THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

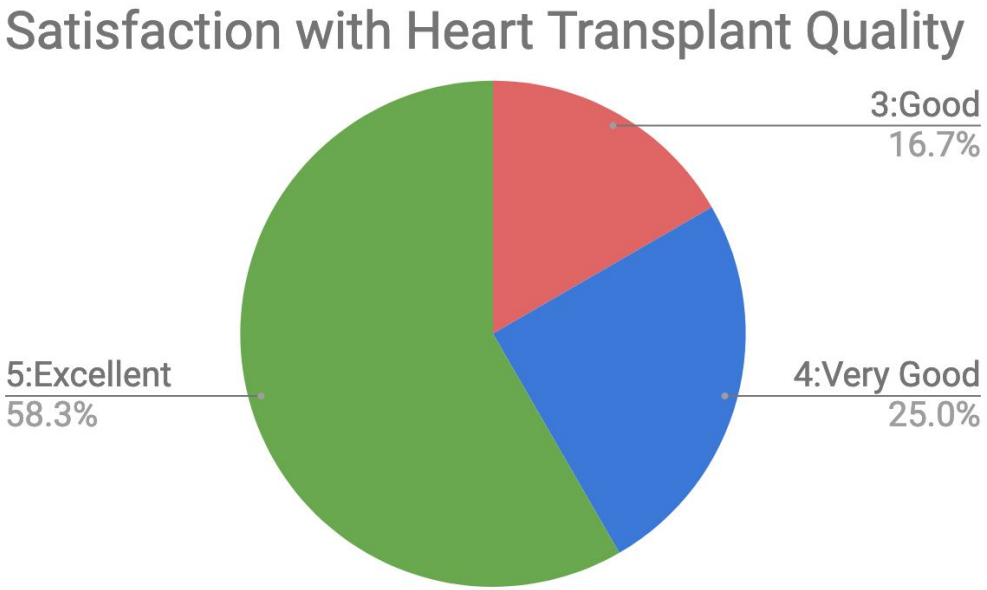
Why Statistics?

Pablo Frank Bolton

Trust me

I have statistics!

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC



In Hospital 1, a survey about user satisfaction with heart transplants showed that the average obtained was 4.4 stars out of 5!!!

Math major or Business major?

Tom is shy

Math major or Business major?

Tom is shy
Students in GWU Business: 4000
Students in GWU Mathematics:~400

Why Study Probability?

After all, we know obvious when we see it, right?

Let’s Play a Game!

Flipping for a Grade?

Rules:

- We will flip a coin 3 times
- You each go against me
- You choose a 3-toss combination
- I choose a 3-toss combination
- Whichever triad comes first, wins

Conditions:

- If I don't win more than ⅔ of the time, you get an A on the course.
- If I do win more than ⅔ of the time, you buy me a car.

A Fair Coin

Tally Results

Pablo VS Students

You don’t believe it?

Other examples:

- Monty Hall Problem
- Single coin tossing HH VS HT
- St. Petersburg Casino

Indubitably!

Dear Watson!

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

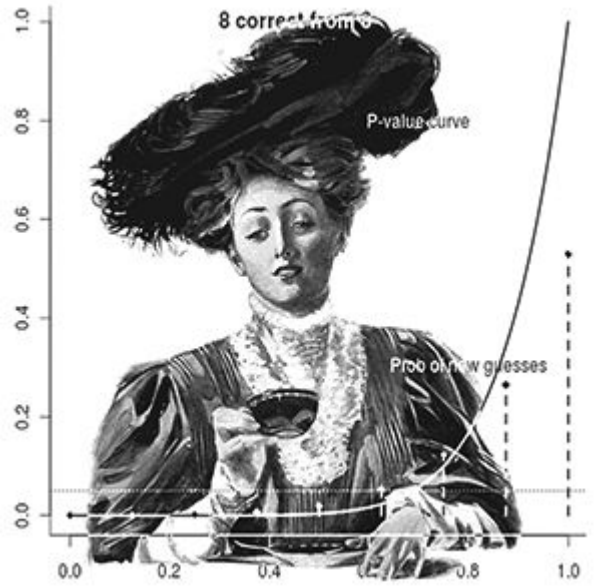
Why Use Statistics?

After all, a thing either works or it doesn’t, right?

A clear-cut result

Lady Tasting Tea

- What questions could we possibly ask about this?



Why Use Statistics?

It’s just Academic, no?

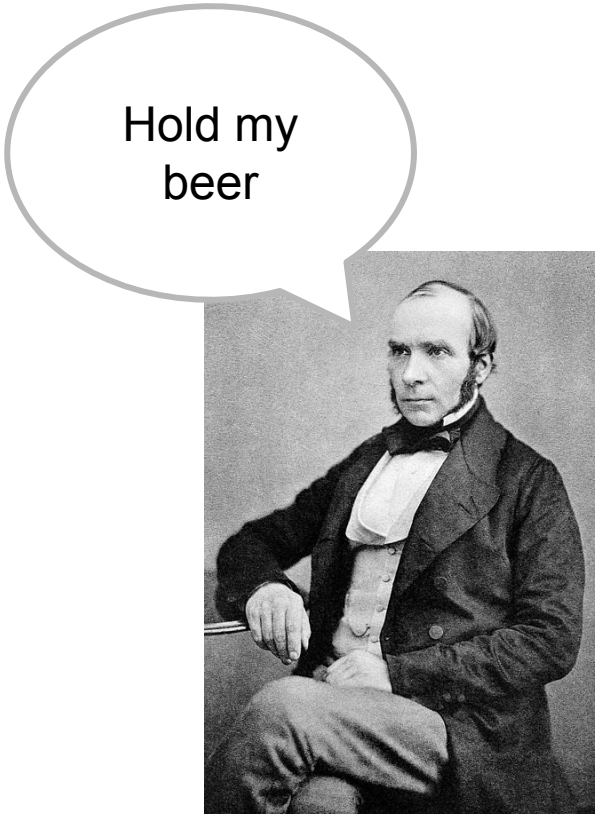
London 1854

A killer lurks under the London fog. So far 127 have been claimed, and by the end of its reign of terror, 616 would be gone.

Sherlock Holmes has failed. Who should we call?

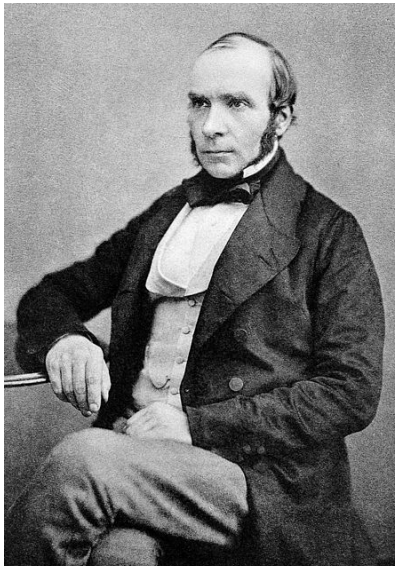
John Snow, Obviously!

Who you gonna call?

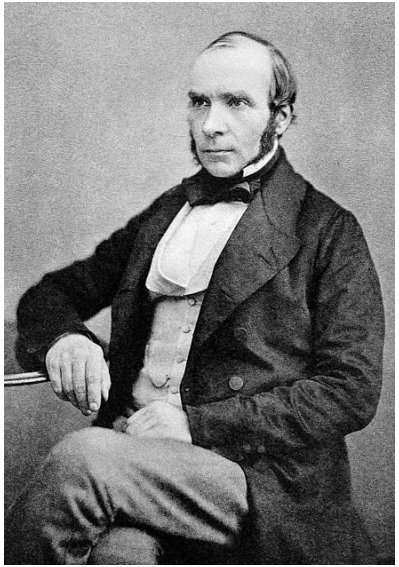


The Prevailing Theory: Miasma from the river Thames is causing the Cholera outbreak.

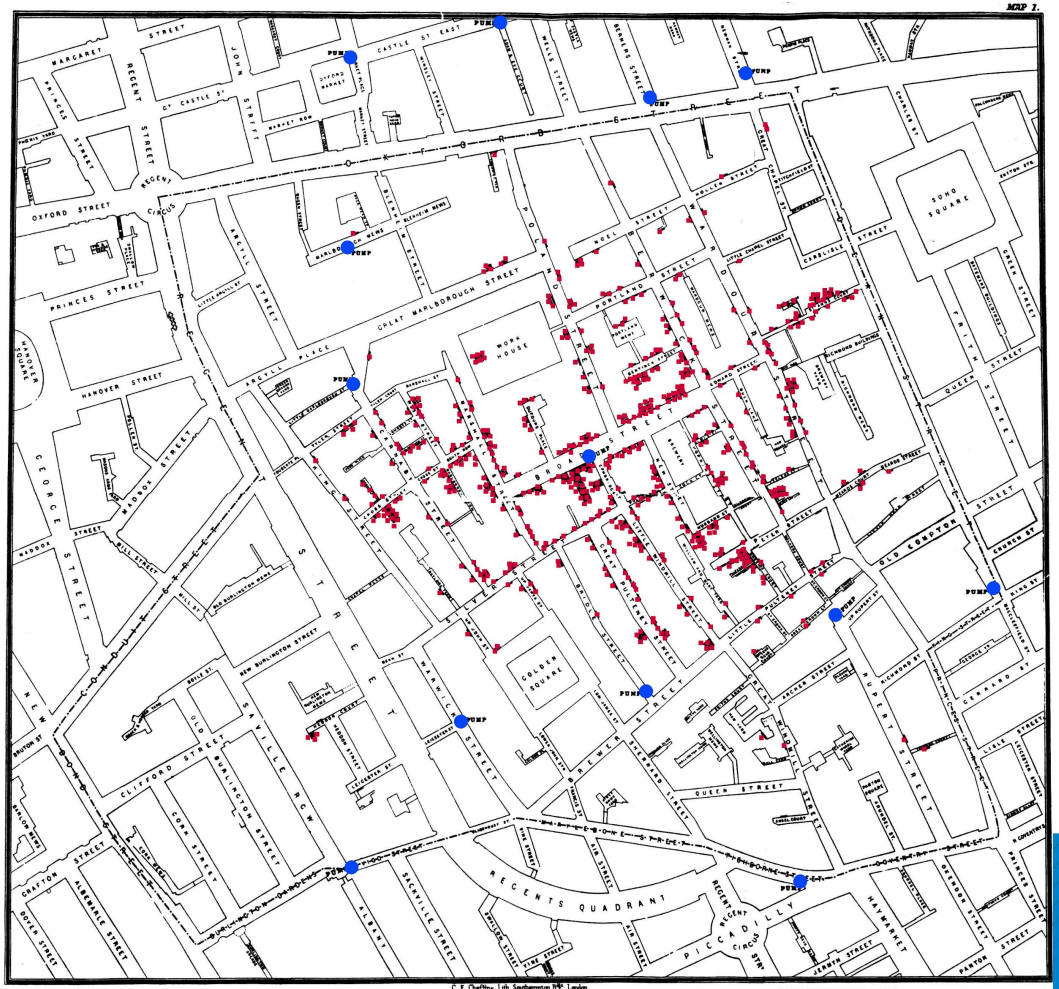
Nobody can stop the air!!
But maybe, just maybe, we don’t need to.

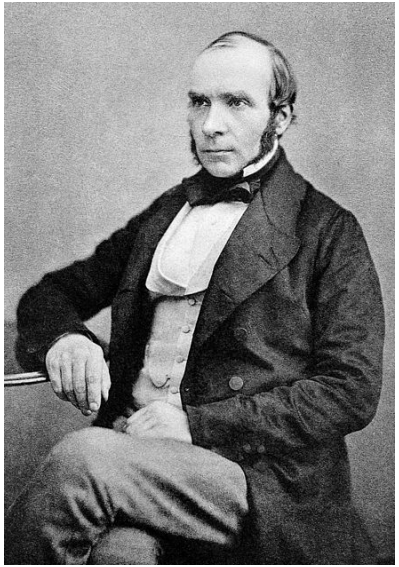


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC





Government officials replaced the Broad Street pump handle. They had responded only to the urgent threat posed to the population, and afterward they rejected Snow's theory. The route of disease transmission would have been too unpleasant for most of the public to contemplate.

It wasn't until 1866 that William Farr, one of Snow's chief opponents, realized the validity of his diagnosis when investigating another outbreak of cholera at Bromley by Bow and issued immediate orders that unboiled water was not to be drunk.

Farr denied Snow's explanation of how exactly the contaminated water spread cholera, although he did accept that water had a role in the spread of the illness. In fact, some of the statistical data that Farr collected helped promote John Snow's views.

Research Objective

Objective	Justification
● Do something better ○ How much better? ○ In which cases?	● Faster, less waste, more gain!
● Explain a phenomenon ○ Associate variables? ○ Prove Causality?	● Make better decisions (predict)

What are (is) statistics?

Statistics: Field of study

- Data collection
- Analysis
- Interpretation

Statistic: Measurement

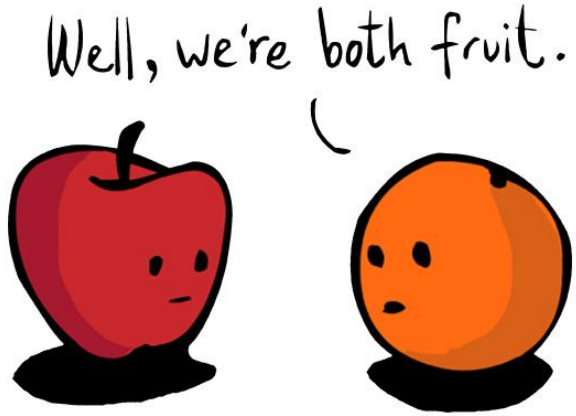
- Of a characteristic in the observed data

Two main Flavors:

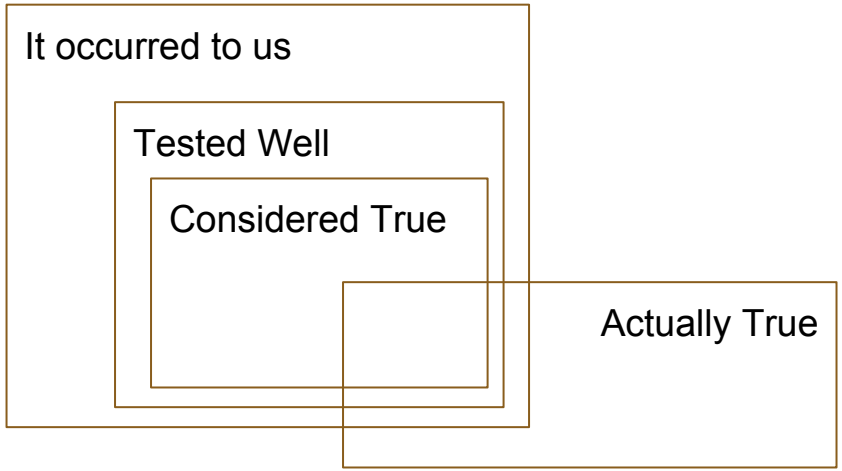
- Descriptive Statistics
- Inferential Statistics

Two “practical” reasons to use statistics

Comparing



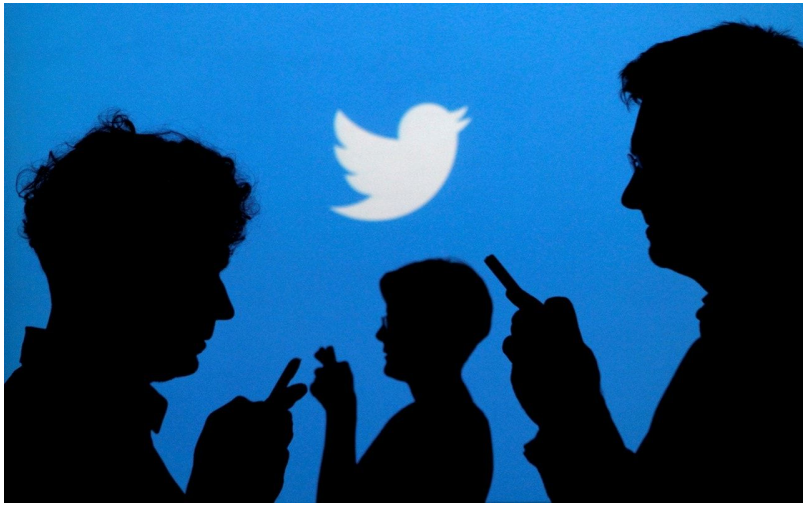
Explaining



What Data do we Statisticizisize?

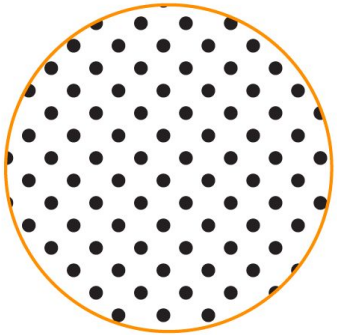
Example:
You've created a Twitter filter that estimates the age of the author.
Question: How good is your filter?

What Data do we analyze?

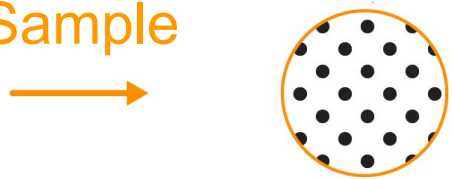


Population VS Sample

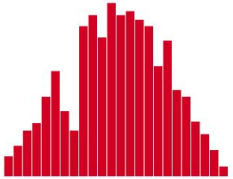
Population



Obtain Sample

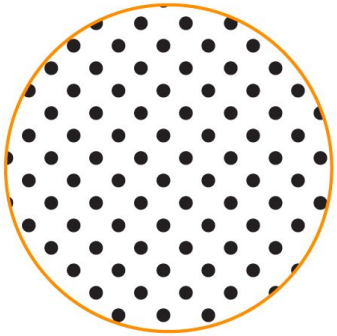


Sample

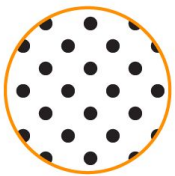


Population VS Sample

Population

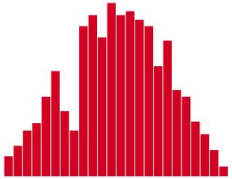


Obtain Sample

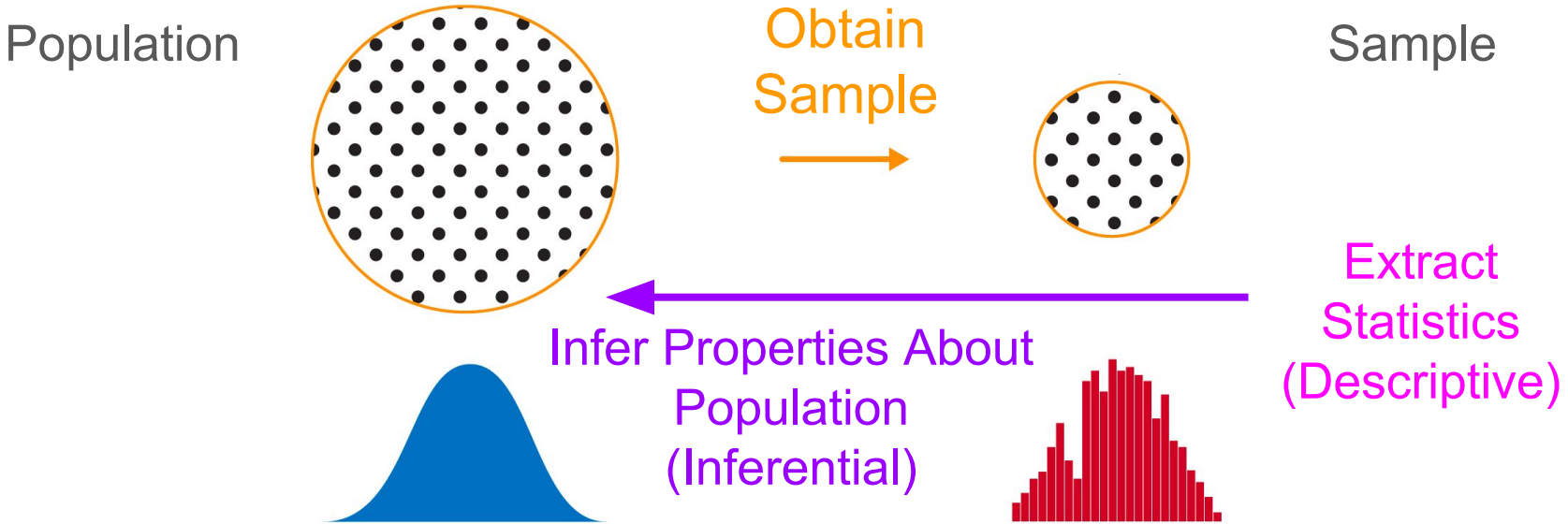


Sample

Extract Statistics (Descriptive)



Population VS Sample



Picking what to sample

- Probability Sampling
  - Simple Random Sampling
  - Systematic Sampling
  - Stratified Sampling
- Non-Probability Sampling
  - Convenience Sampling
  - Quota Sampling

Other Considerations

Scales:

- Interval vs Ratio(Continuous)
- Ordinal (Discrete)
- Categorical/Nominal (Discrete)

Examples

- Heights in students in class
- Letter grades in class
- Color of clothing in class

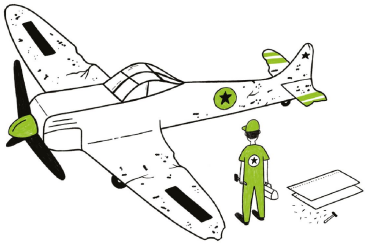
Why is Sampling an Issue?

Comparison Case

FDA Anecdote

Explanation Case

WWII Challenge

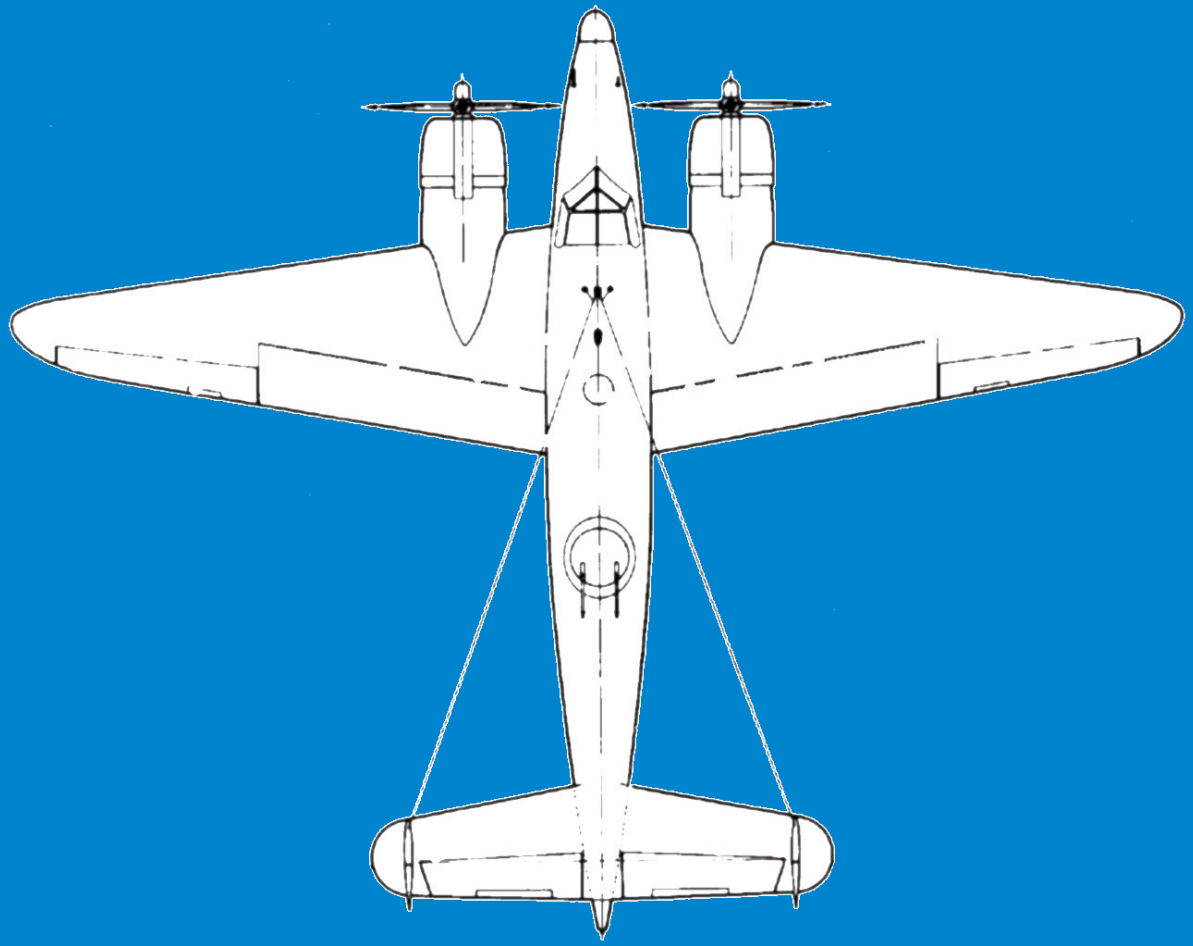


Scenario 1 (Easy Puzzle)

INDIVIDUAL Mission:

- You are a WWII Engineer.
- Given the damage seen on planes that returned from reconnaissance runs:
- Make Modifications

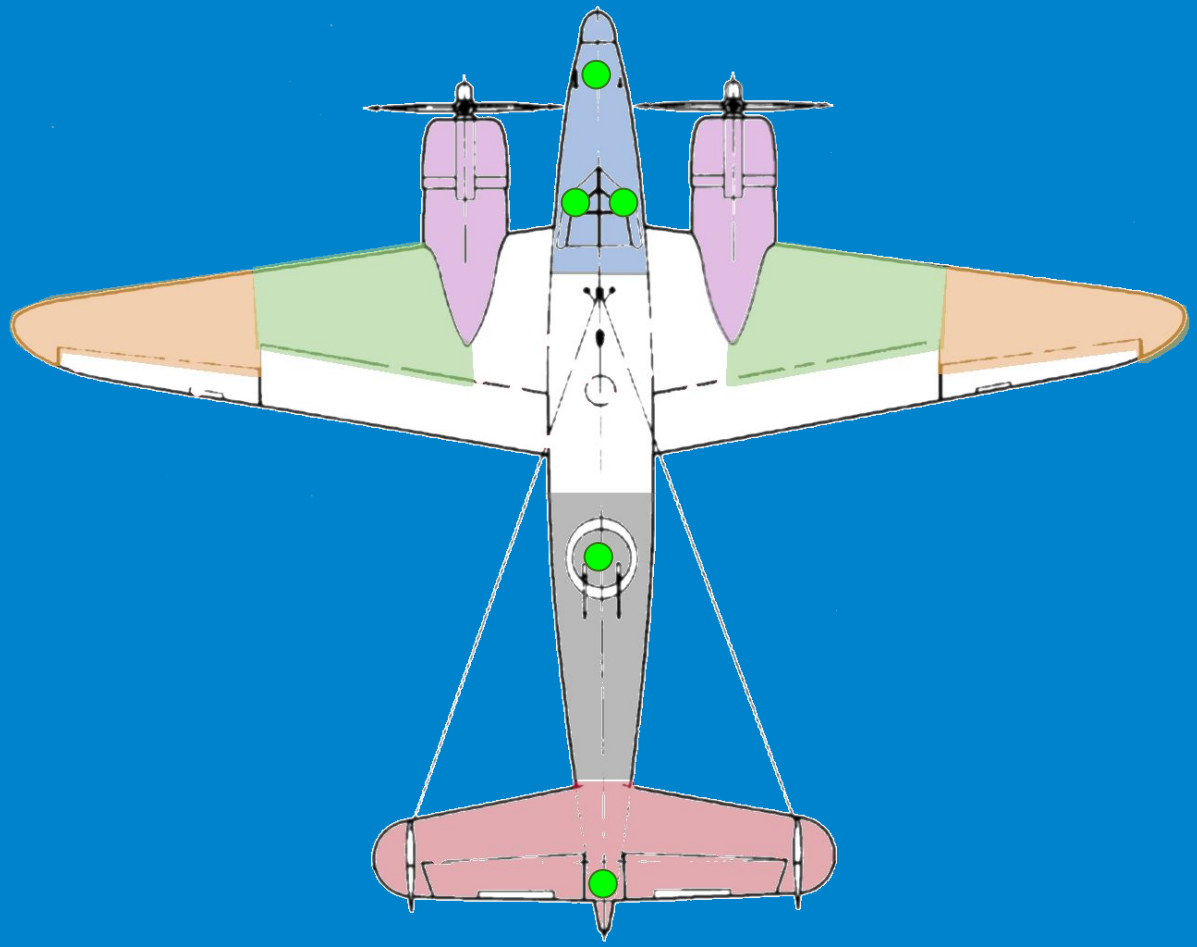
WWII Recon Runs



The Plane

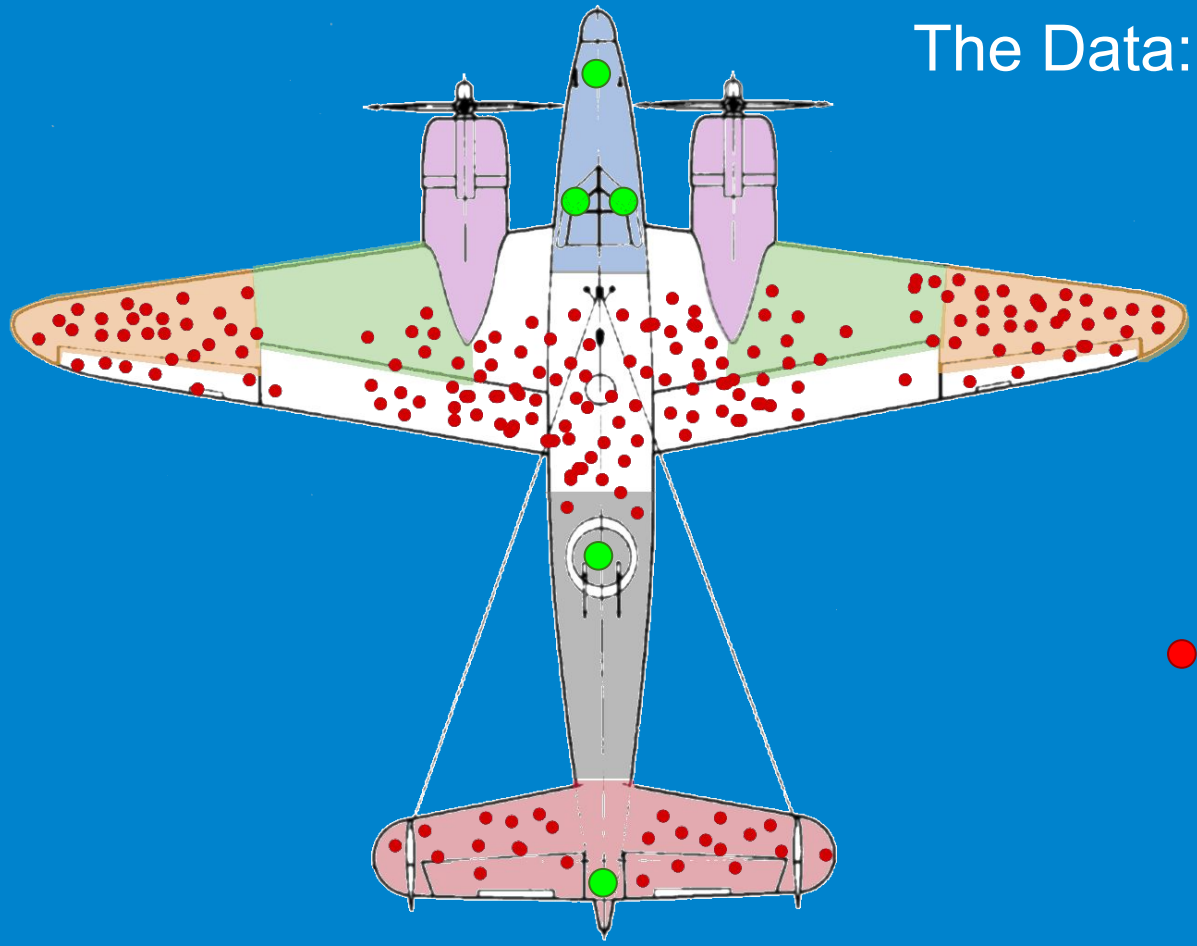
The Swiss cheese

The Parts



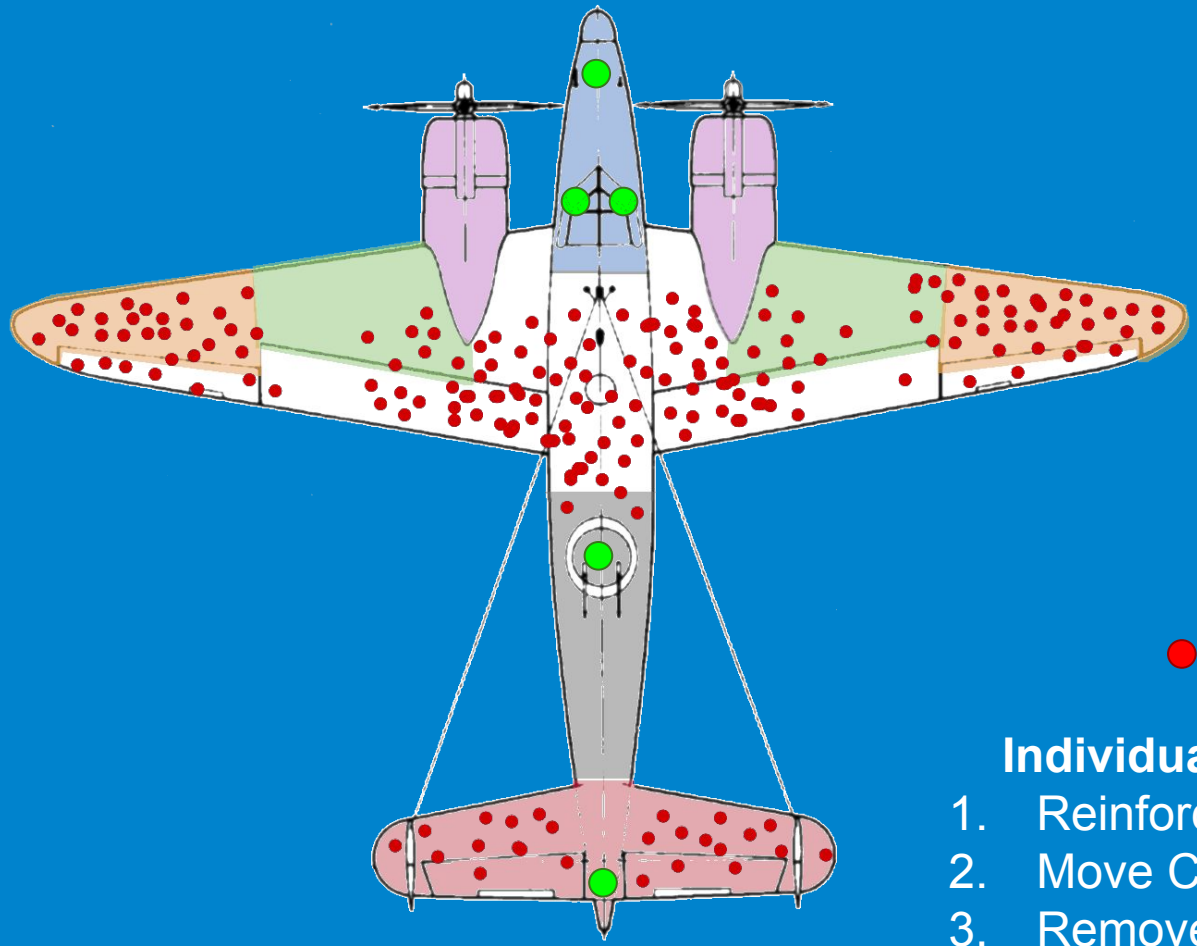
- Crew Member
- Cockpit
- Engines
- Fuel
- Wing Tips
- Front fuselage
- Back fuselage
- Tail

The Data: Scenario 1 (obvious)



- Crew Member
- Cockpit
- Engines
- Fuel
- Wing Tips
- Front fuselage
- Back fuselage
- Tail
- Hit From Air Defenses

The Data



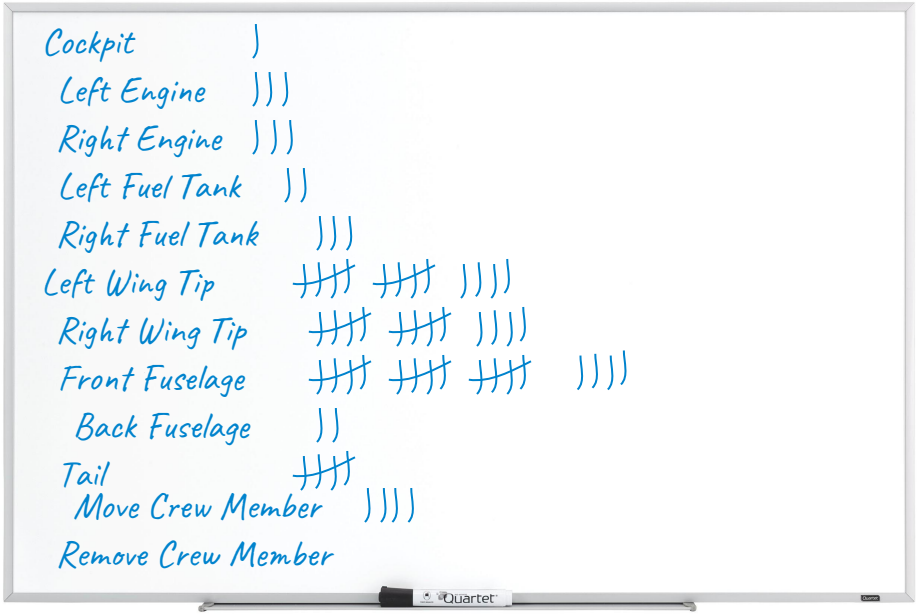
- Crew Member
- Cockpit
- Engines
- Fuel
- Wing Tips
- Front fuselage
- Back fuselage
- Tail
- Hit From Air Defenses

Individually: Pick Any 4 Actions

1. Reinforce 1 Area (choose)
2. Move Crew member
3. Remove Crew Member

Count Actions

- For each action count:
  - Votes if action taken
- Show tally

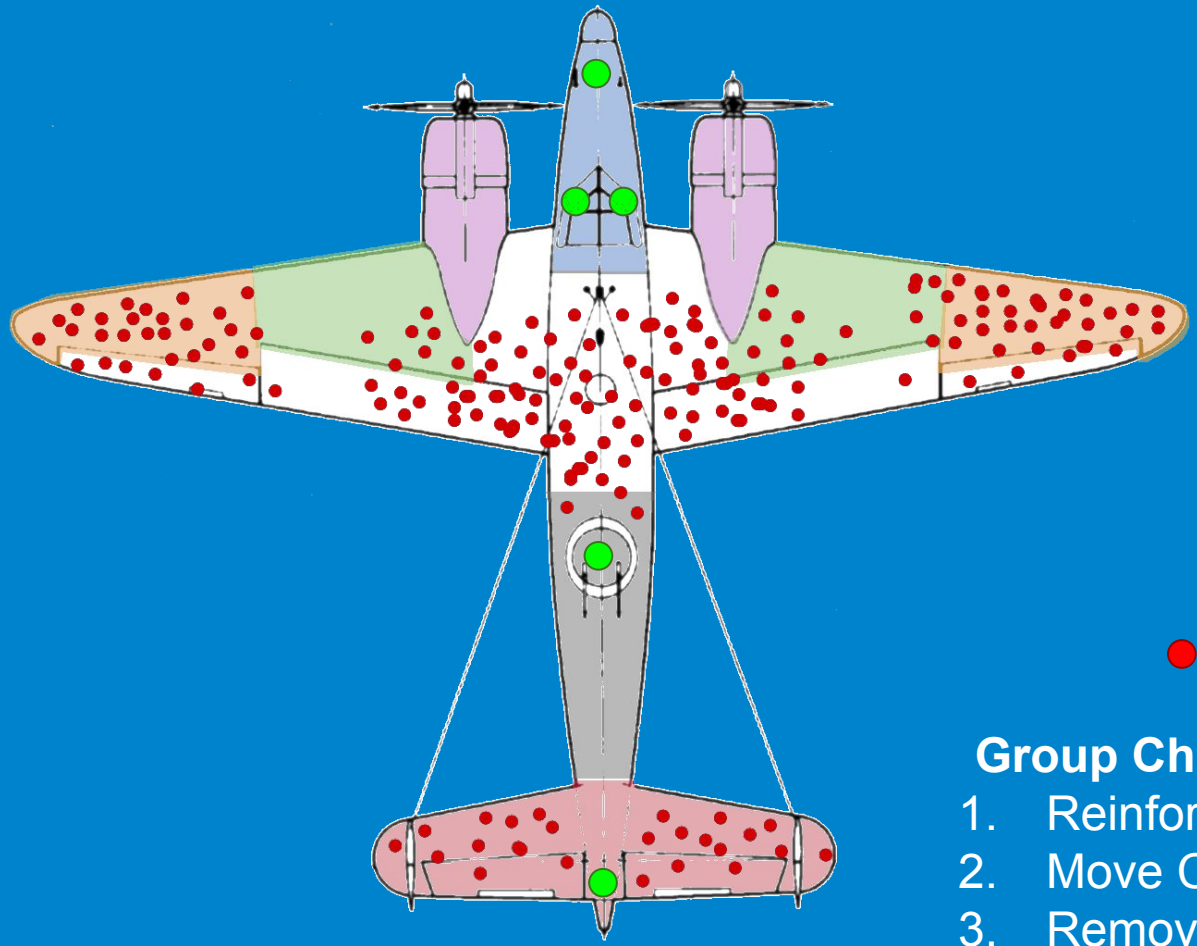


Scenario 1 (Obvious Puzzle)

GROUP Mission:

- Given the same data
  - Form Groups of 3.
  - Discuss your reasoning
  - New Vote.

WWII Recon Runs



The Data

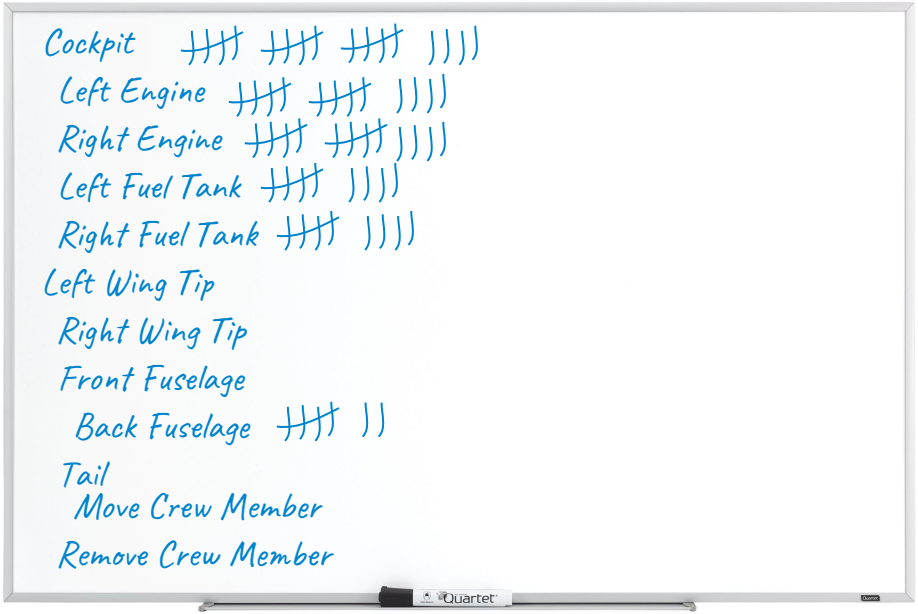
- Crew Member
- Cabin
- Engines
- Fuel
- Wing Tips
- Front fuselage
- Back fuselage
- Tail
- Hit From Air Defenses

Group Choice: Pick Any 4 Actions

1. Reinforce 1 Area (choose)
2. Move Crew member
3. Remove Crew Member

Count Actions

- For each action count:
  - Votes if action taken
- Show tally



Puzzle 2 Easier

INDIVIDUAL Mission:

- In WWI the british used simple cloth hats.
- There were a LOT of head injuries.
- Metal helmets were introduced.
- Surprisingly, more soldiers were hospitalized with head injuries.
- Should you go back to the caps?

Cap Vs Helmet

FDA Opioid Testing

What Should Be:

- Split Group Into Control and Experimental Groups
- Give Opioid to Experimental
- Give Placebo to Control
- Check how many have less pain without side effects

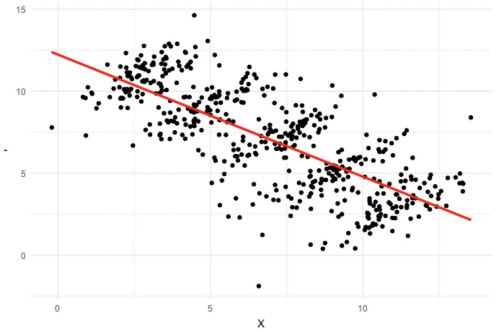
What Was:

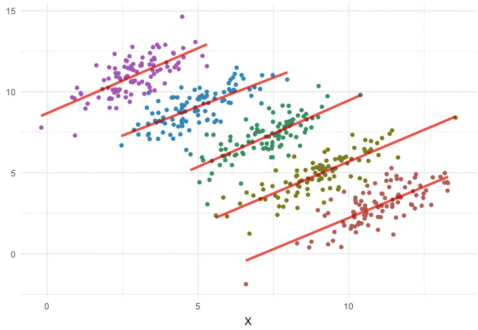
- Give Opioid to Experimental
- Let people drop out.
- “Wean” people from opioids.
- Split Group Into Control and Experimental Groups.
- Give Placebo to Control
- Check how many have less pain without side effects

Back to Basics: Two objectives.

Comparing

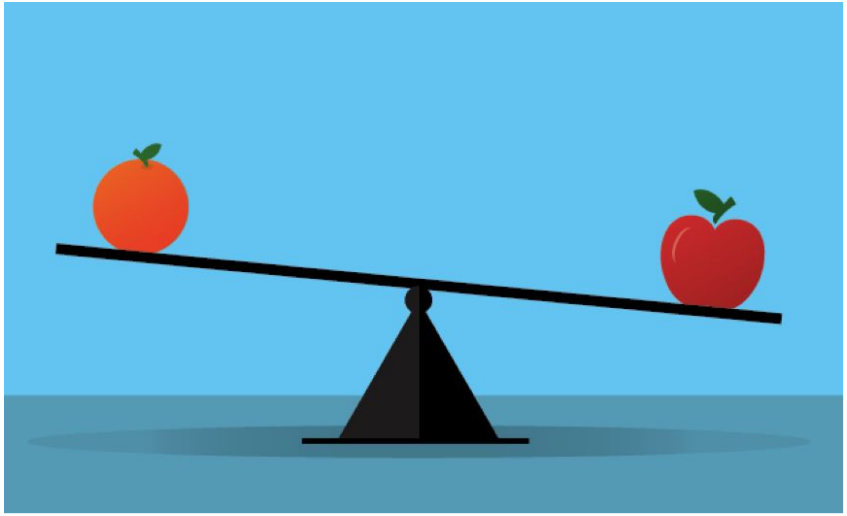
Explaining



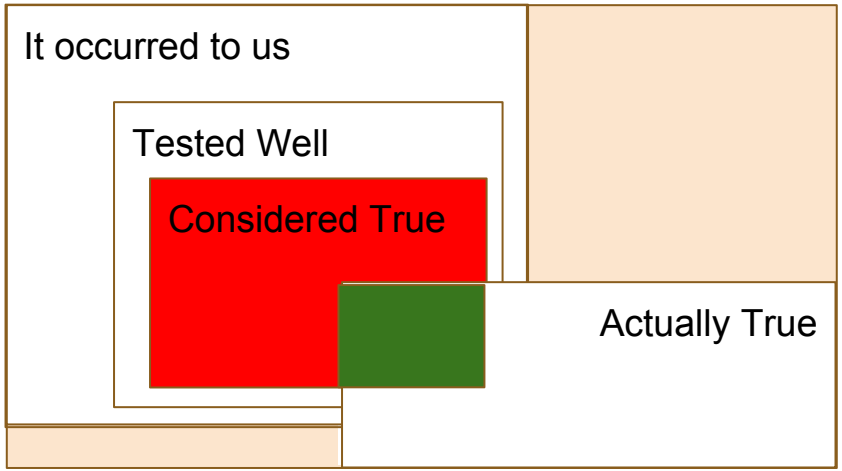


Comparing

- Your method VS the competition
- Decision between alternatives
- Before-After a change



Explaining



"It ain't so much the things we don't know that get us into trouble. It's the things we know that just ain't so."

— Artemus Ward

Descriptive Statistics

Basics

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Considerations about Data

- **Stochasticity** of a process and comparing processes
- **Complexity** of a process and comparing processes
- **Interdependence** between alternatives (variables or processes)
- **Information** that you have at your disposal to compare

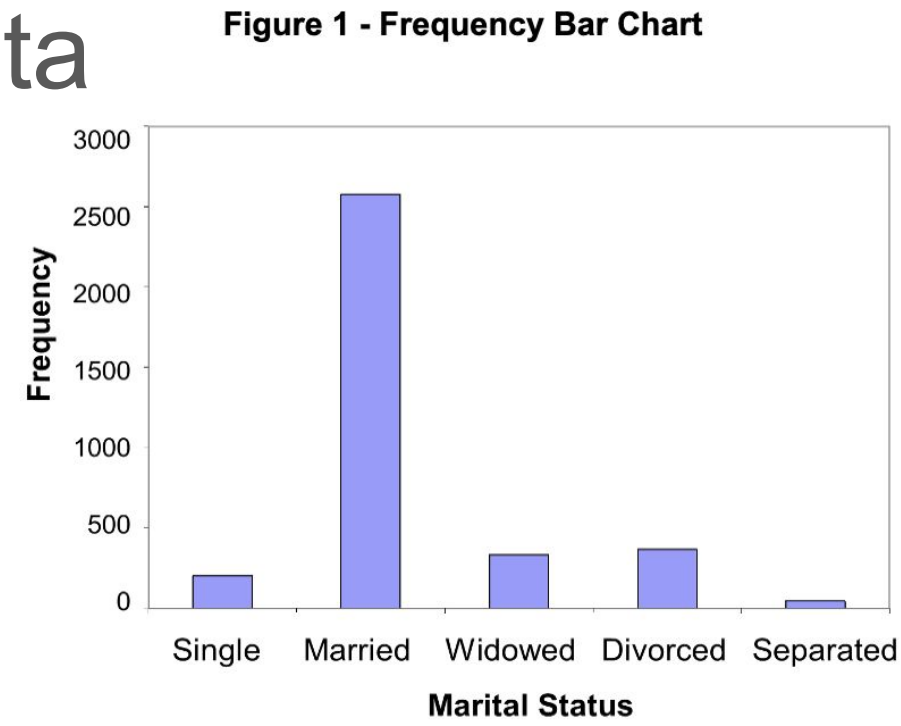
Summarizing Data

Frequency Distribution Table

Smoking Status	Frequency	Relative Frequency, %
Non-Smoker	1,330	37.6
Former	1,724	48.8
Current	482	13.6
Total	**3,536**	**100.0**

Summarizing Data

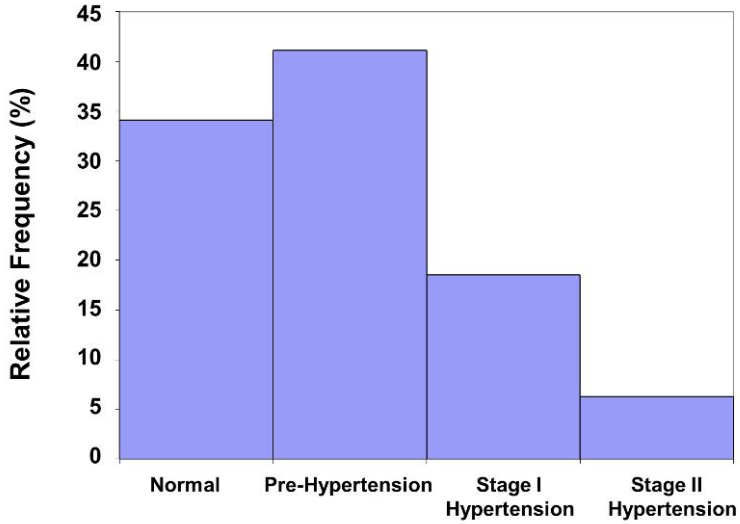
Frequency Distribution Chart



Summarizing Data

Histogram

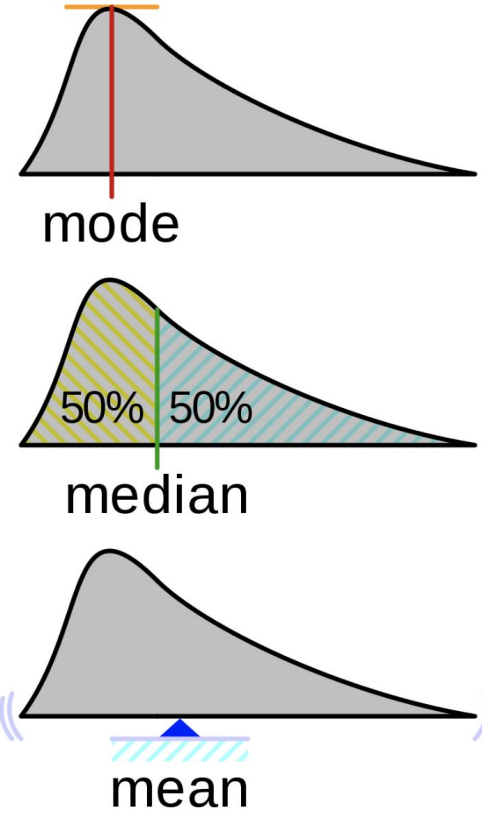
Figure 8 - Relative Frequency Histogram for Blood Pressure



Mean, Median, Mode

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

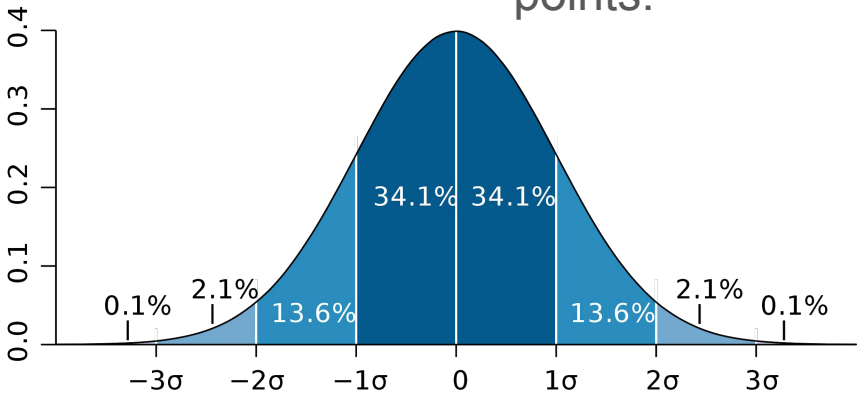
Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**



Standard Deviation Vs Variance

Standard deviation looks at how spread out a group of numbers is from the mean, by looking at the square root of the variance.

The variance measures the average degree to which each point differs from the mean—the average of all data points.



Standard Error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

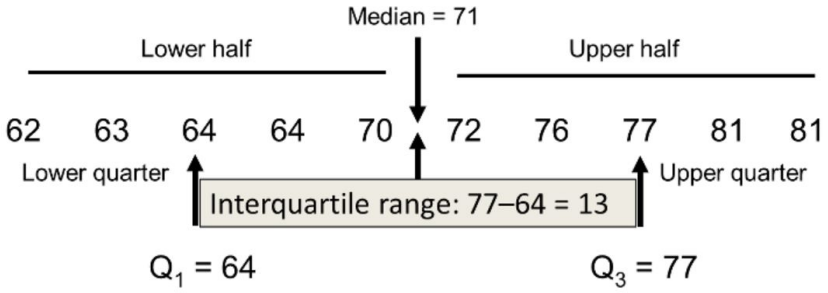
$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

$$\mathrm{s}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Inter-Quartile Range

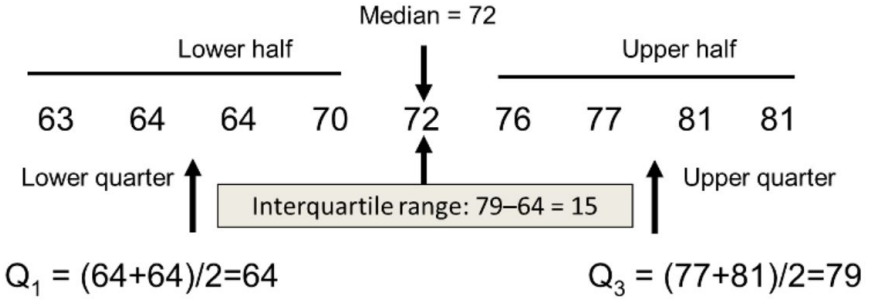
EVEN

Figure 9 - Interquartile Range with Even Sample Size

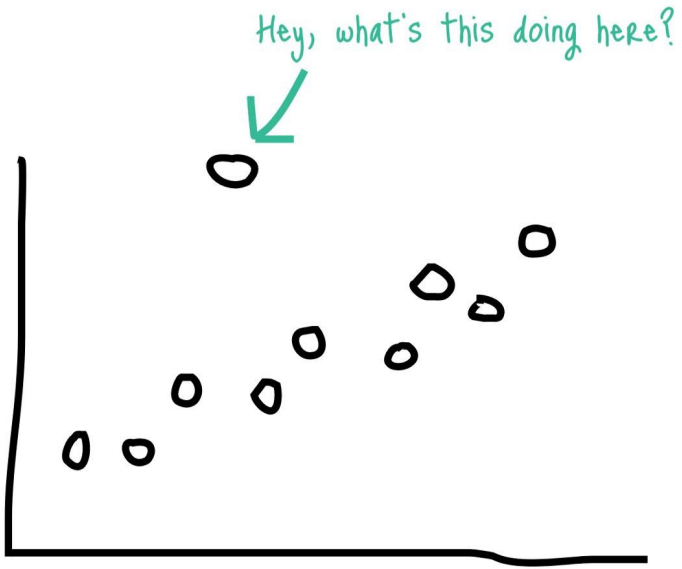


ODD

Figure 10 - Interquartile Range with Odd Sample Size

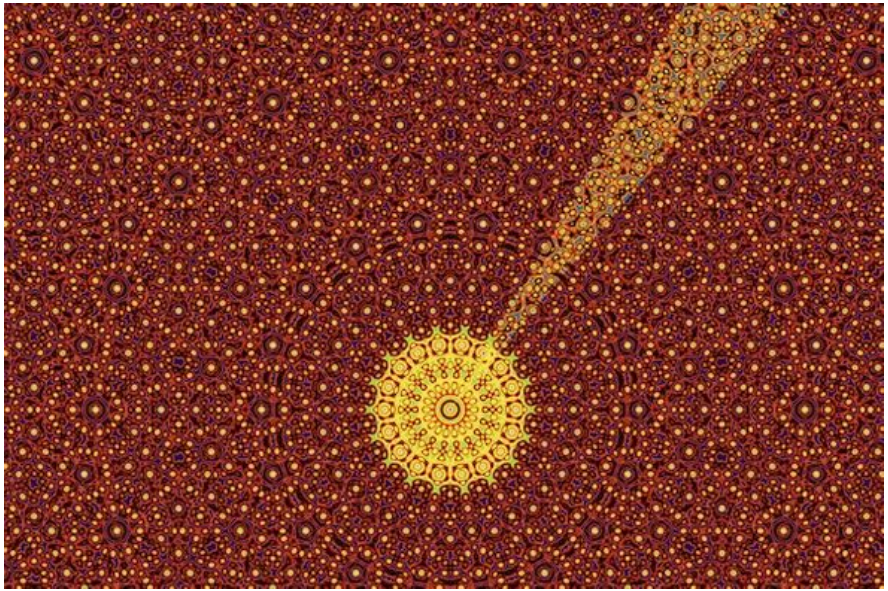


Outliers



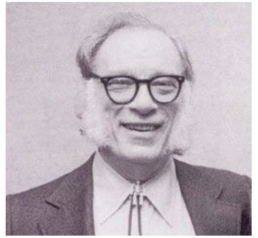
- Tukey Fences: Outliers are values below Q1-1.5(Q3-Q1) or above Q3+1.5(Q3-Q1)
- RANSAC

Don’t Discard to Quickly



The most exciting phrase
to hear in science,
the one that heralds new discoveries,
is not ‘Eureka!’ (I found it!)
but ‘That's funny ...’”

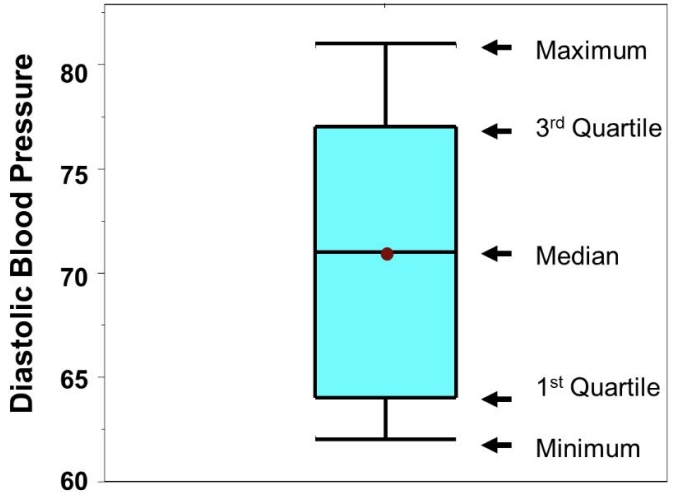
- Isaac Asimov



Box & Whiskers

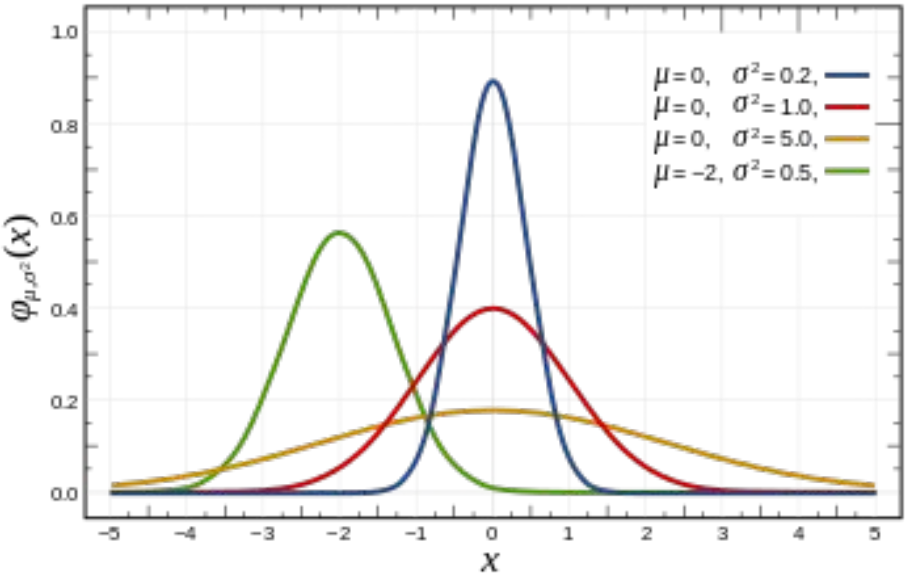
Minimum:	62
Q_1:	64
Median:	71
Q_3:	77
Maximum:	81

Figure 11 - Box-Whisker Plot of Diastolic Blood Pressures in Subsample of n=10.



One Key Idea

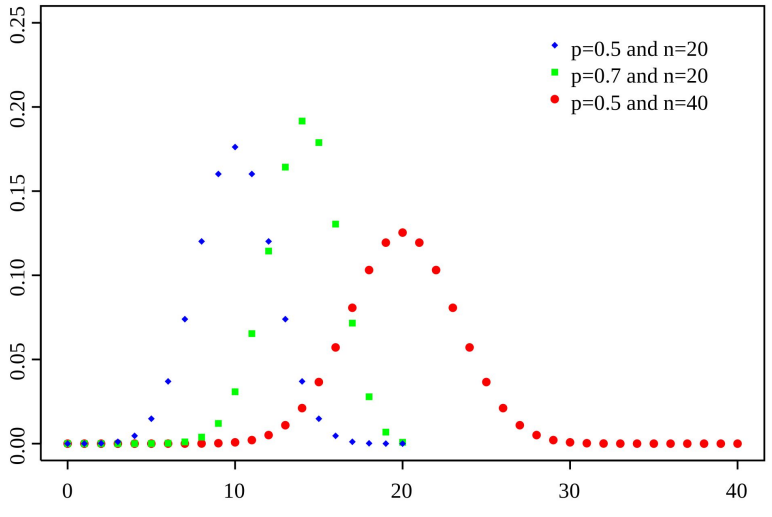
Compare sample distribution models rather than the sets themselves.



Probability Models: Binomial

- Objective: Count # “successes”
- Two possible discrete outcomes
- Same probability per attempt
- Independent events
- n is finite

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Exercise!

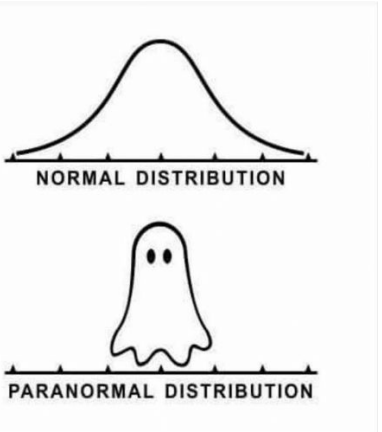
What is the probability of getting 2 heads in 5 fair coin flips?

Exercise!

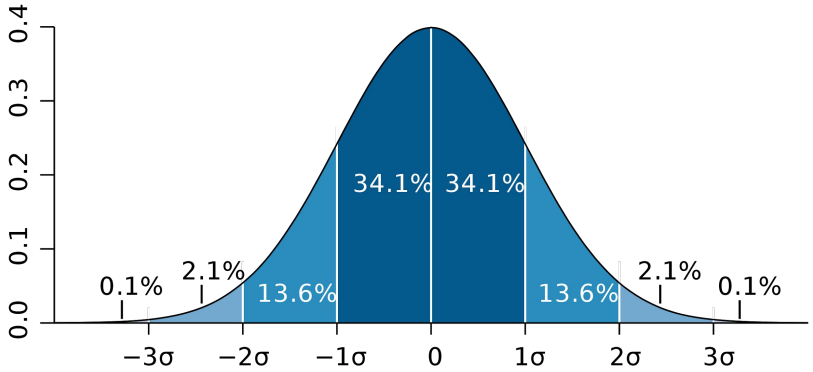
What is the probability of getting 2 heads in 5 fair coin flips?

Do we care about order?
Where do I put the 1st Coin? (5 places)
Where do I put the 2nd Coin? (4 places)
Where do I put the 3rd Coin? (3 places)
How many ways of arranging 3 coins? (3!)
Result: $(5*4*3)/3! * (0.5)^{(5-3)} = 5/16 = 0.3125$

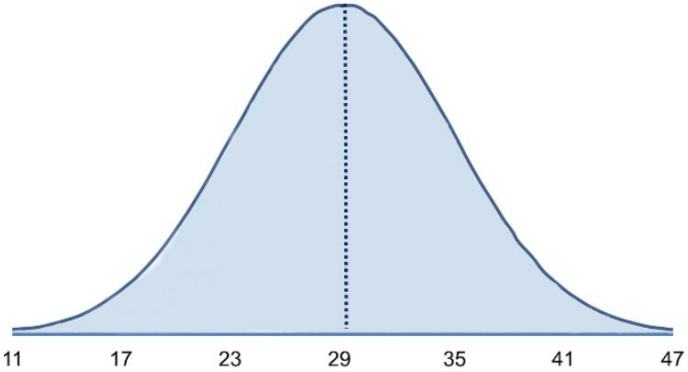
Probability Models: Normal



- Many possible continuous outcomes
- “Small” variations around a central value



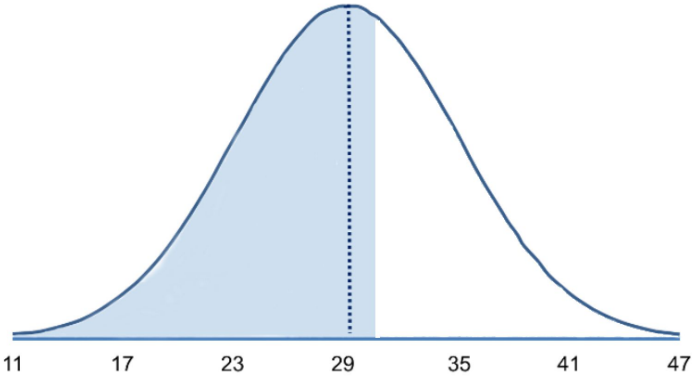
Example Question:



Consider body mass index (BMI) in a population of 60 year old males in whom BMI is normally distributed and has a mean value = 29 and a standard deviation = 6.

What is the probability that a randomly selected male from this population would have a BMI less than 30?

Example Question:



A BMI of 30 is 30-29=1 BMI unit above the means. The standard deviation is 6, so 1 BMI unit above the mean is 1/6 = 0.166667 standard deviations above the mean.

We call this a Z-Score

Standardized-Score

$Z = (X - \mu) / \sigma$

The probability that a male aged 60 has BMI less than 30 is 56.75%.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

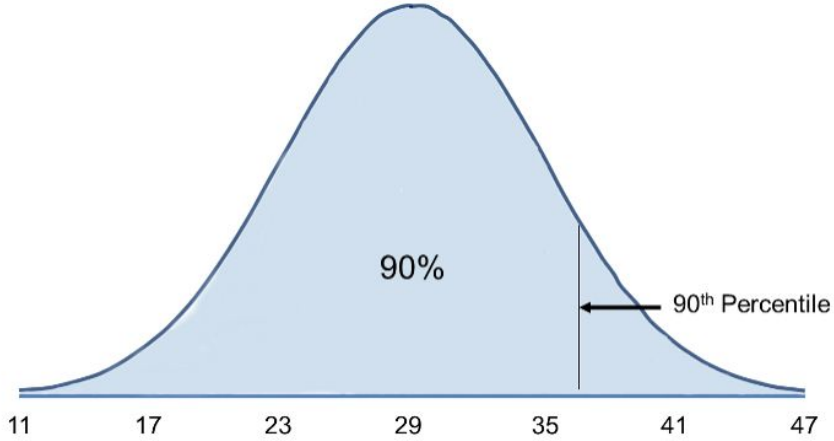
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409

Percentiles

To compute the 90th percentile, we use the formula $X=\mu + Z\sigma$
Find the corresponding Z value for 90% (~1.282)

$X = 29 + 1.282(6) = 36.69.$

Ninety percent of the BMIs in men aged 60 are below 36.69.



Estimating Parameters (Comparing)

	Parameters Being Estimated	
	Continuous Variable	Dichotomous Variable
One Sample	mean	proportion or rate, e.g., prevalence, cumulative incidence, incidence rate
Two Independent Samples	difference in means	difference in proportions or rates, e.g., risk difference, rate difference, risk ratio, odds ratio, attributable proportion
Two Dependent, Matched Samples	mean difference	

Confidence Intervals

A range of likely values for the population parameter based on:

- the point estimate, e.g., the sample mean
- the investigator's desired level of confidence (most commonly 95%, but any level between 0-100% can be selected)
- the sampling variability or the standard error of the point estimate.

$$\text{Upper 95\% limit} = \bar{x} + (\text{SE} \times 1.96), \text{ and}$$

$$\text{Lower 95\% limit} = \bar{x} - (\text{SE} \times 1.96).$$

Confidence Intervals B

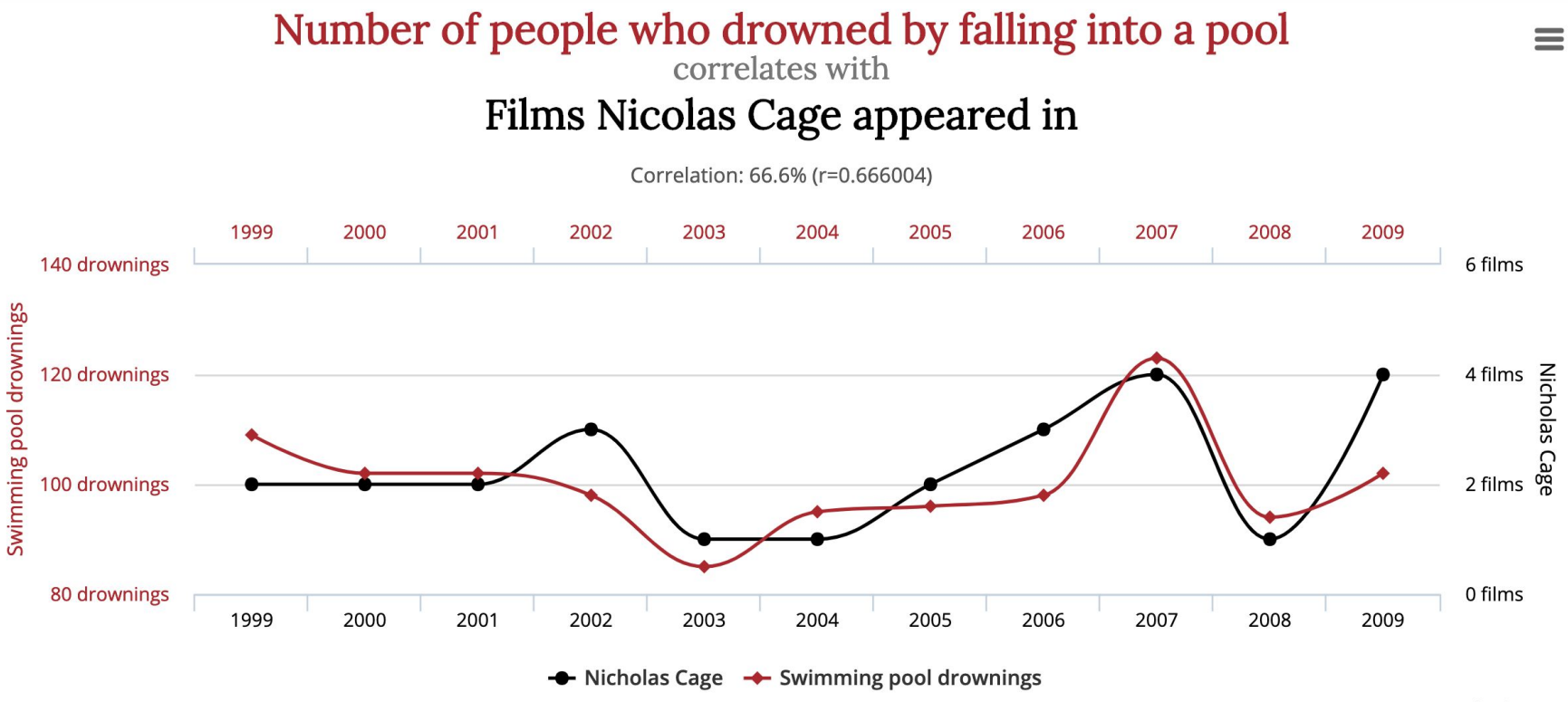
- Another way of thinking about a confidence interval is that it is the range of likely values of the parameter (defined as the point estimate + margin of error) with a specified level of confidence (which is similar to a probability)
- There are many ways of setting confidence intervals for different point samples (single, difference, proportion, ...)

Inferential Statistics

Basics

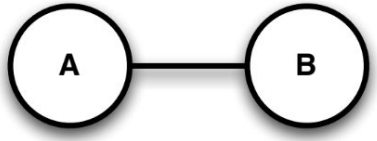
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

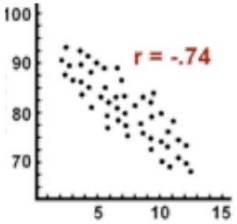
Association, Correlation, Causation



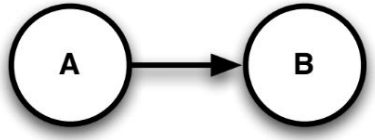
Conditions for causal inference

- Association

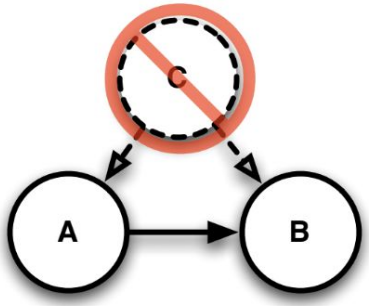


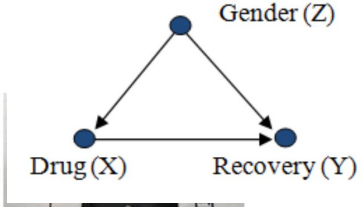


- Direction



- Elimination of potential common causes





Deduction, Induction, Abduction

Deduction

(P)All men are mortal
(P)Socrates is a man

(C)Socrates is mortal

Induction

Several Instances → General Rule

All biological life forms that we know of depend on liquid water to exist.
Therefore, if we discover a new biological life form it will probably depend on liquid water to exist.

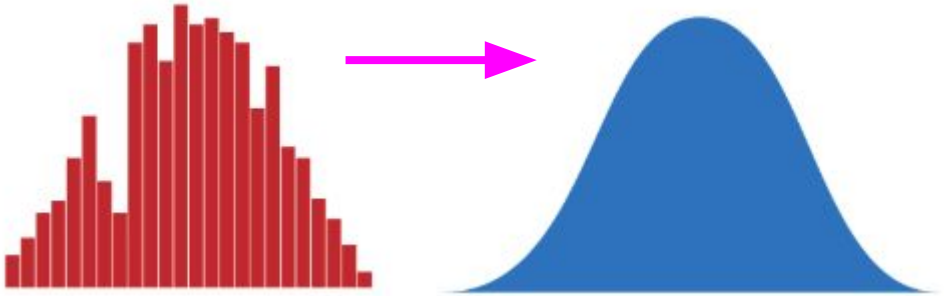
Abduction

Observations → Best explanation

“when you have eliminated the impossible, whatever remains, however improbable, must be the truth”

Our Objective

Infer something about the population from analyzing the sample.



Hypothesis Testing

A specific statement or hypothesis is generated about a population parameter, and sample statistics are used to assess the likelihood that the hypothesis is true.

Hypothesis Testing Process

1. The process of hypothesis testing involves setting up two competing hypotheses, the null hypothesis and the alternate hypothesis.
2. Select a random sample (or multiple samples when there are more comparison groups), compute summary statistics
3. Assess the likelihood that the sample data supports the research or alternative hypothesis.

Example H_0 and H_1

An example is where water quality in a stream has been observed over many years, and a test is made of the null hypothesis that "there is no change in quality between the first and second halves of the data", against the alternative hypothesis that "the quality is poorer in the second half of the record".

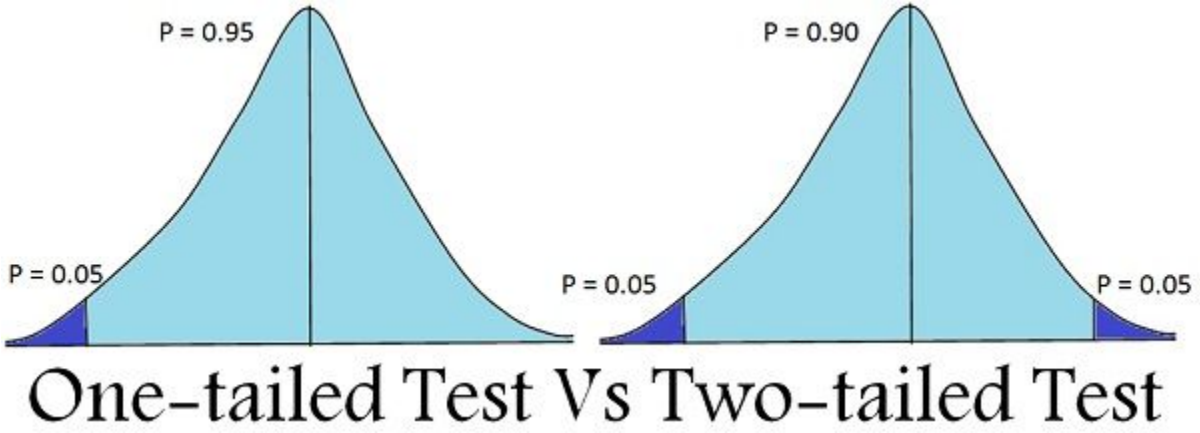
Significance

1. We set a value α beforehand as a desired significance level, the probability of rejecting the null hypothesis, given that it were true (Type I error, or false positive).
2. We find p, the probability of getting a result at least as extreme, given that the null hypothesis were true.
3. The result is statistically significant, by the standards of the study, when $p < \alpha$.

Significance Test

Value varies in one direction

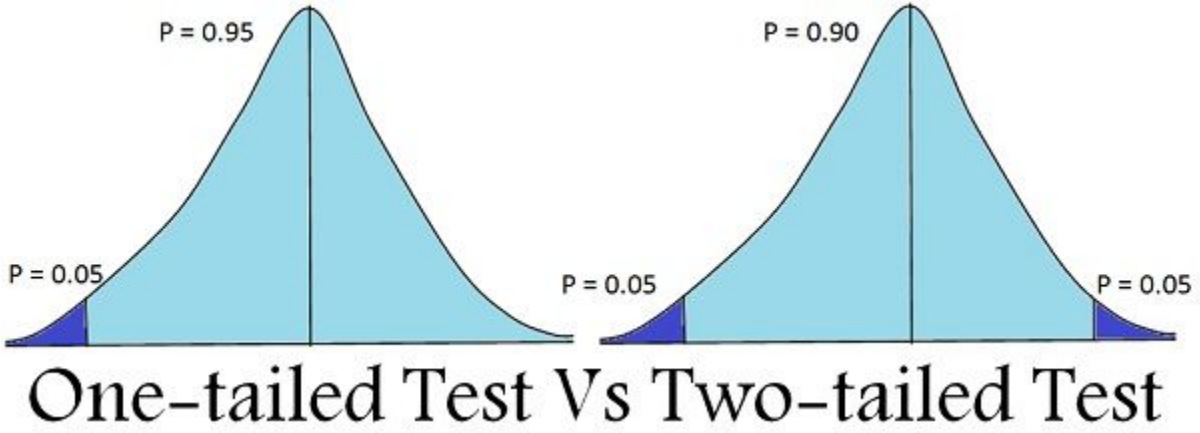
Value varies in two directions



Significance Test Example

Coin biased towards Tails

Coin biased towards Heads Or Tails



Errors

Table - Conclusions in Test of Hypothesis

	Do Not Reject H_0	**Reject H_0**
H_0 is True	Correct Decision	Type I Error
H_0 is False	Type II Error	Correct Decision

Power

The power of a binary hypothesis test is the probability that the test rejects the null hypothesis (H0) when a specific alternative hypothesis (H1) is true.

The statistical power ranges from 0 to 1, and as statistical power increases, the probability β of making a type II error (wrongly accepting the null) decreases.

Hypothesis testing methods

Many different ones depending on

- the type of outcome variable being analyzed
- the number of comparison groups in the investigation
- whether the comparison groups are independent or not

Cookbook:

		Continuous		Discrete
		Parametric	Non-parametric	
One sample		t-test	Wilcoxon test	Chi-square or binomial test
Two samples	Unpaired	t-test	Mann-Whitney test	Fisher's or chi-square test
	Paired	Paired t-test	Wilcoxon test	McNemar's test

Summing Up

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Why Understand Statistics?

Data is Truth, no?

NO! Interpretation and POV matter

Remember:

72.97% of all statistics are made up

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Thank You!

Questions?

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Resources

- http://sphweb.bumc.bu.edu/otlt/MPH-Modules/Menu/
- https://people.cs.umass.edu/~jensen/courses/index.html
- https://callingbullshit.org/index.html
- https://en.wikipedia.org (obviously)
-