - regarding the sampling, randomization, measurements, independence, etc
- They are often violated for big data.
- Data scientists investigate these assumptions and propose solutions.

Learning activity I:
understand mathematical notations

Notations

- **Statistics rely on computation of numerical summaries of data.**
- **Mathematical notation and equation formally describe such computation.**
- **Data are organized by individuals and variables.**
- **Variables, denoted by letters close to the end of the English alphabet such as X, Y.**
- **X with a subscript i is X's value for individual i.**
- **Summation sign.**

Statistics&Probability I

Display numerical data

Displaying categorial variable

- For categorical variable, we summarize the data using the **counts** of observed occurrences of each value.
- Alternatively, we can use ***percentage*** or *proportion*.

Pie chart

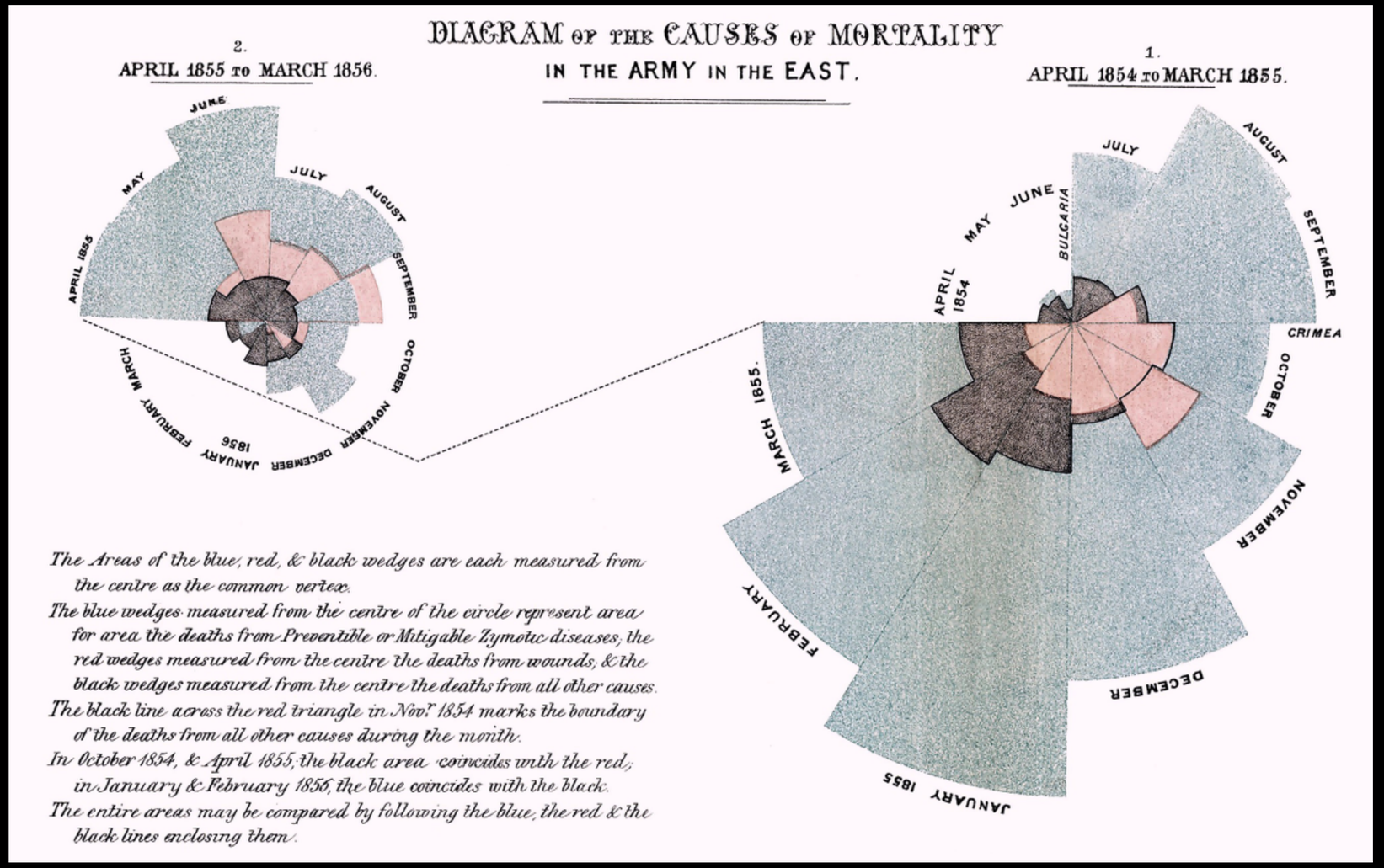
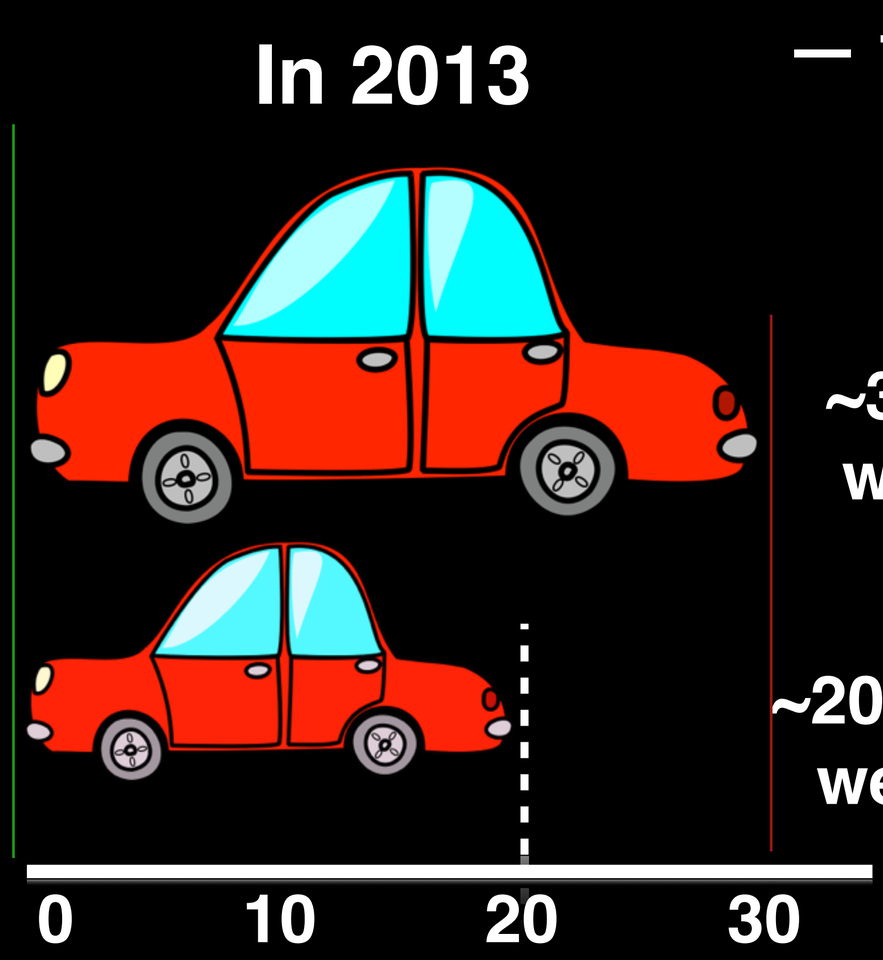


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.
2.
APRIL 1855 TO MARCH 1856.
1.
APRIL 1854 TO MARCH 1855.
JUNE
MAY
JULY
AUGUST
APRIL 1855
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
1856
FEBRUARY
MARCH
BULGARIA
CRIMEA
APRIL 1854
MARCH 1855.
JANUARY 1855
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Area principle

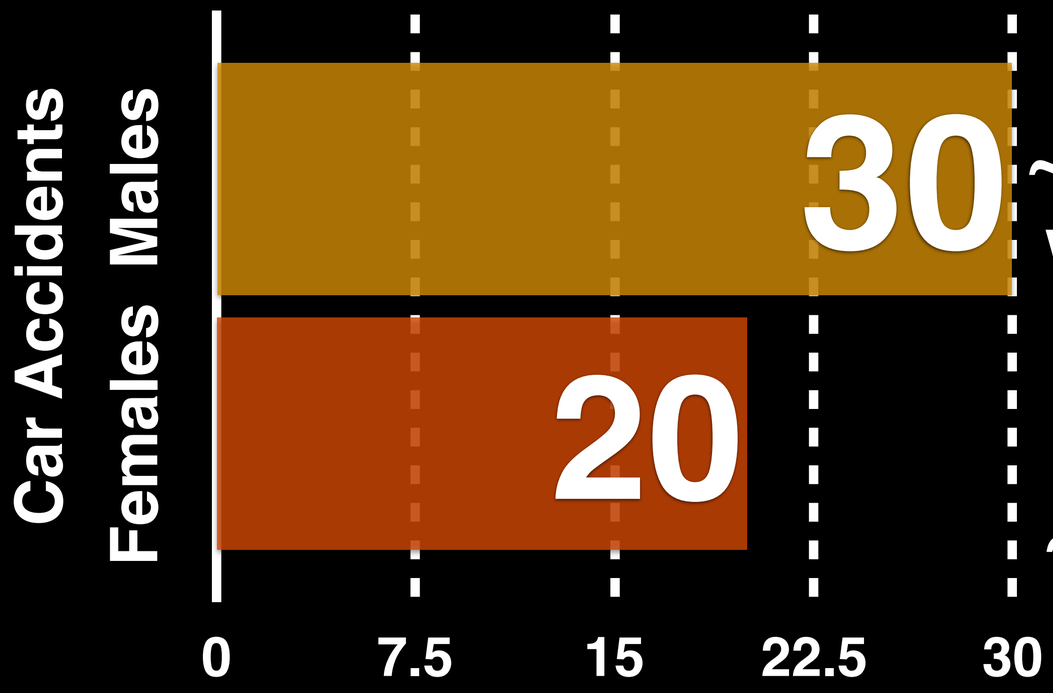
The area principle
— the size of the area correlates with the data summaries.



~30% of accidental deaths of males were due to automobile accidents.

~20% of accidental deaths of females were due to automobile accidents.

Area principle

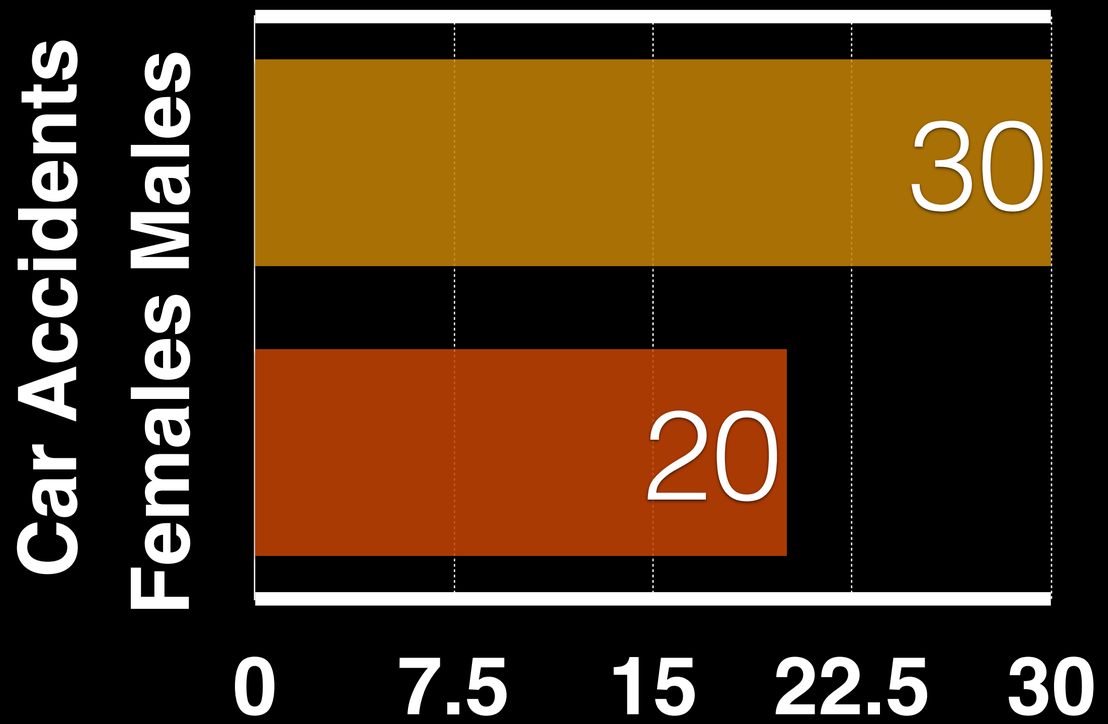


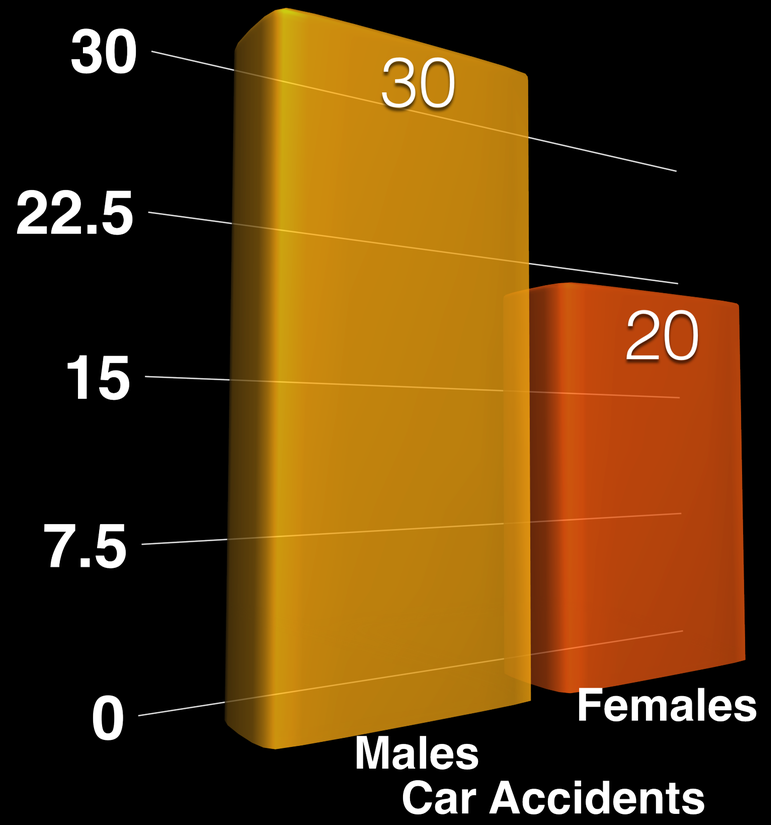
The area principle
— the size of the area correlates with the data summaries.

~30% of accidental deaths of males were due to automobile accidents.

~20% of accidental deaths of females were due to automobile accidents.

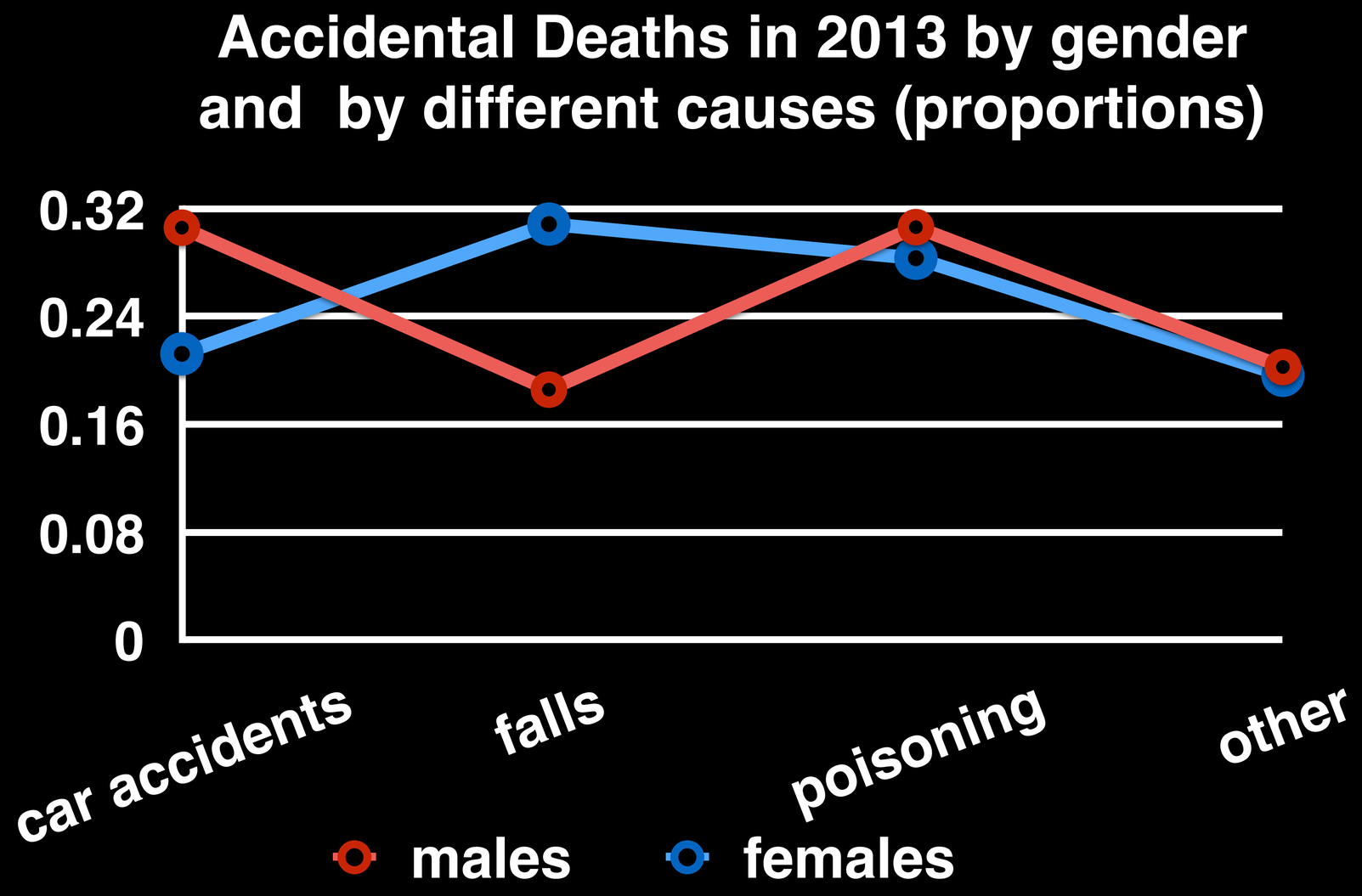
3D effects?



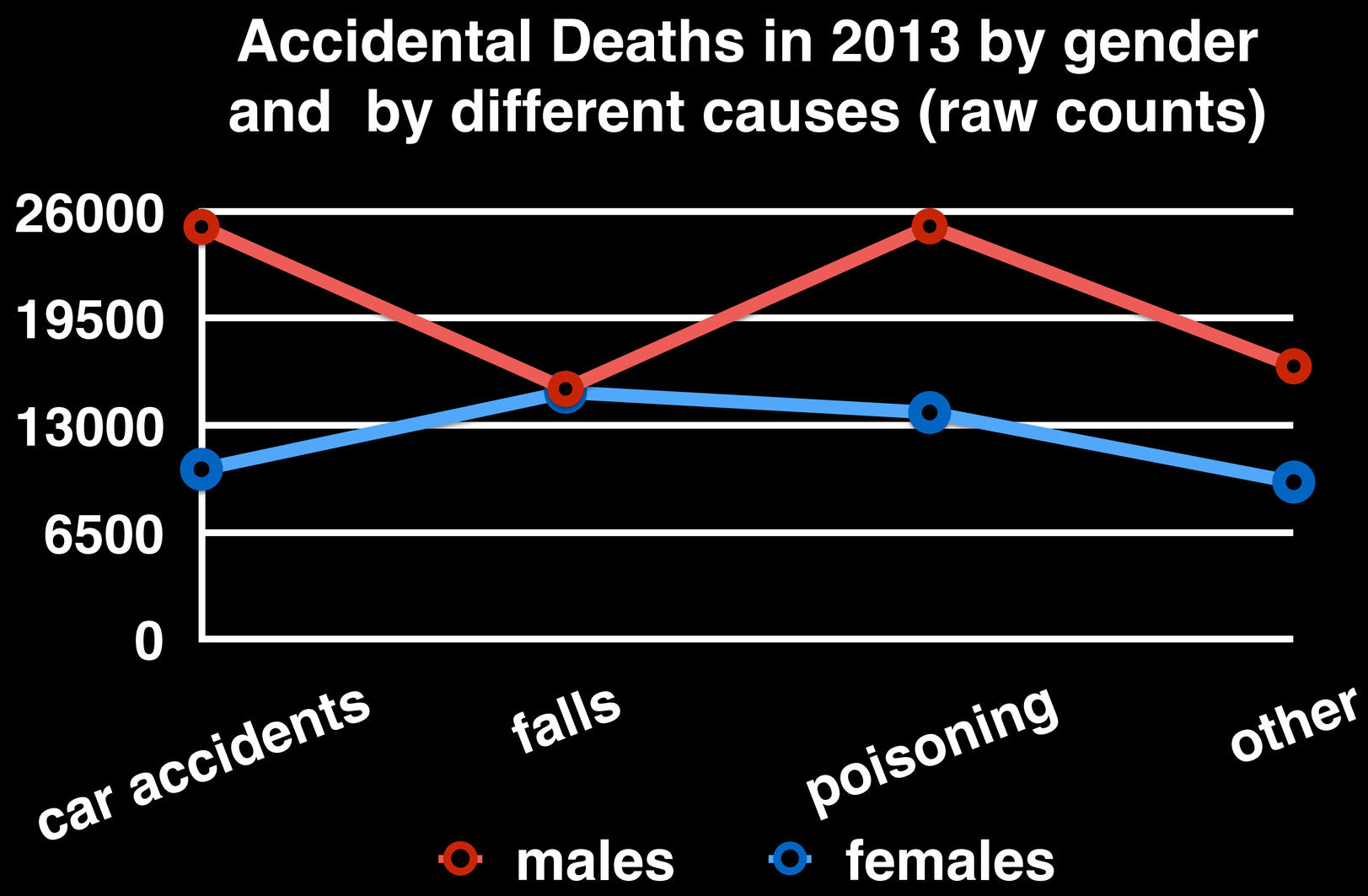


3D effects distort the visualization and often violate the area principle.

Side by side comparison



Create Meaningful visualization



Displaying quantitative variable

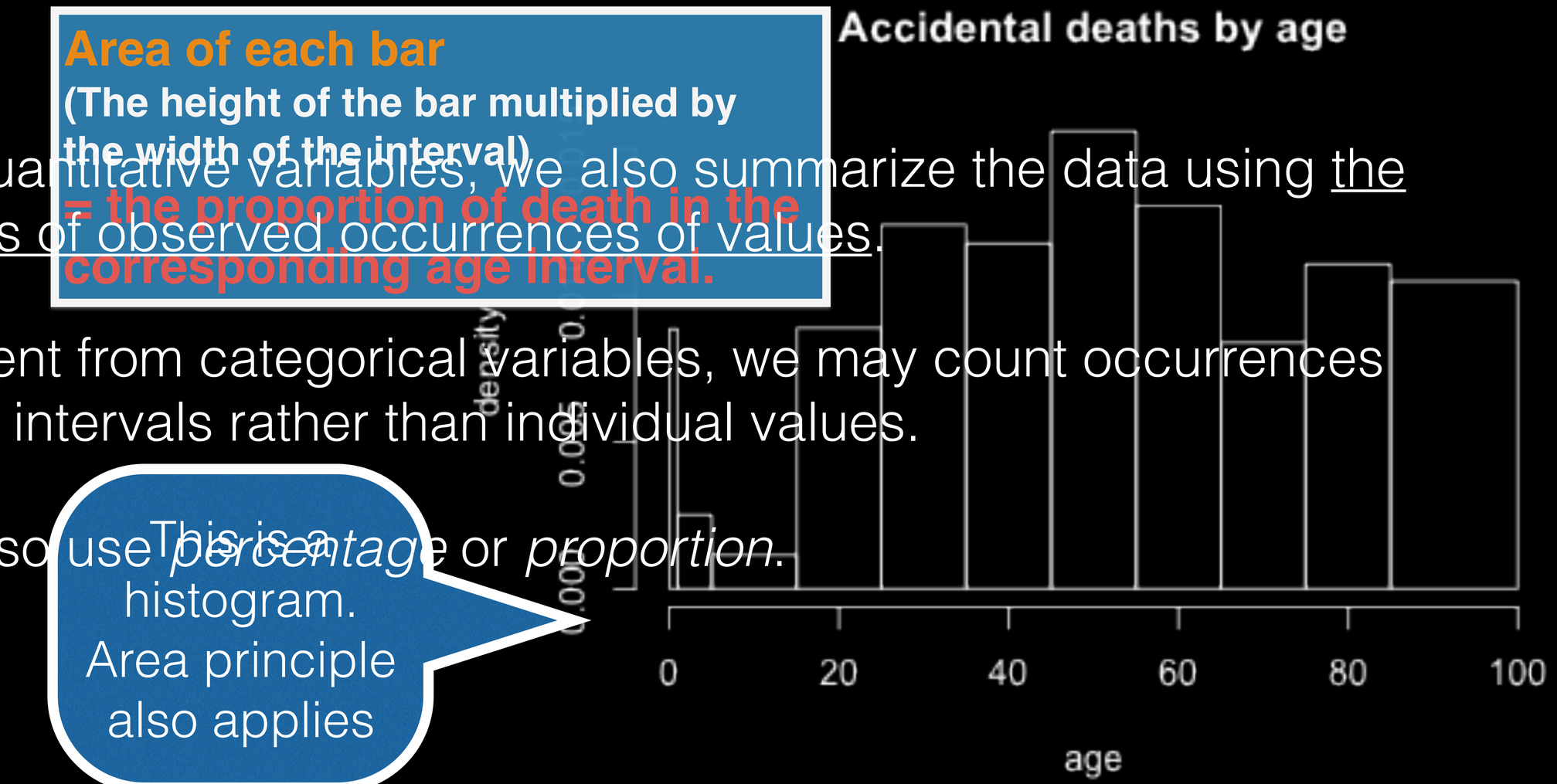
- For quantitative variables, we also summarize the data using the counts of observed occurrences of values.
- Different from categorical variables, we may count occurrences **within intervals** rather than individual values.
- We also use percentage or proportion.

Displaying quantitative variable

Area of each bar
(The height of the bar multiplied by the width of the interval)
= the proportion of death in the corresponding age interval.

- For quantitative variables, we also summarize the data using the counts of observed occurrences of values.
- Different from categorical variables, we may count occurrences within intervals rather than individual values.
- We also use *percentage* or *proportion*.

This is a histogram.
Area principle also applies



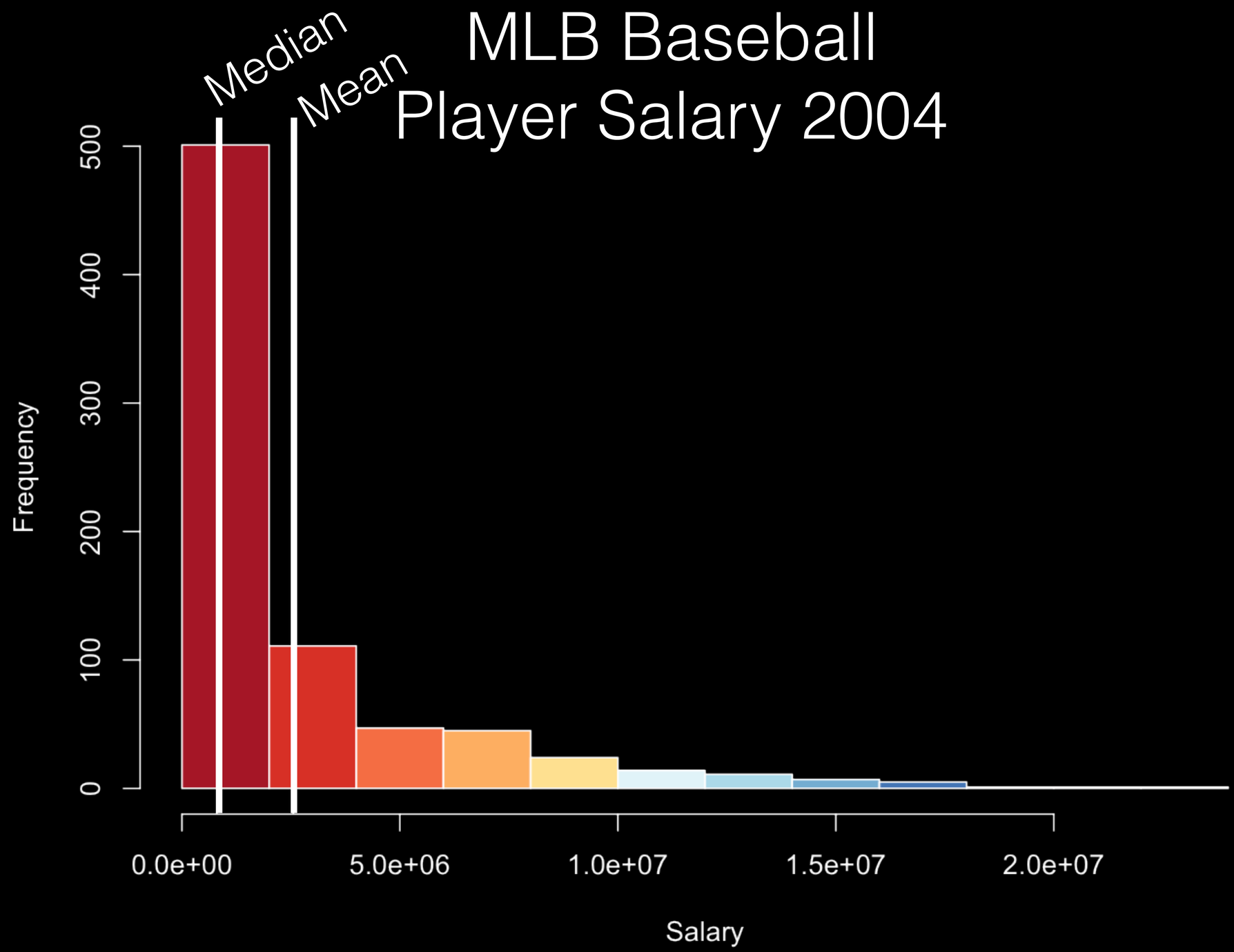
Statistics&Probability I

Summarize numerical data

Center of Variation

- Summarizing center of variation:
  - **mean** (numerical average)
  - **median** (mid-point)
- When the data come with a few extremely large values
  - mean is more affected by them than median.
  - Sensitive to outliers.

MLB Baseball Player Salary 2004



Summarizing variation
Standard Deviation

- For multiple observed values, **variation** is quantified by their **deviation from their center**.
- **Standard deviation**
  - deviation from the **sample mean**
  - the square root of variance—the average squared deviation.
- Standard deviation is a parameter for normal distributions.
- It is used as a "**yard stick**" for variation.
- It ***standardizes*** variation to make random values from different variables comparable.

variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

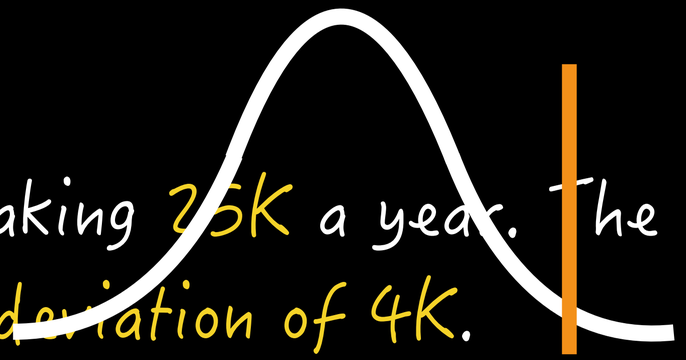
(or)

standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Standardization using mean and standard deviation

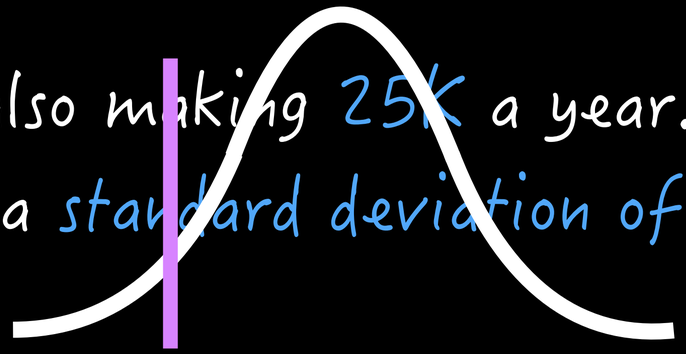
- X: a value observed
- We calculate **how many "standard deviations"** X is above/below from the mean

Standard deviation as a yard stick



- Luis is making 25K a year. The income in his city has a mean 20K and standard deviation of 4K.

Luis is 1.25 standard deviation ABOVE the mean in his city



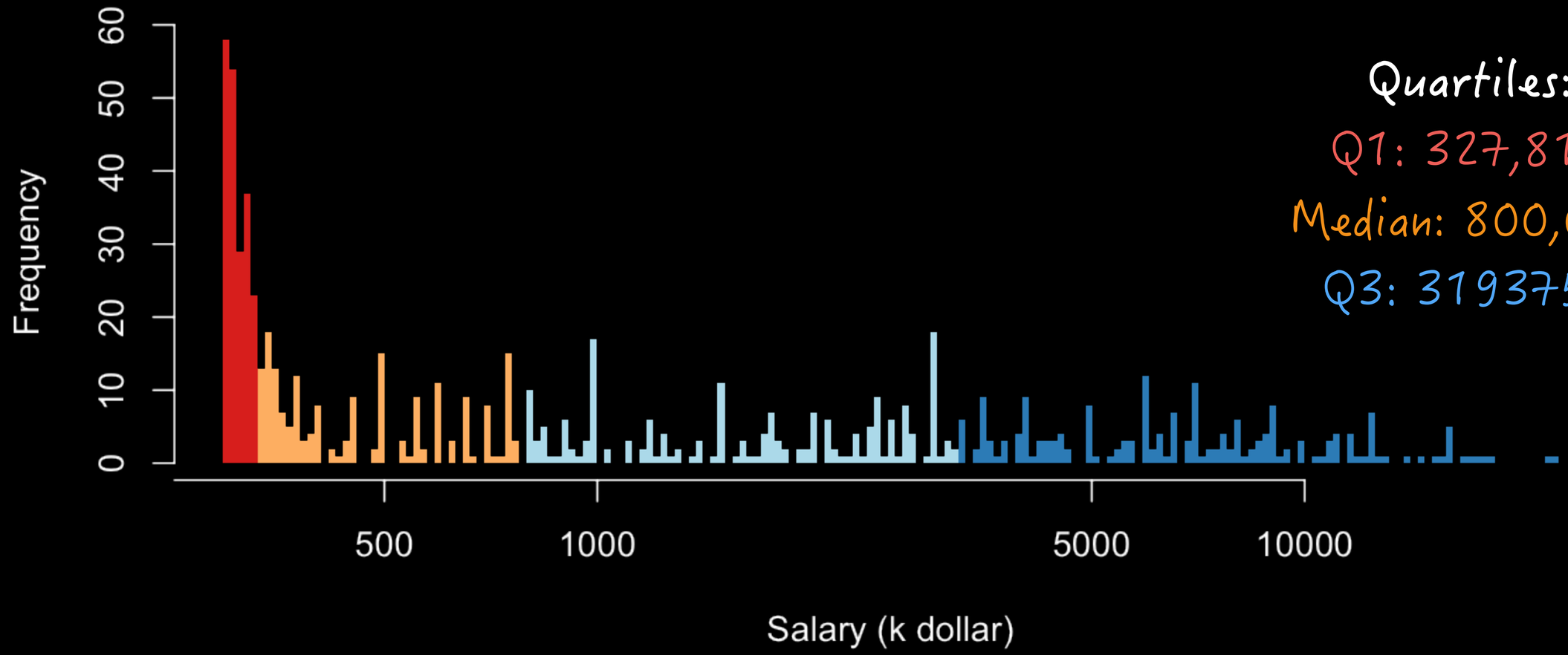
- Miles is also making 25K a year. The income in his city has a mean 30K and a standard deviation of 5K.

Miles is 1 standard deviation BELOW the mean in his city

Summarizing variation
Quantiles

- Quantiles (or percentile): a value threshold of a variable that is defined to have a percent of data below it.
  - SAT critical reading, a score 600 is the 79th percentile.
- A set of special percentiles are called **quartiles**, which corresponds to 25%, 50% and 75% percentiles.
  - Quartiles divide data into quarters

MLB Baseball
Player Salary 2004



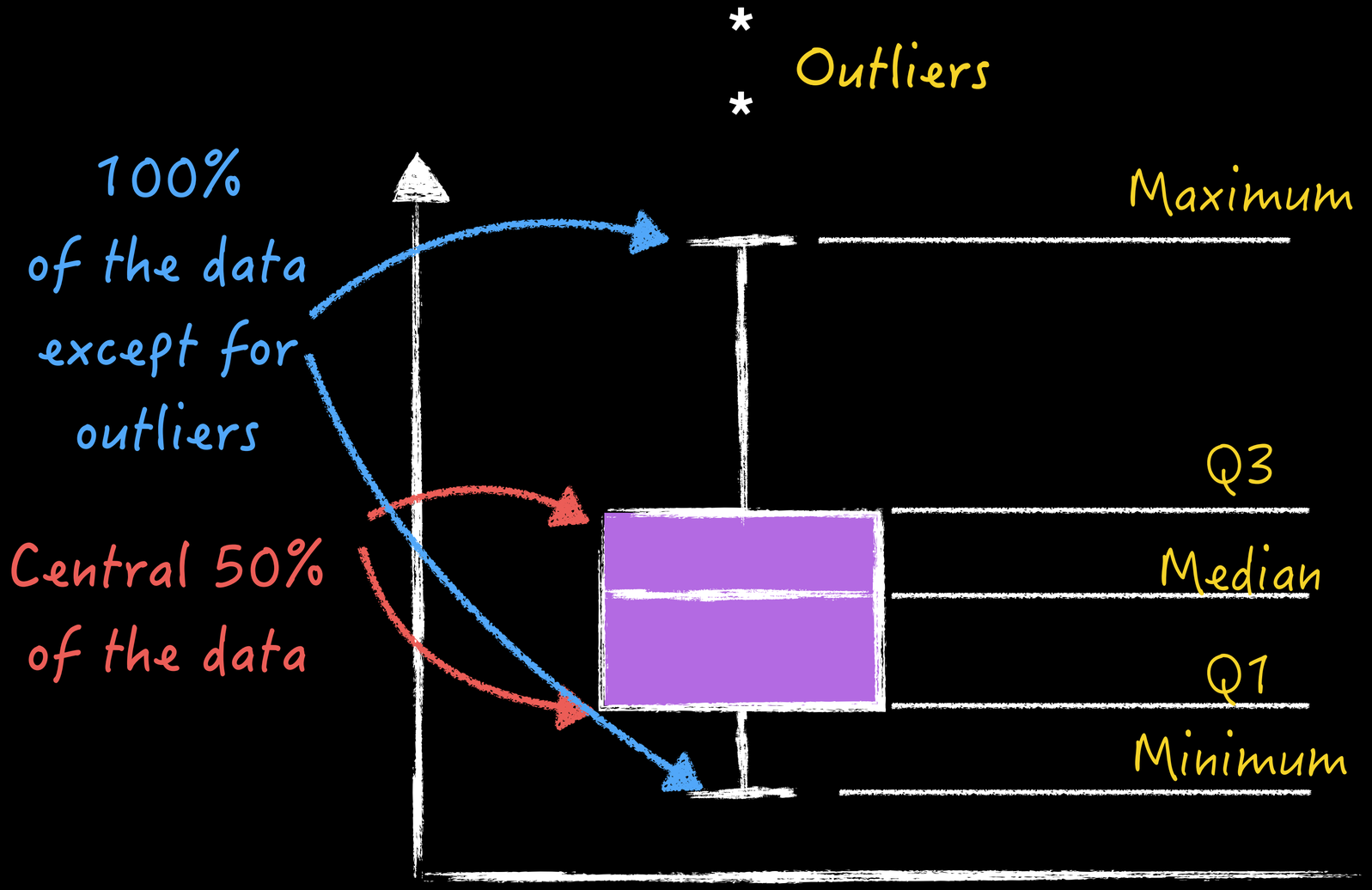
Quartiles:

Q1: 327,81

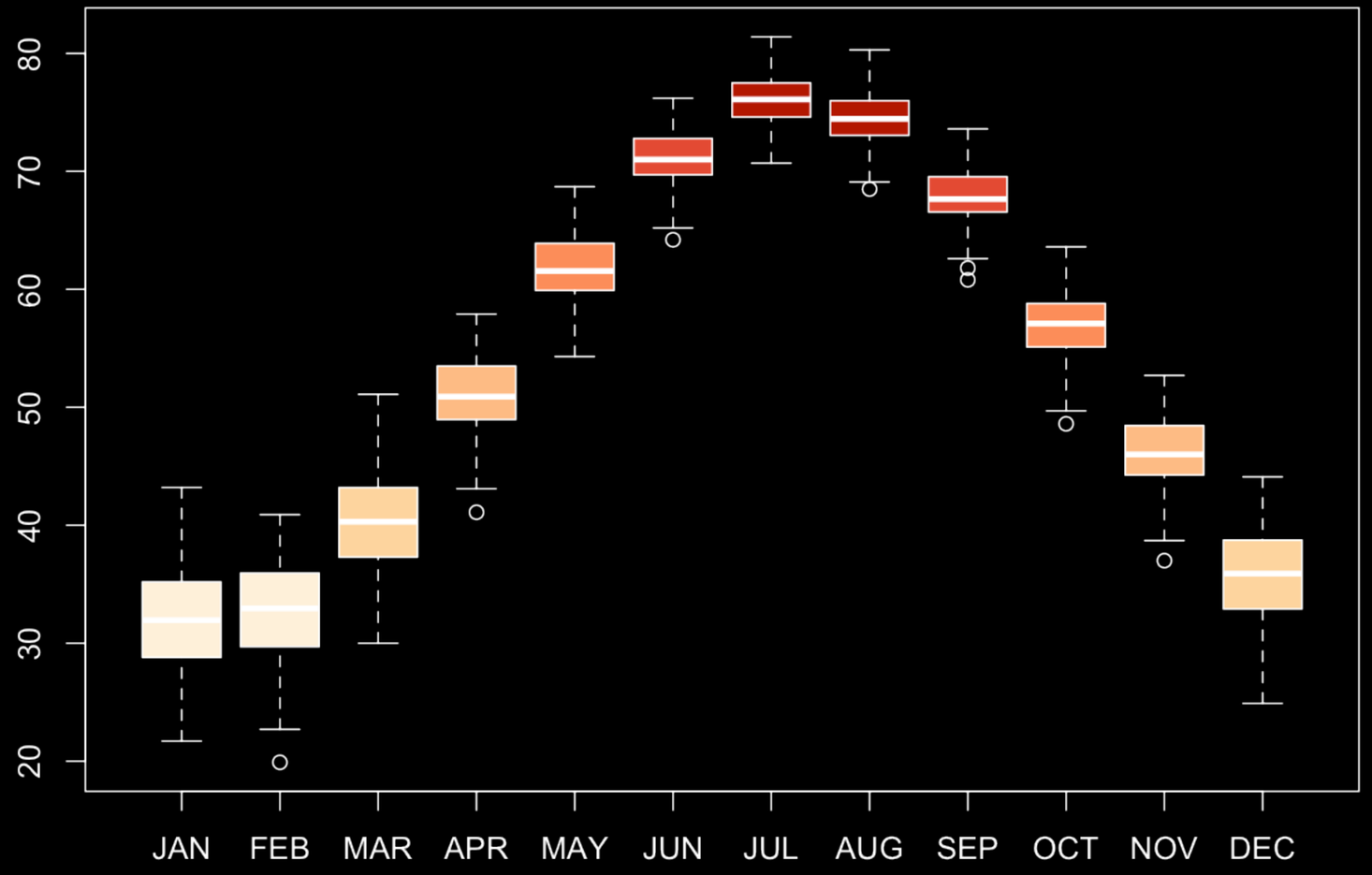
Median: 800,

Q3: 319375

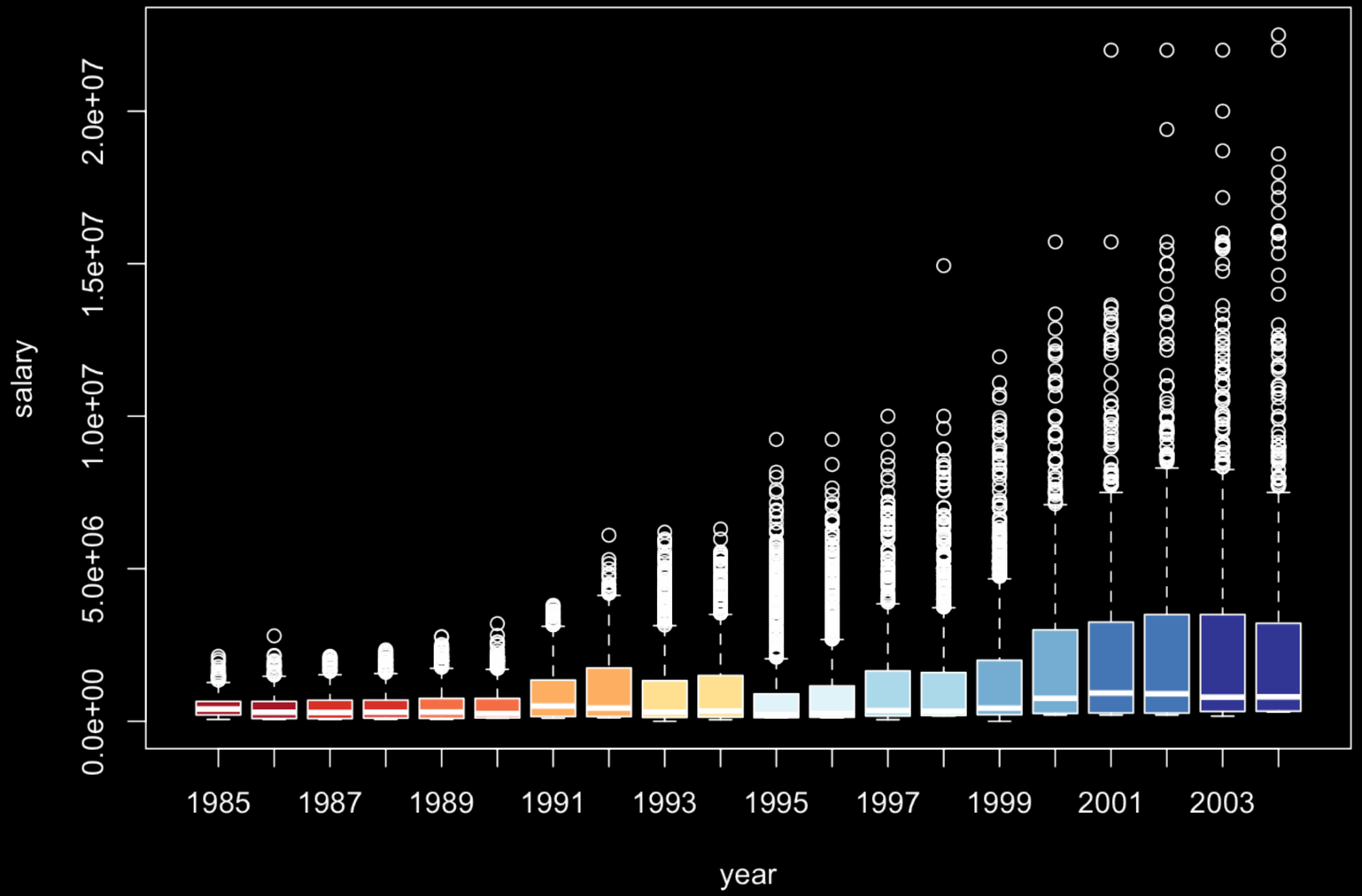
Box plot


Outliers
100%
of the data
except for
outliers
Central 50%
of the data
Maximum
Q3
Median
Q1
Minimum

NYC Monthly Average Temperature 1869-2012



MLB Salary
1985-2004



Statistics&Probability I

Association between variables

Association

- Certain values of one variable are observed more frequently with certain values of another variable.

Display and summarize association

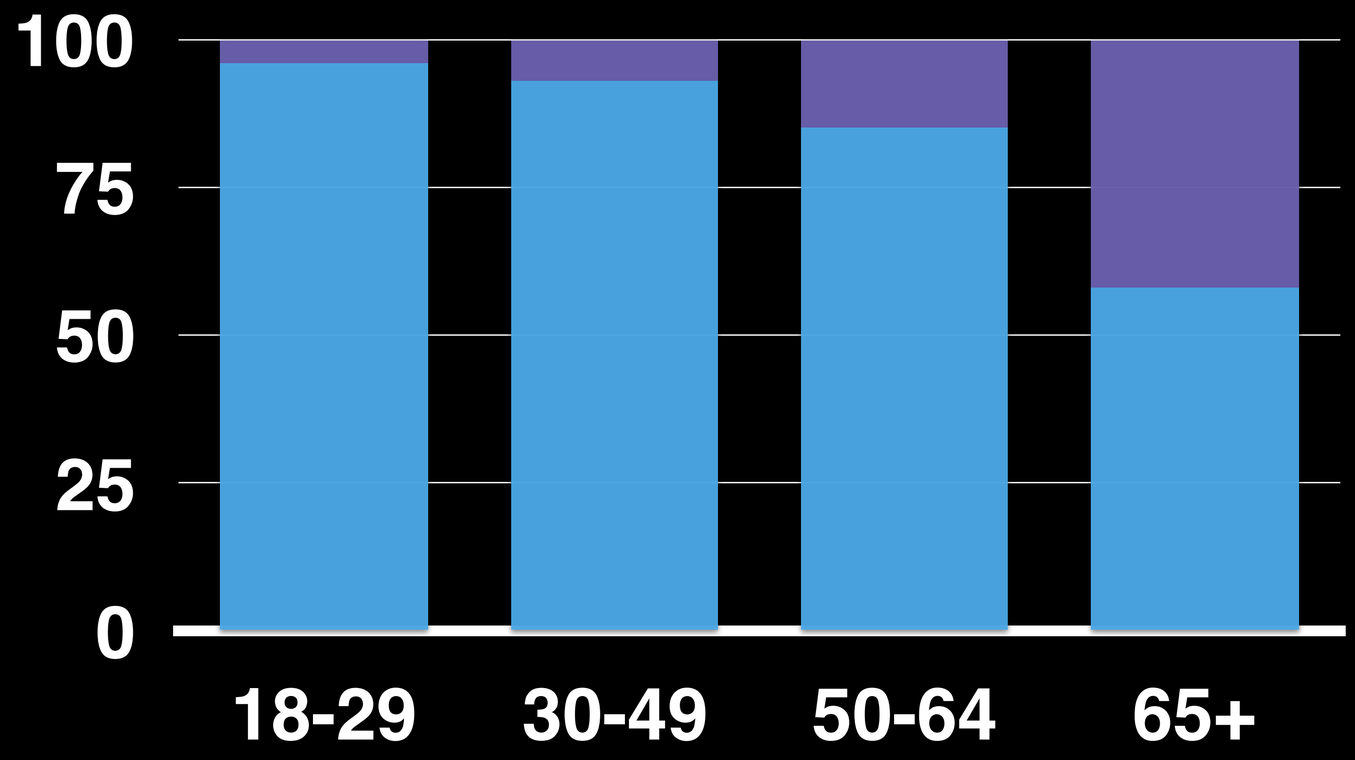
Association: categorical variables

- The association of two categorical variables is summarized using counts of joint occurrences.

	Use Internet	Do not use
18-29	48	2
30-49	93	7
50-64	85	5
65+	29	21

Hypothetical example based on findings from
Pew Research Center on Internet Use (August 2012)

Internet Use versus Age



This is a stacked bar chart.

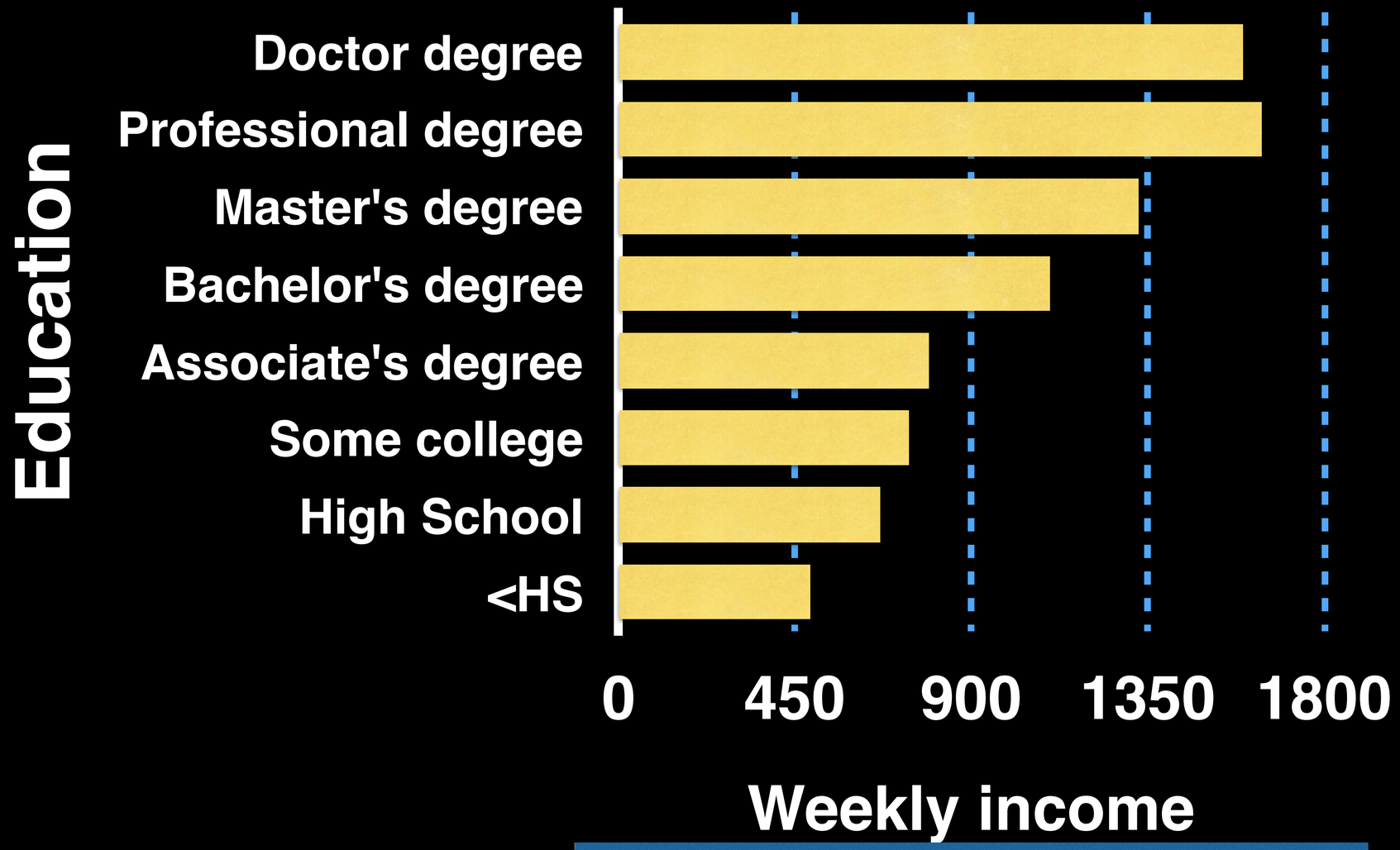
Measuring association

- Young adults: 96% use internet
- Senior adults: 58% use internet
- Difference in proportions
- Relative risks
- Odds ratio

Association: categorical variable versus quantitative variable

- Can be summarized by
  - *distribution* of the quantitative variable (Y)
  - given (or *conditioning*) on each value of the categorical variable (X).
- If the two variables are associated, the distribution of Y will be *dependent* on the value of X.

Education versus Income



Data source: Bureau of Labor Statistics (2014)

Measuring association

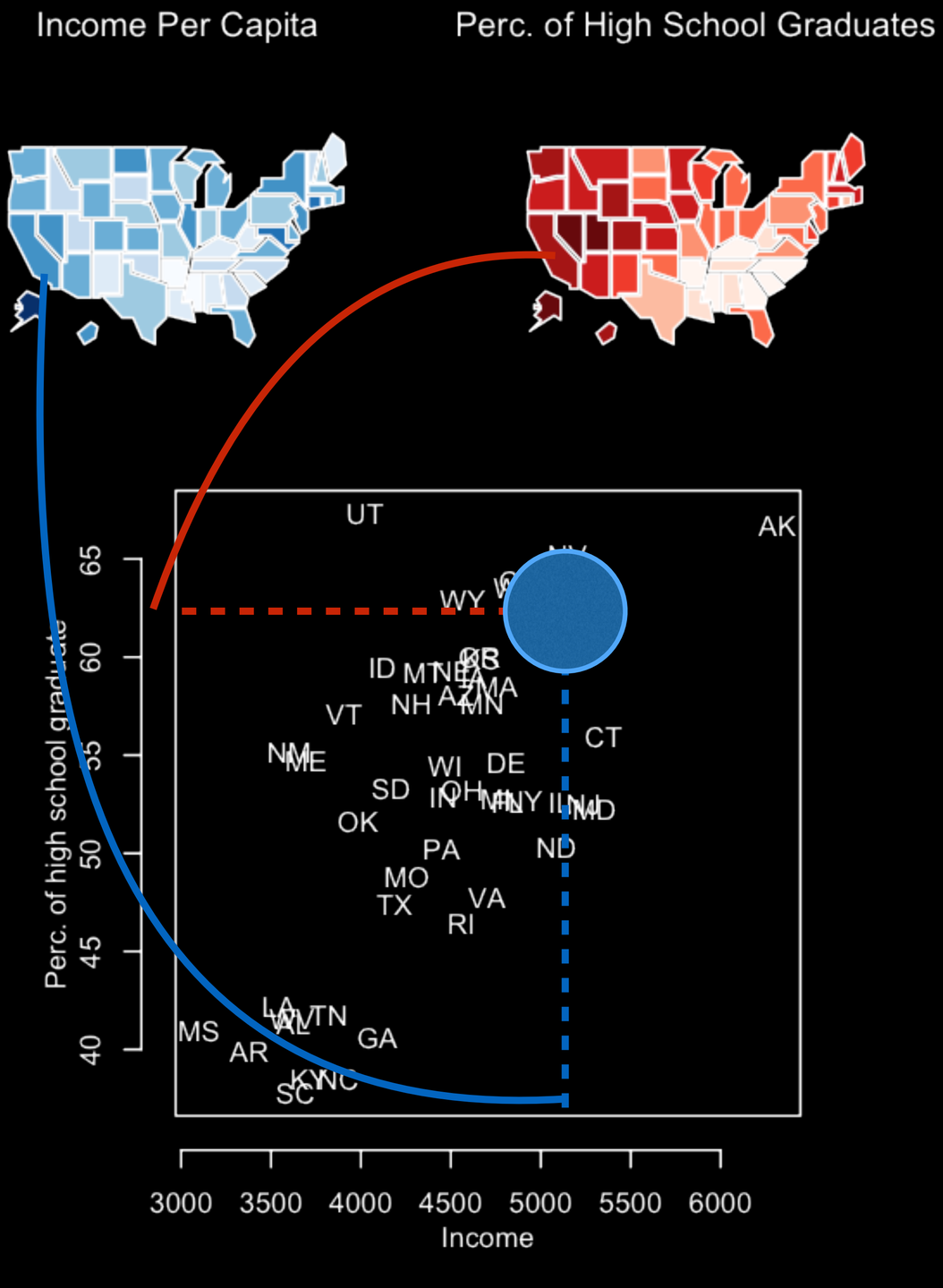
- <HS: on average $488 a week
- HS: on average $668 a week
- Difference of the averages
- Factor to consider in this comparison
  - sampling variation

Association: quantitative variables

- Patterns in the joint occurrence of values from the two variables.

SCATTERPLOT

displays the relation
of two variables



Measuring association

- Correlation:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

Correlation

- **No distinction between X and Y**
- **Between –1 and 1**
- **Value 0 indicates weak LINEAR association**
- **Correlation measures direction & strength for LINEAR relations**
- **Correlation may be more accurate, but it does not tell the whole story.**

Confounding and Simpson’s paradox

Scenarios of Association

Association = Causation?

- The answer is NO!
- Association is symmetric between X and Y, causation is not.
- Different scenarios of association

Now let’s consider a hypothetical example

The cell phone example

- In this example, we consider the association between
  - Y: price of a cell phone
  - X: brand of a cell phone
- We compare average prices (Y) giving X.

- We then looked at a third variable.
- Z: technical specification of the cell phone.
- We found if we consider the values the Z, the association pattern is reversed.

Establish cause-effect relation

- Randomized experiment
  - A/B testing is a randomized experiment with one treatment group and one control group.
  - Why control group? — Placebo effect
  - Double-blinded experiment
  - Control, Randomization and Repetition
- Causal inference for observational data

Estimation and Inference

Sampling

Population

- Population is the entire collection of individuals of interest for a study
- Why not study the whole population then?

A representative sample?

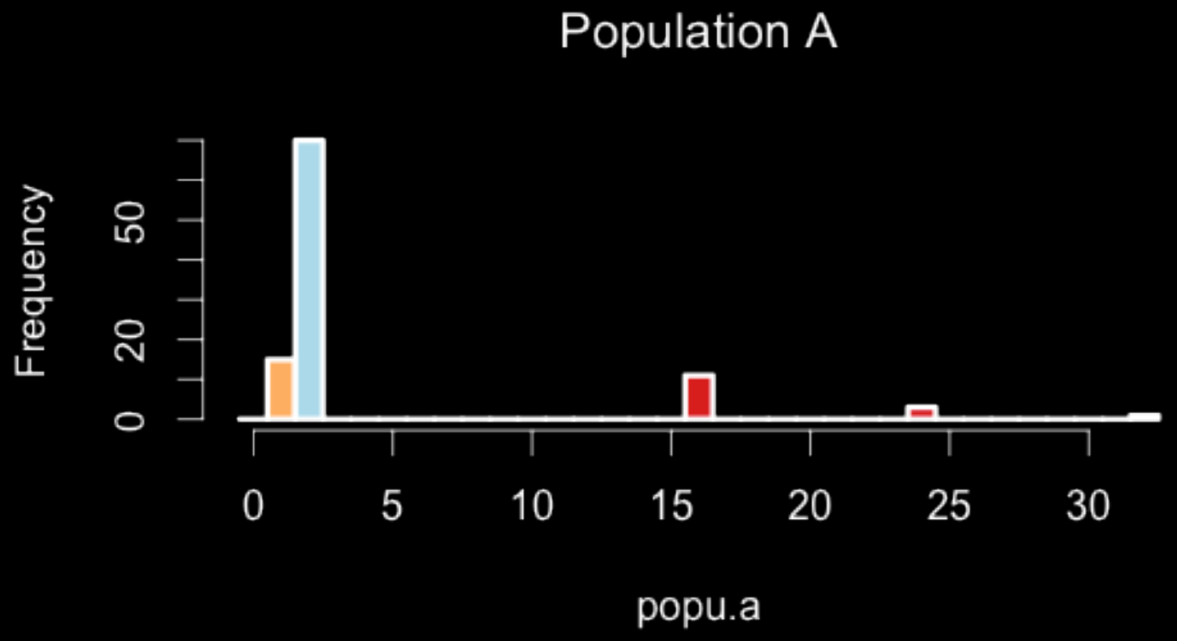
- **What happens if we don’t have a good (representative) sample?**
  - **Misleading outcomes**
  - **Biased results**
  - **Difficult to analyze**
  - **Waste of time and $ $ $**

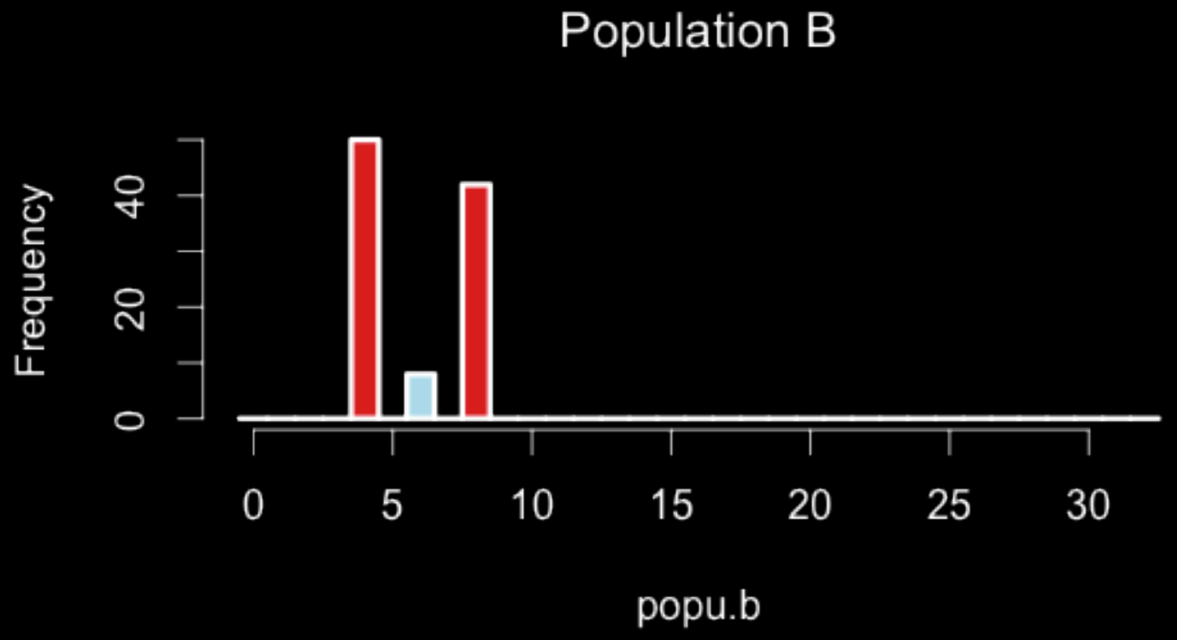
What makes a representative sample?

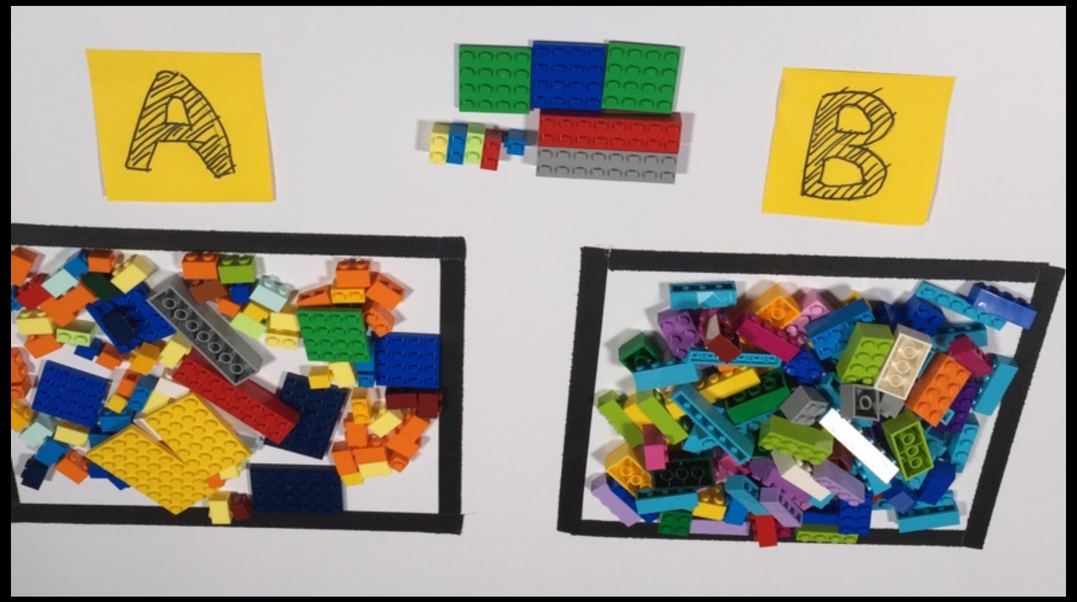
- **A sample that carries correct images of the population**
- **Biased design: sampling method with systematic errors;**
  - **The sampling chance of an individual is associated with the outcome of interest.**
- **Why not matching?**
  - **Matching is to have you sample composition exactly proportional to that of the populatio**

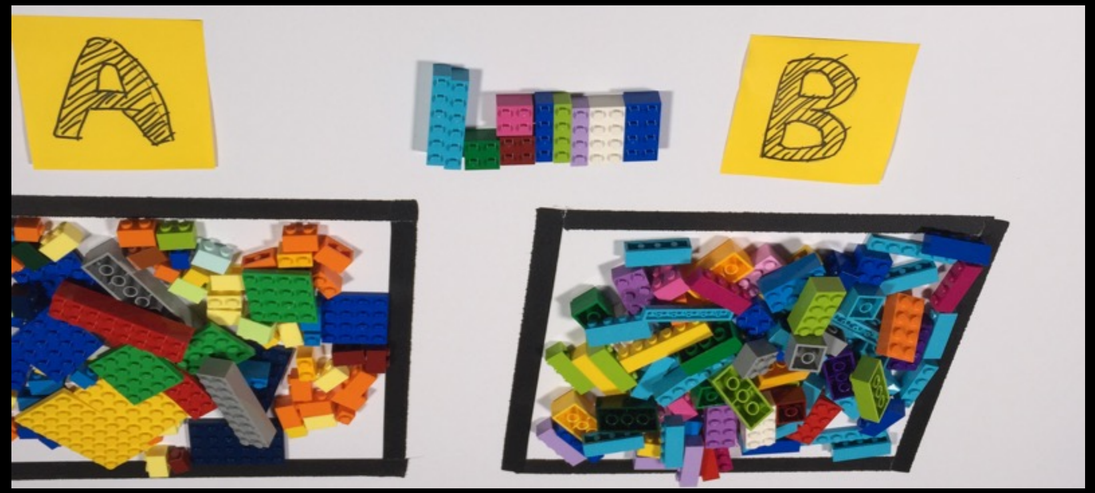
Lego demo

- From a population of 200 lego pieces, we create two populations.
- Each population has 100 pieces.
- X, Variable of interest is the number of points on a sampled lego piece.
- Population A: a mixed population of very small pieces and very large pieces
- Population B: a relatively more homogeneous population
- Population A: the average for X is 4.35
- Population B: the average for X is 5.84
- For illustration, we randomly sample pieces from these two populations by hand.

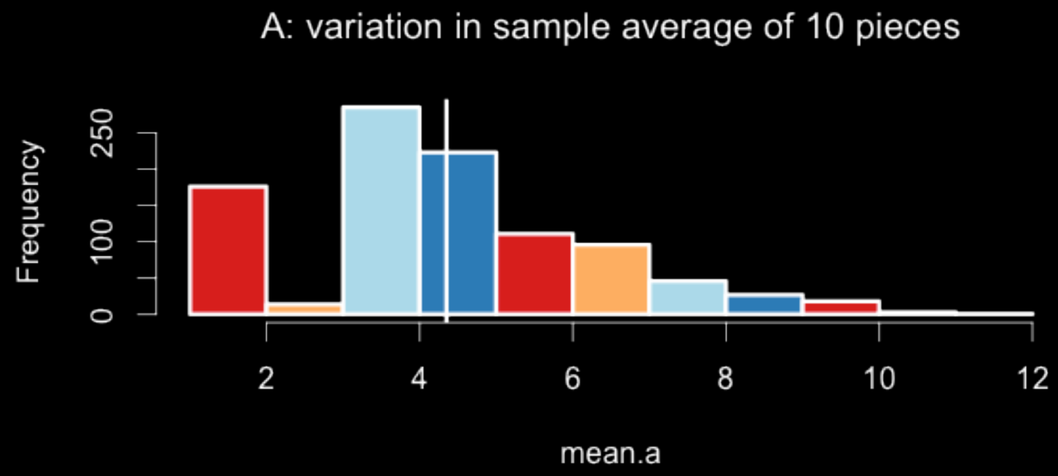

Population A
Frequency
0
20
50
0
5
10
15
20
25
30
popu.a

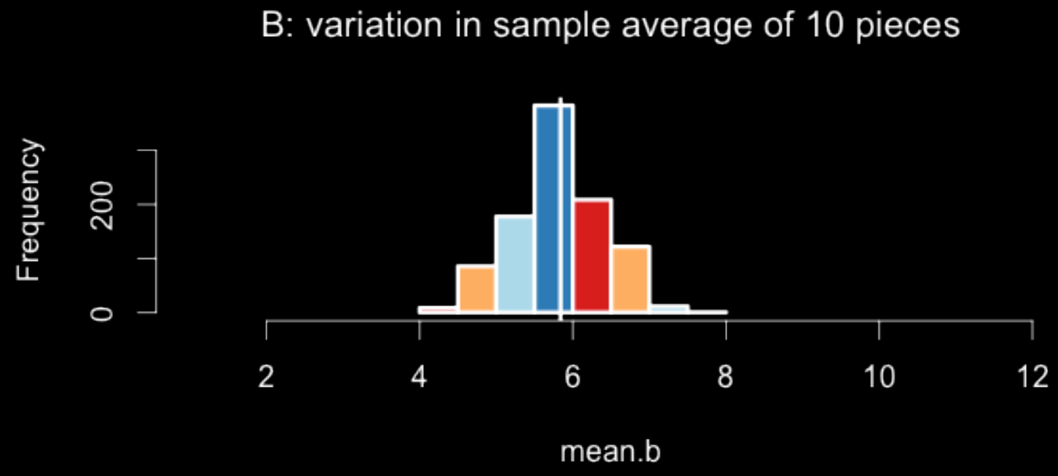

Population B
Frequency
0
20
40
0
5
10
15
20
25
30
popu.b


A
B

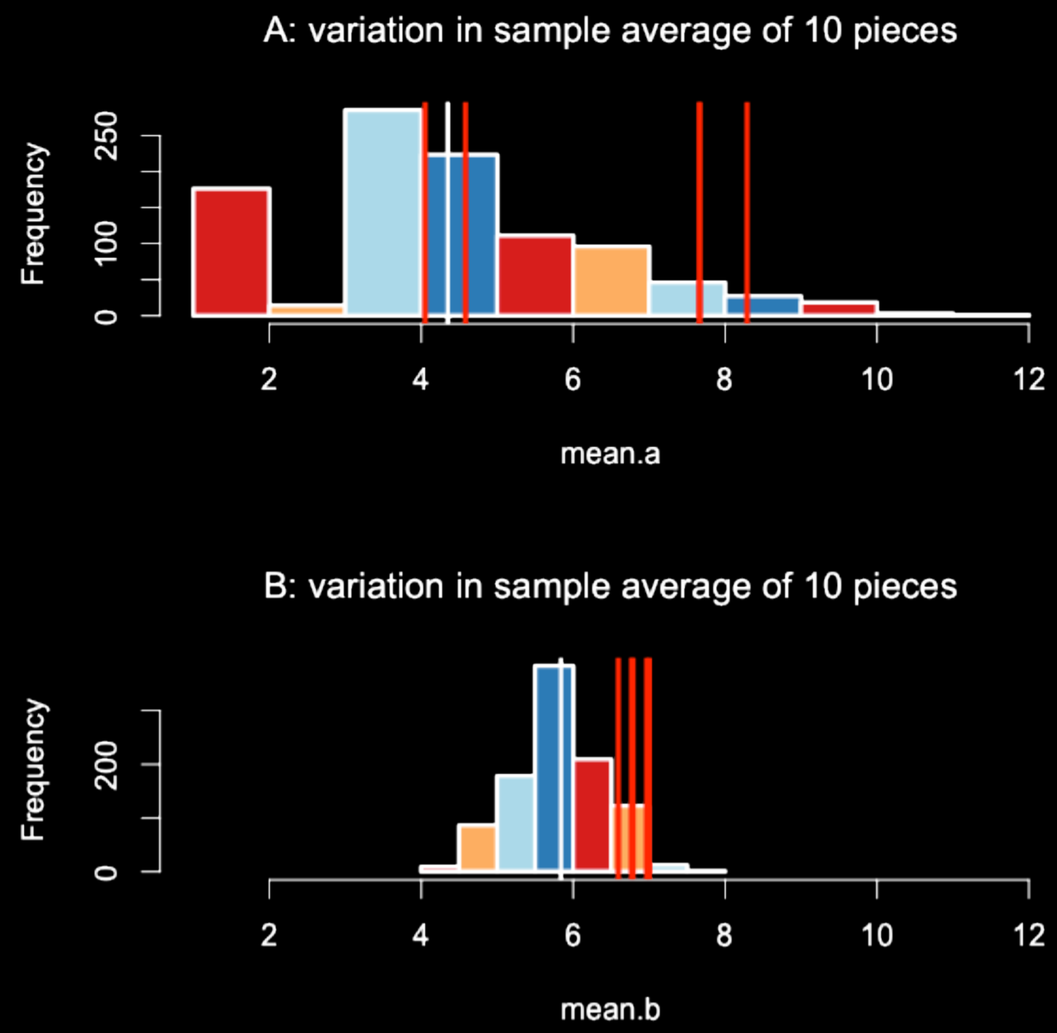

A
B

Sampling distribution of sample means (computer simulation)





Sample means from real samples



Bias in Sample estimates

- Estimates based on a sample differ from the truth.
- *Sampling bias* or *Selection bias* in sampled observations on a variable
  - *o*ccurs when the probability that *an individual is selected* is associated with the variable's value.
  - Example: education level using an internet survey
  - Example: political opinions on social media.
- We need to model the selection process for understanding sampling bias, in addition to sampling variability.

From population to Sample (animation)

Intro to Probability

Probability and randomness

Probability studies randomness

- Unpredictable
- Trends

Randomly pick one:

1 2 3 4

Did you pick 3?

- About 75% of all people pick the number 3.
- 20% pick either 2 or 4
- Only 5% pick 1

How to describe randomness

- **All possible outcomes**
- **Chance (or probability) of observing each of the outcomes.**
- **Observing an event is observing one of a set of specific outcomes.**

Basic Probability Rules

- Notation for event: A, B, C, etc.
- For any event, Prob(A) or P(A) is a value between 0 and 1, including 0 and 1.
- The probabilities of all possible outcomes add up to 1.
- $P(A^C)=1-P(A)$;
  - Here A^c is the "Not A" event.
- If A and B don't overlap,
  - P(A or B) = P(A) + P(B)
- If A and B are independent,
  - P(A and B) = P(A)P(B)

Probability example

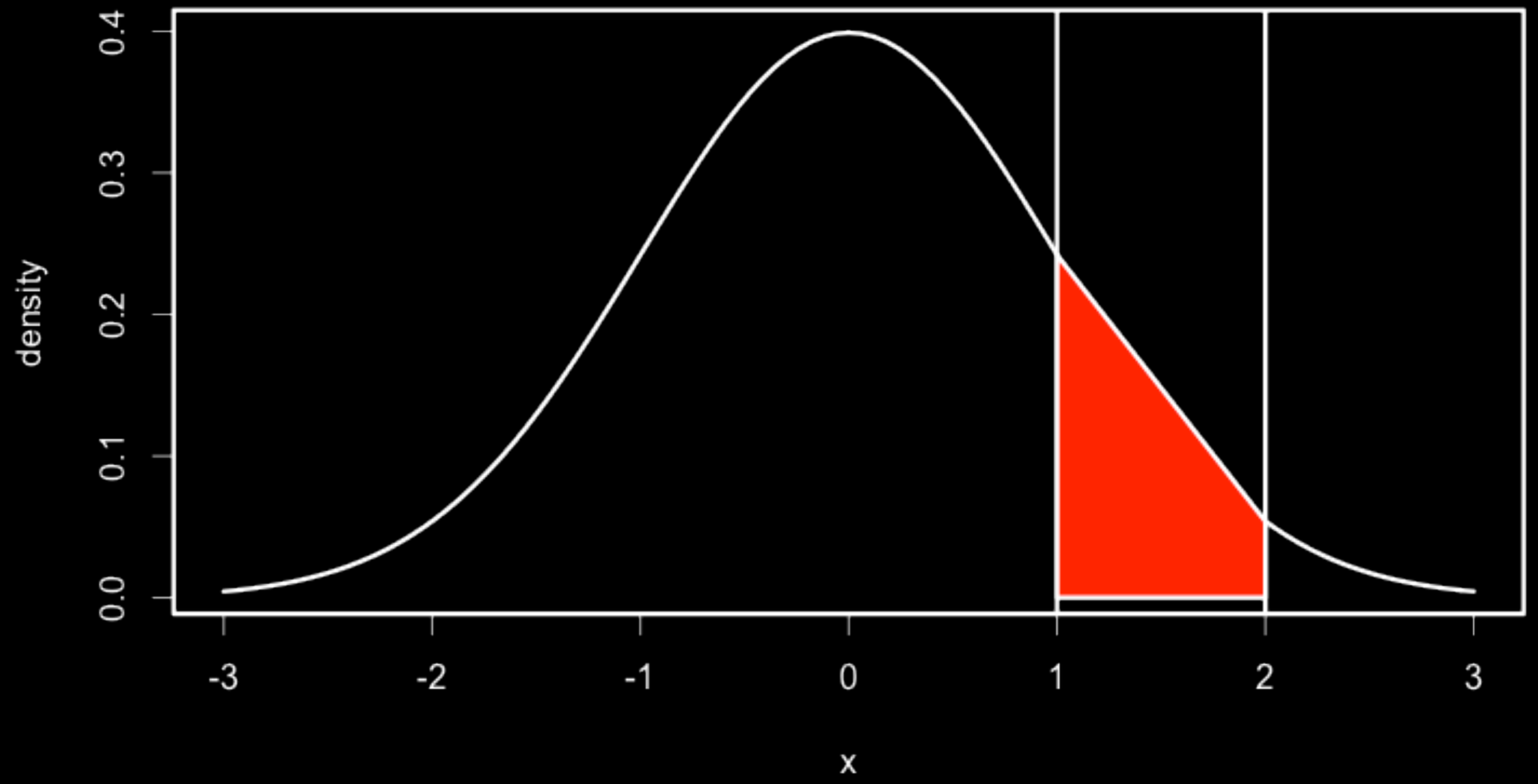
- **Flip a coin 3 times, what is the probability of getting fewer heads than tails?**

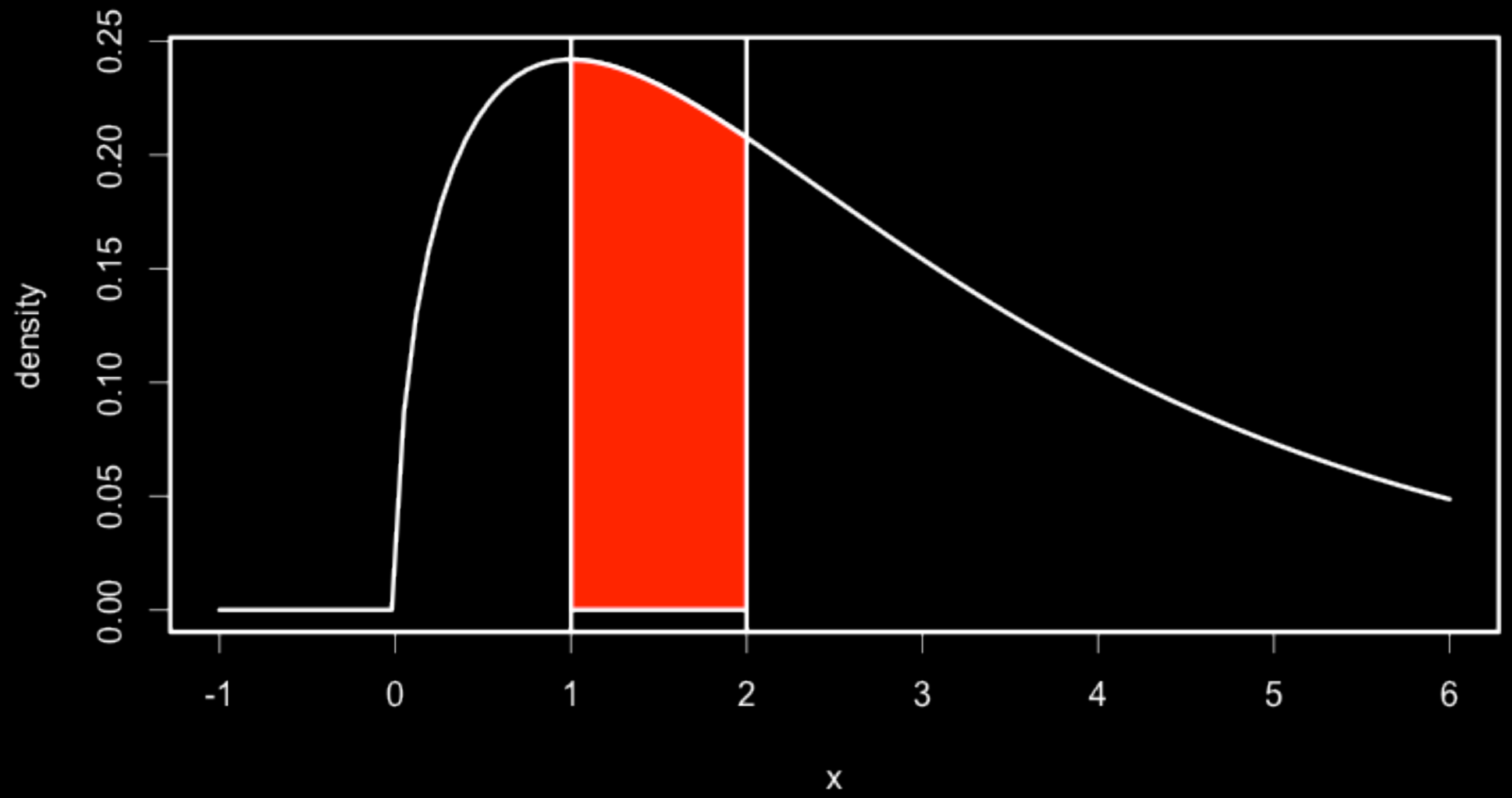
Probability example

- **Flip a coin 10 times, what is the probability of having heads on both the first and the last flip?**

- **What is the probability of having 10 heads?**

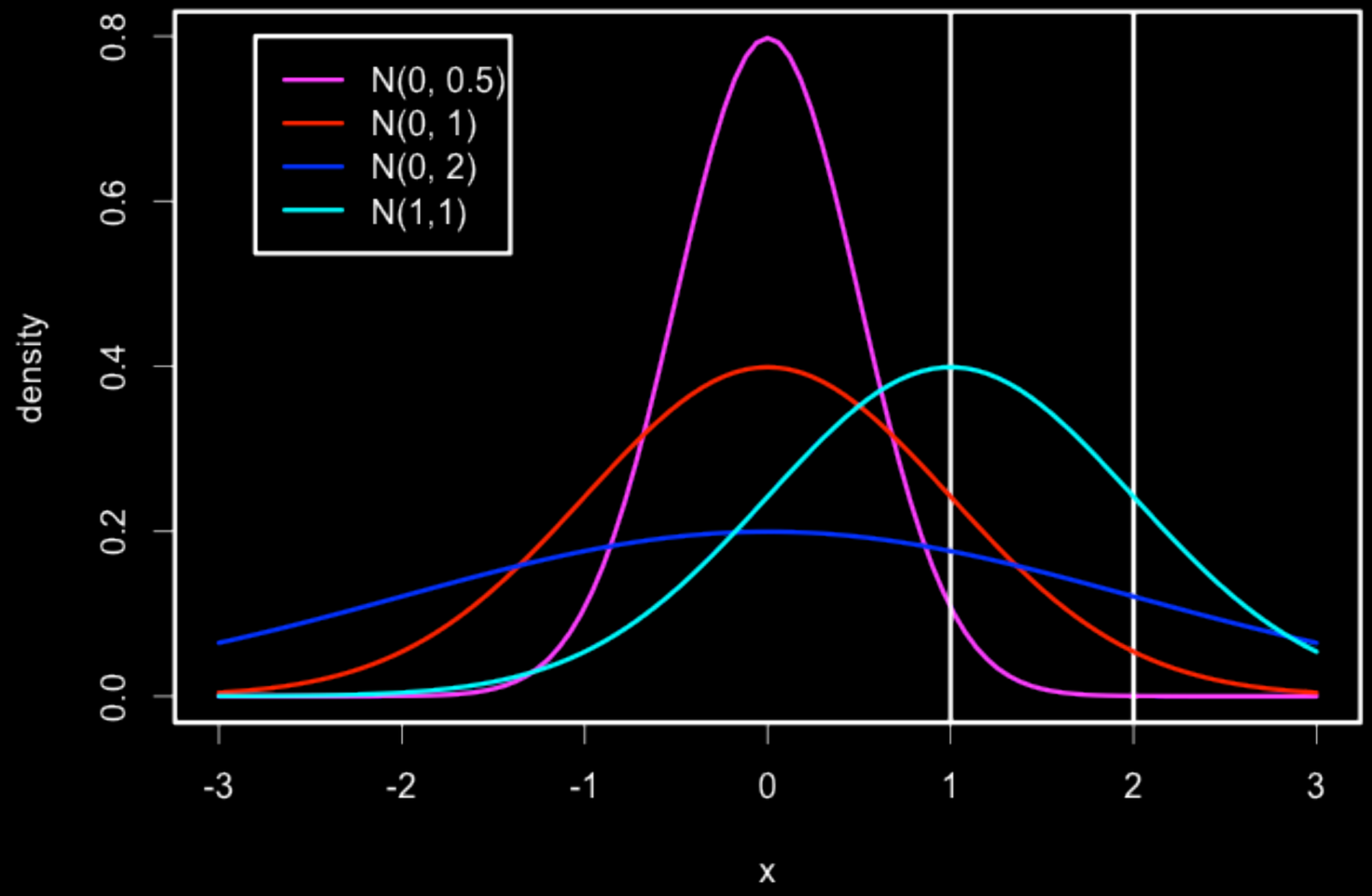
Probability distribution


density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x


density
0.00
0.05
0.10
0.15
0.20
0.25
-1
0
1
2
3
4
5
6
x

Normal distributions

N(mean, standard deviation)



Center and Spread

- Probability distributions
  - models for variation!
  - Mean (Expected value): average outcome value weighted by probability.
  - Variance: average squared departure from mean weighted by probability
    - Standard deviation: square root of variance.
  - Median: mid point of a probability distribution
  - Percentiles: cut-off threshold for the (random) outcome values corresponding to specific percent values of a probability distribution.

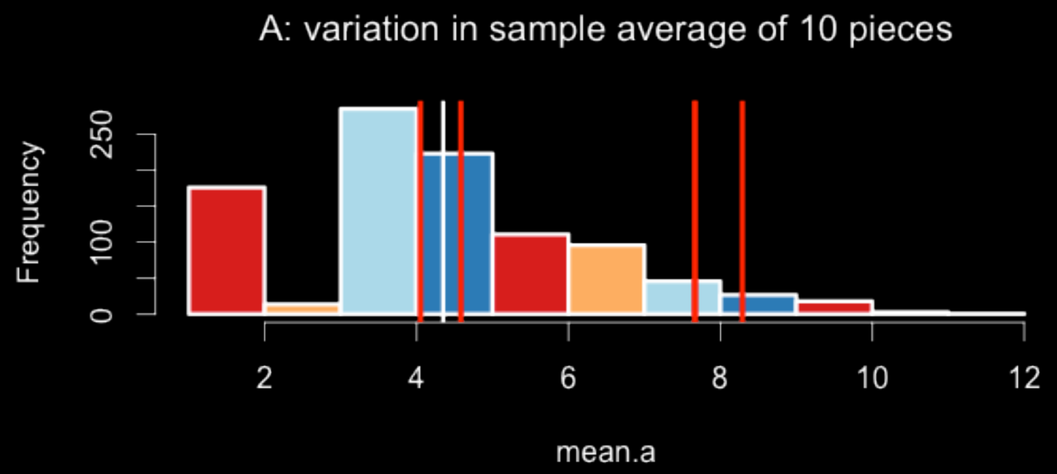
Statistical Inference

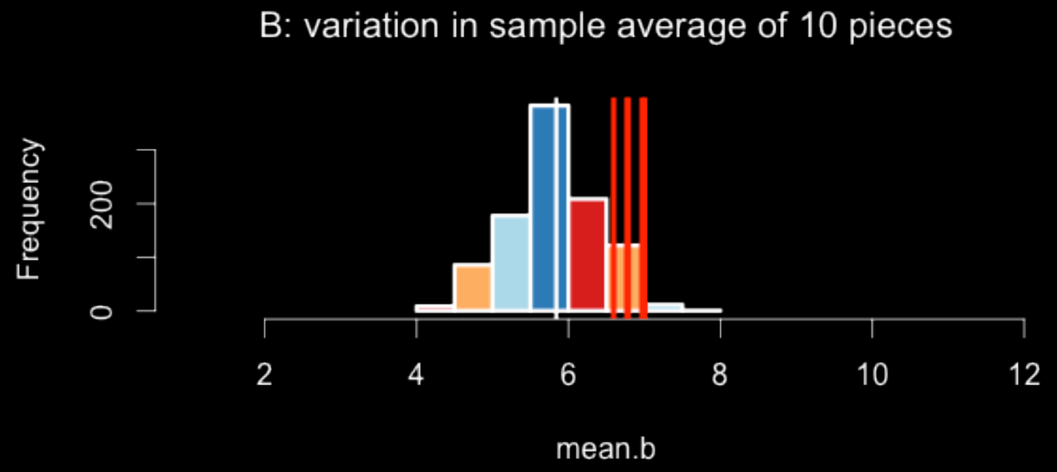
Confidence Interval

Recall the lego example

Sampling distributions

- by simulations
- by experiments
- by mathematical model

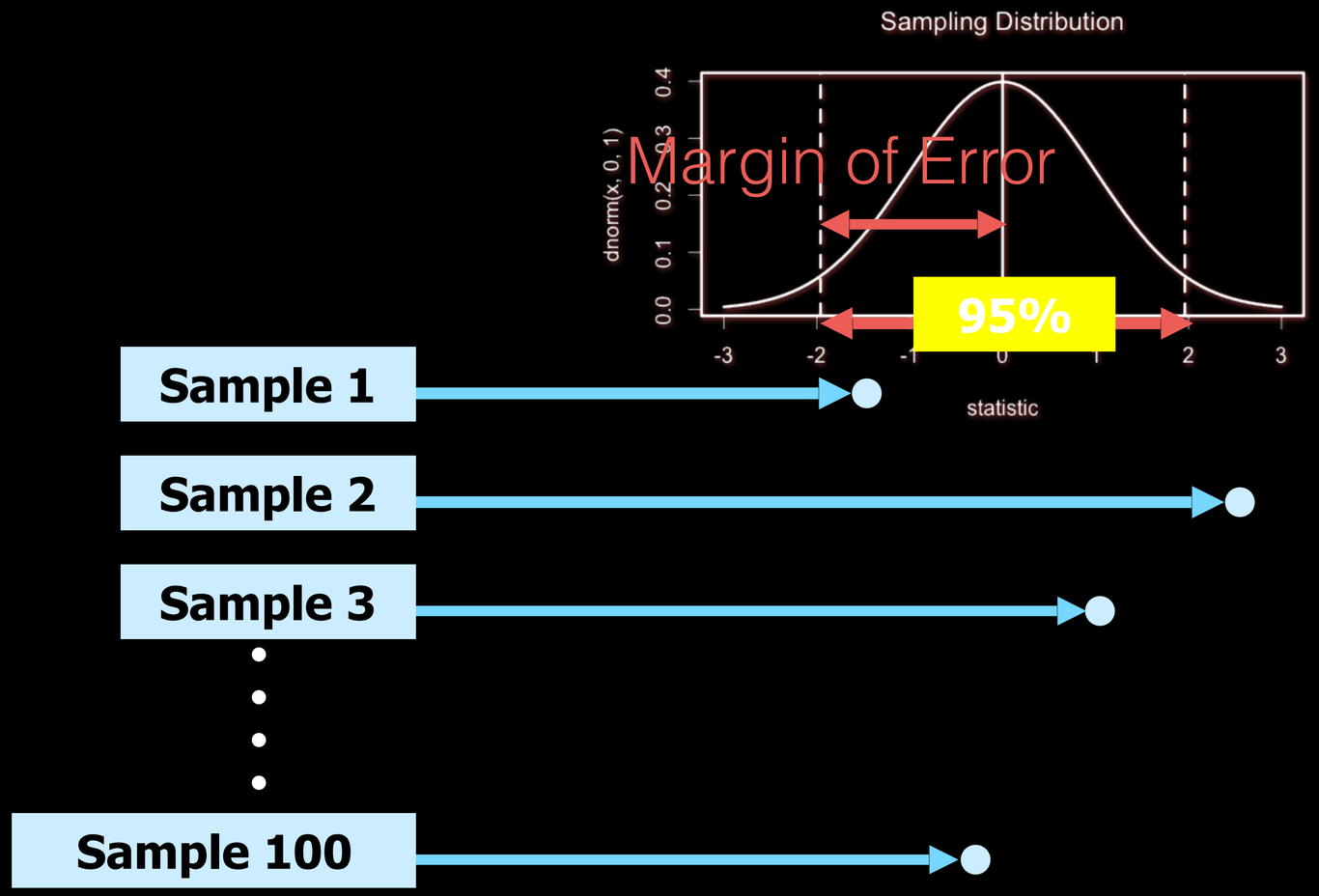




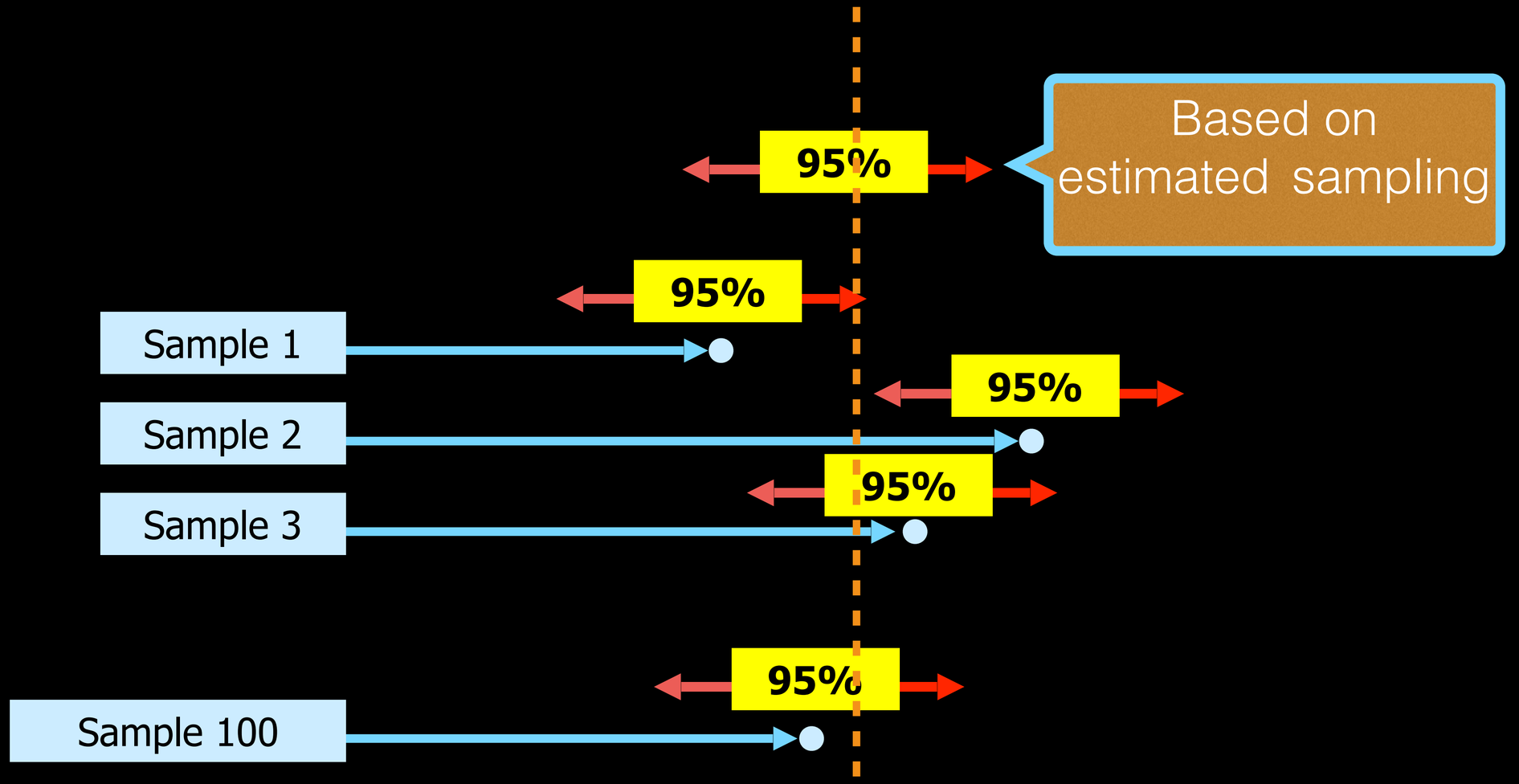
Sampling distribution model

- Probability model
- Describe the sampling distribution of a statistic (e.g., an estimate of a population quantity)
  - Center
  - Variability
- Factors affect sampling distributions
  - Population distribution for the variable of interest
  - Data generation process
  - Sample size

Sampling distribution and confidence interval



Sampling distribution and confidence interval



Confidence interval

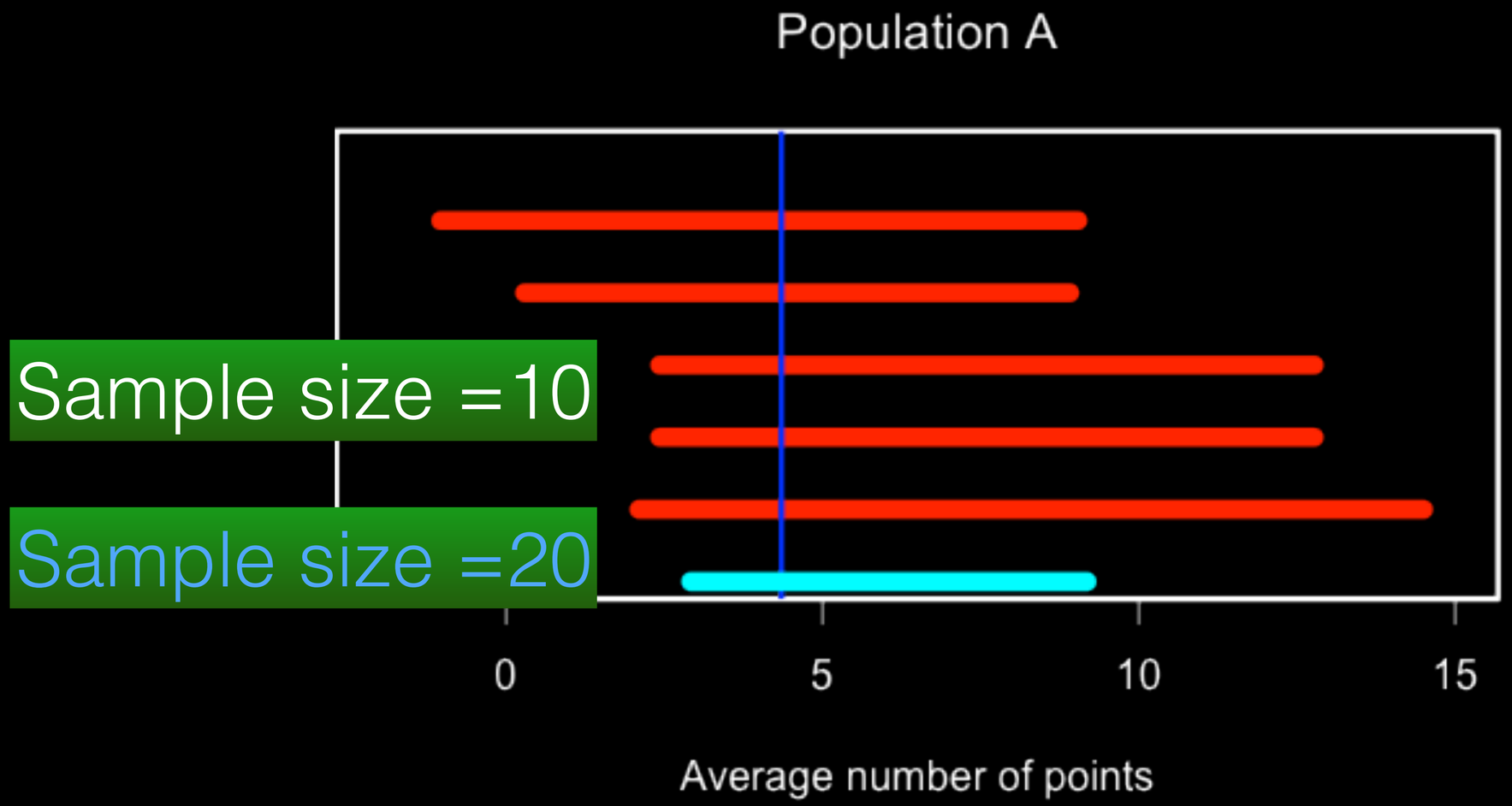
estimate +/- margin of error

- "**Confidence**" is a probability value assigned to the computing process of confidence intervals under assumptions about the data.
- Confidence can take values other than 95%.
- Validity of a confidence interval (in terms of true *coverage probability* depends on the validity of the assumptions.

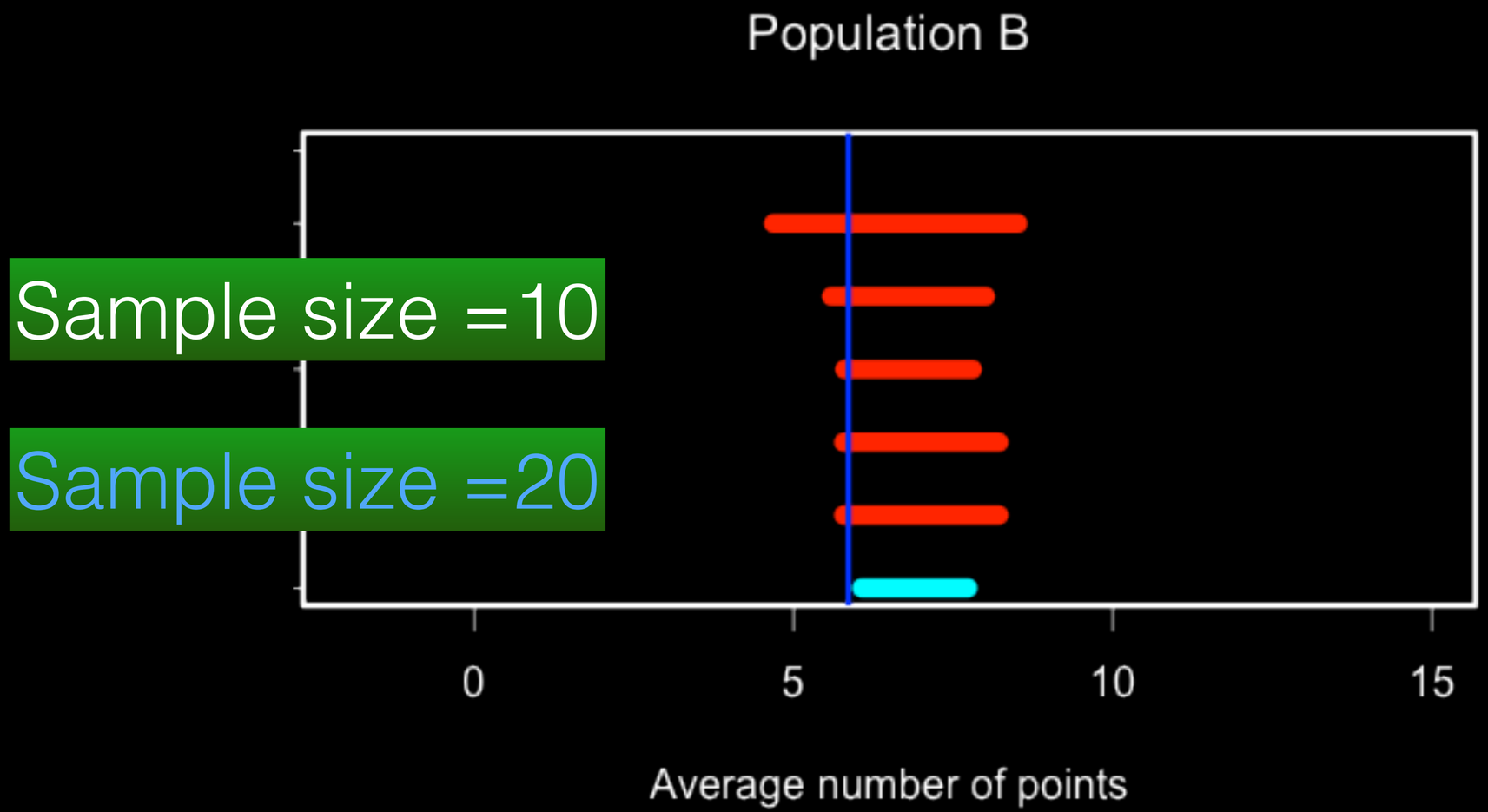
Margin of error

- It equals half the width of a confidence interval.
- Decided by
  - overall variability in the population.
  - confidence level
  - sample size

Lego example confidence intervals



Lego example confidence intervals



Reduce bias is more important.

Significance test

Hypotheses

- **A statement of the value of a population quantity of interest.**
- **It usually *proposes* a model for the population, which is to be evaluated.**
- **The consistency between the observed data and the model is to be evaluated,**
  - **which will be used to decide whether the hypothesis should be rejected or not.**

Example: a biased coin?

- A coin is being check for fairness.
- A fair coin produces heads 50% of the time.
- n=10 tosses are observed for the coin under test.
- The coin is actually biased and produce heads 90% of the time.
- In the 10 tosses, 10 heads are observed for this coin.
    - If the coin is fair, probability of observing 10 heads is <0.001.
    - For this biased coin however, the probability is actually 0.35.
- Statistical significant?

Testing a hypothesis

- **Hypothesis proposes a model for the sampling distribution of a statistic.**
- **Data generate a value of that statistic.**
- **Evidence that is against or supporting a hypothesis should be**
  - **a measure of "distance" (or inconsistency) between the model and the data**
  - **Such distances are usually computed as "test statistics".**
  - **Inconsistency is measured using p-values.**

P-value

is defined as

- the probability,
  assuming the null hypothesis is true
- of the event that
  - we observe evidence
    - against the null hypothesis
    - and support the alternative
  - at the level that is *at least as strong as what is observed (i.e., the **data**)*.

Back to the coin example

- Null hypothesis: a fair coin, P(head) = 50%
- Alternative hypothesis: a biased coin P(head) ≠ 50%
- Data from 10 tosses: number of heads observed 10 (100% heads)
- Evidence against the null: too many or too few heads.
- “Distance”: 100%-50%.
  - We normalize this distance using **standard error** for this sample proportion.
  - p-value, how likely it is for us to observe a sample proportion this “far” from the value stated in the null hypothesis?

Statistical methods aim to have maximum power while controls type I error.

	Truth	
	H_0 is true	H_0 is false
Test conclusion		
Accept H_0		
Reject H_0		

		Truth	
		H_0 is true	H_0 is false
Test conclusion	Accept H_0	True negative	False negative
	Reject H_0	False positive	True Positive

What does “reject the null hypothesis” mean?

- [wrong!] The null hypothesis is not true?
- [wrong!] That the alternative hypothesis is true?
- Here is what we mean: *the difference between the data and the null hypothesis is statistical significant at a pre-decided level SO THAT we reject the null and accept the alternative.*
  - *(the strength of the evidence is beyond reasonable doubt).*
- Rejecting the null hypothesis does NOT mean the real value is much different from the claimed value in the null hypothesis.

Google

bacon carcinogen

Web Images News Shopping Videos More Search tools

The cancer-causing effect of bacon has been found to be statistically significant.

Bacon is Carcinogenic, but So is the Air You Breathe
Newsweek - Oct 30, 2015
When the World Health Organization officially designated processed meats like **bacon** and sausage a definite (Group 1) **carcinogen** in

The awful truth about yummy **bacon**

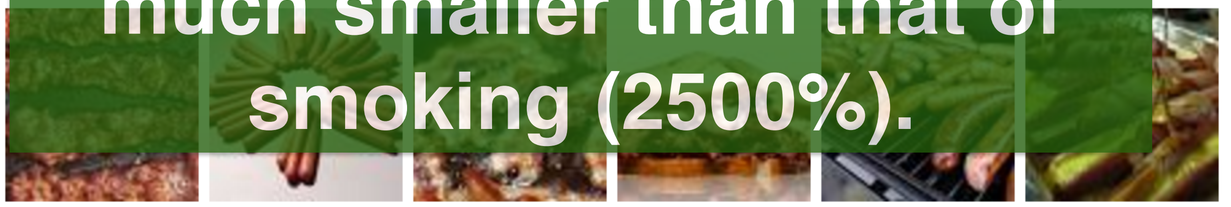
New Health Warning Explained: How Processed Meat Is Linked to ...
Live Science - Oct 30, 2015
Bacon, Cell Phones, Glyphosate. Which Potential **Carcinogen** ...

So Will Processed Meat Give You Cancer?
Opinion - New York Times - Oct 31, 2015

In-Depth - TIME - Oct 29, 2015

WHO puts it in the same category as smoking!

The estimated effect size of processed meat (18%) is much smaller than that of smoking (2500%).



Statistics and Probability I
Concluding remarks

- Good practices on
  - sampling
  - studying central trends, variation and association
- measuring sampling variation for statistical inference
- experimentation for establish a cause-effect link

Statistical challenges

- **Design data generation and analysis for nontraditional data such as social media, customer behaviors, etc.**
- **Handle huge amount of existing data that were not necessarily collected in the most ideal way.**
- **Need to study the behaviors of our statistical methods and models when applied to such data**
- **Create adjustments to derive reliable findings**