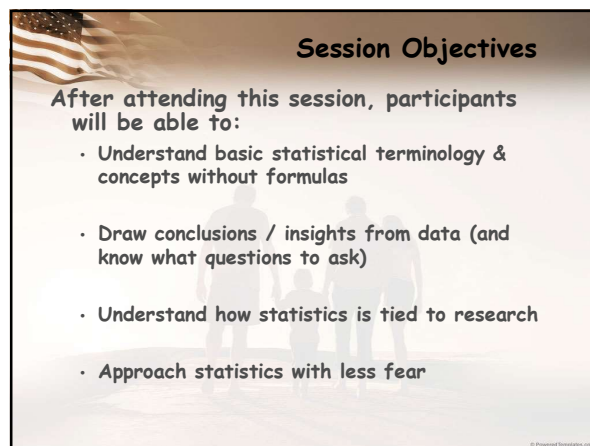


Statistical Thinking

Dr. Jim Mirabella

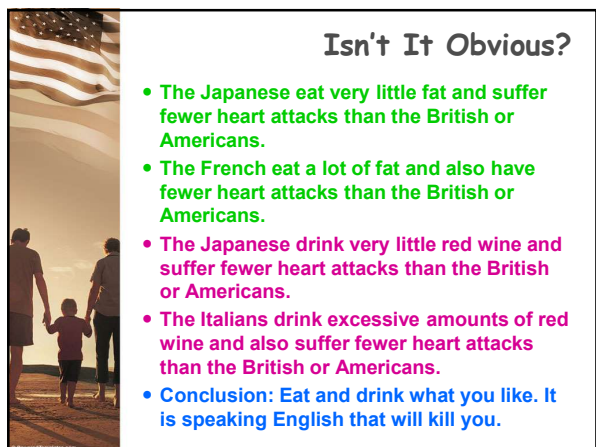
Jacksonville University



Session Objectives

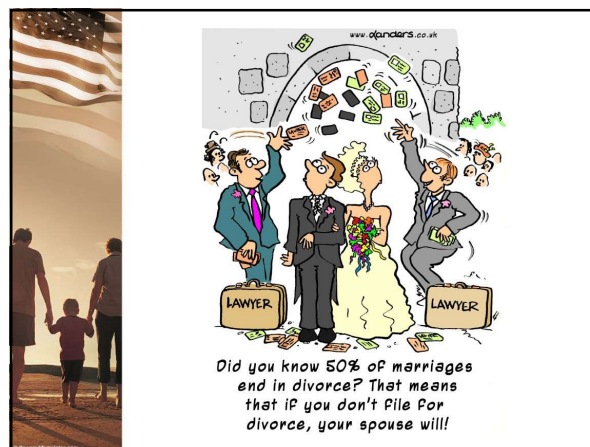
After attending this session, participants will be able to:

- Understand basic statistical terminology & concepts without formulas
- Draw conclusions / insights from data (and know what questions to ask)
- Understand how statistics is tied to research
- Approach statistics with less fear

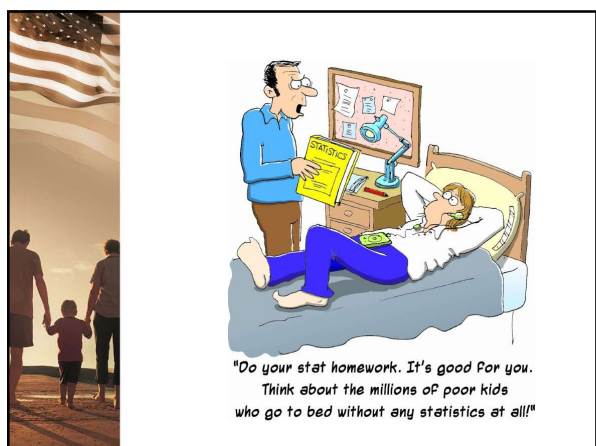


Isn't It Obvious?

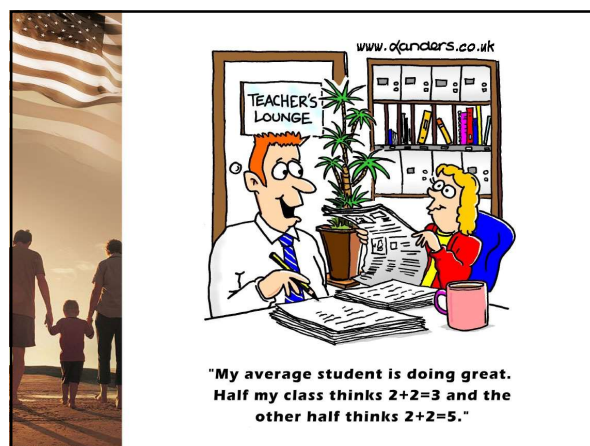
- The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.
- The French eat a lot of fat and also have fewer heart attacks than the British or Americans.
- The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans.
- The Italians drink excessive amounts of red wine and also suffer fewer heart attacks than the British or Americans.
- Conclusion: Eat and drink what you like. It is speaking English that will kill you.



Did you know 50% of marriages end in divorce? That means that if you don't file for divorce, your spouse will!



"Do your stat homework. It's good for you. Think about the millions of poor kids who go to bed without any statistics at all!"



"My average student is doing great. Half my class thinks $2+2=3$ and the other half thinks $2+2=5$."

Learning from Misuse

- It is common to sneer at statistics with: "you can make statistics say anything"
- Only through misuse--by either those presenting or those consuming statistics--is this true
- Many introductory statistics courses focus on how to use statistics rather than how to avoid misusing statistics
- Textbooks offer recipes without advising of the dangers in leaving an ingredient out
- As scholar-practitioners we have an obligation to use statistics as a tool to increase understanding and gain perspective, not to mislead

Learning from Misuse

- Most driving accidents occur near the home. So are you really safer when you leave town - can you unbuckle your seat belt with reduced fear?
- "Healthiest place to live" and other reports.
- Can Exit Polls really predict results accurately?
- If an ad read "Anacin has twice as much pain reliever as Aleve", does it make it twice as strong?
- Sports statistics: a lesson in futility
- In 1992 Presidential election, at one point in the polls, Clinton had 41% of the vote, Bush had 39% and Perot 20%
 - One newspaper reported that Clinton leads the way.
 - Another reported that the majority of voters did not support Clinton.
 - Both were correct but seemingly contradictory.

Learning from Misuse

- Tire shredding cars
- Gun Shows
- Government definitions:
 - Unemployment
 - Homelessness
 - Hunger
 - Living in Poverty
- Healthy Infants
- Arm-wrestling for income

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."

Common Mistakes New Users of Statistics Make

- Failure to use representative data
 - Garbage in / garbage out
 - Is the data representative of the process (time period, quantity)
 - Free of measurement and sampling biases
- Using the wrong tool
- Using the right tool incorrectly
 - Interpreting results is dependent upon valid analysis
- Missing the boat on interpretations
 - Are the results important beyond statistical significance?
 - What is the benefit of the results?

How NOT to Analyze Results

21% of the boys and 30% of the girls support me; therefore I'll get 51% of the vote.

VOTE FOR ME AS PRESIDENT OF THE MATH CLUB

Always Trust Your Sample???

THE MICE SQUAD

I JUST READ YOUR COLUMN IN TODAY'S PAPER. HOW CAN YOU SAY THAT THE NUMBER OF FUNCTIONALLY ILLITERATE IN OUR SOCIETY IS GREATLY EXAGGERATED?!

I DID A SURVEY - AND OF ALL THE PEOPLE WHO COMPLETED AND MAILED BACK THE QUESTIONNAIRE...

... NOT ONE WAS ILLITERATE .

© Proquest.com

Basic Terminology

- **Observation** - A single piece of data
- **Population** - A collection of all possible observations sharing some common set of characteristics
- **Census** - An investigation of all the individual observations making up a population
- **Sample** - A subset or some part of a larger population; a sample can be the entire population
- **Sampling** - The process of using a small number of items from a larger population to draw conclusions about the whole population
- **Parameter** - Computation based on a population
- **Statistic** - Computation based on a sample

© Proquest.com

Why Sample?

Populations have all of the data. If we have the population, we have everything we need. Then why do we take samples?

- Lower cost
- Greater speed of data collection
- Greater accuracy of results
- Availability of population elements
- Destructiveness of observations

© Proquest.com

Sampling Frame

- **Sampling Frame** - The list of elements from which a sample may be drawn
 - A complete and correct list of population members
 - Ideally, the source should be representative of the population
 - The source should not bias the results

QUESTION: Can a phone book be a valid sampling frame?

© Proquest.com

Types of Samples


- **Probability Sample** - A sample in which items are selected on the basis of known probabilities
 - Simple Random Sample
 - Stratified Random Sample
 - Systematic Random Sample
 - Cluster Random Sample
- **Non-probability Sample** - A sample in which items are selected without regard to their probability of occurrence
 - Convenience Sample
 - Judgment Sample
 - Quota Sample
 - Snowball Sample
 - Voluntary Sample

© Proquest.com

Non-probability Sampling

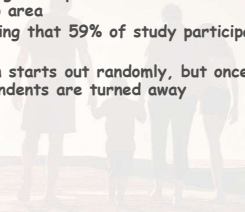
- **Convenience Sample** - A sample of items most readily available
 - Mall surveys
 - Surveys using students
- **Judgment Sample** - A sample selected by an experienced researcher based upon some appropriate characteristic
 - "Market basket" items upon which the CPI is based
 - The Dow Jones Industrial Average

© Proquest.com




Non-probability Sampling

- **Quota Sample** - A sample that ensures that certain characteristics of the population will be represented to the exact intended extent
 - Getting a sample of 100 residents of a specified metro area
 - Deciding that 59% of study participants must be male
 - Often starts out randomly, but once a quota is met, respondents are turned away

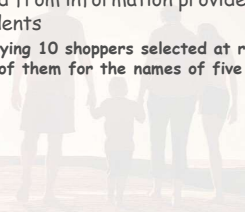


© PoweredTemplate.com




Non-probability Sampling

- **Snowball Sample** - A sample in which initial respondents are selected using probability methods, and then additional respondents are obtained from information provided by the initial respondents
 - Surveying 10 shoppers selected at random and asking each of them for the names of five friends

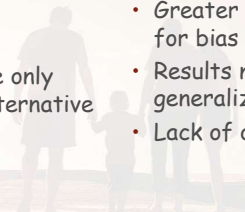


© PoweredTemplate.com




Non-Probability Sampling

<p><u>Advantages</u></p> <ul style="list-style-type: none"> • Lower cost • Less time • May be the only feasible alternative 	<p><u>Disadvantages</u></p> <ul style="list-style-type: none"> • Greater opportunity for bias • Results not generalizable • Lack of objectivity
--	--




© PoweredTemplate.com




Probability Samples

- **Simple Random Sample** - A sampling procedure that ensures that each element of the population has an equal chance of being included in the sample
 - Random drawing
 - Random numbers

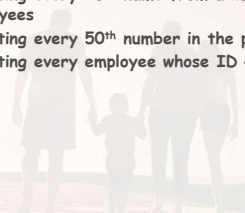


© PoweredTemplate.com




Probability Samples

- **Systematic Sample** - A sample in which every n th number is selected
 - Selecting every 25th name from a list of company employees
 - Selecting every 50th number in the phone book
 - Selecting every employee whose ID # ends in 3

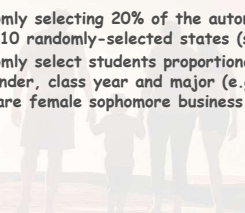


© PoweredTemplate.com



Probability Samples

- **Stratified Sample** - A subsample drawn from samples within different strata that are essential equal on some characteristic
 - Randomly selecting 20% of the automobile dealers from 10 randomly-selected states (strata)
 - Randomly select students proportionally in subgroups by gender, class year and major (e.g., choose X% that are female sophomore business majors)



© PoweredTemplate.com

Probability Samples

- **Cluster Sample** - A sample in which the primary sampling unit is not an individual element in the population but a large cluster of elements
 - Most common type of cluster sample is an "area" sample, in which the primary sampling unit is a geographic area
 - Can also randomly select several small clusters and choose all elements in the clusters

Probability Sampling

Advantages

- Minimization of bias
- Generalizability of results

Disadvantages

- More costly
- More time consuming

Sample Sizes

- A sample does not have to be large to be useful, as long as it's representative
- What is the "right" sample size?
 - > Is it a percentage of the population?
 - > Is population size a factor?
 - > Is there a magic minimum?
- According to Dr. George Gallup:

"You do not need a large sampling proportion to do a good job if you first stir the pot well."

Sample Sizes vs. Error

Sample Sizes vs. Error (95% Confidence)

Sample Size	Margin of Error
100	9.8%
200	6.9%
385	5.0%
500	4.4%
1,000	3.1%
1,500	2.6%
2,000	2.2%

Respecting Ockham's Razor

- Ockham's Razor: *What can be done with less is done in vain with more.*
- Modern statistics is often in need of a shave
- The simplest procedures that can be used to solve a problem are preferred.
- Deliberately complicating solutions is a misuse of statistics -- it obscures the analysis
- **Keep It Statistically Simple And Statistically Sound** (or KISS ASS)

Problem 1: Hangover

- Approach the problem scientifically
 - Identify and state the problem
 - Collect data
 - Determine the root cause of the problem
 - Remove the cause
- Design a data sheet

Date	Input	Output
Monday	Gin & Tonic	Drunk
Tuesday	Vodka & Tonic	Drunk
Wednesday	Rum & Tonic	Drunk

Problem 1: Hangover

- No need for further data
- Common results each day = **drunk**
- Common input each day = **tonic**
- Conclusion: **Eliminate the TONIC!**

Lessons learned :

- Do not blindly follow results of statistical analysis
- If analysis contradicts years of experience, find out why?

Problem 2: Equality

Is this university's admissions process unfair to women?

Schools	Females	Males	Total
Business			
Nursing			
Total	30%	40%	35%

Problem 2: Equality

Is this university's admissions process unfair to women?

→ Not necessarily. Each school treats males and females equally, but... more females apply to the tougher school.

Schools	Females	Males	Total
Business	50/100	100/200	50% for both
Nursing	40/200	20/100	20% for both
Total	30%	40%	35%

→ What happens if they are required to treat genders equally at a university level?

Statistical Toolbox

Operating instructions

- use the right tool for the right job
- misuse of tool leads to disaster
- proper use of tools results in success

The real difficulty with statistical tools is knowing:

- where and when to use each tool
- more importantly, when **not** to use it

It is usually more detrimental to take action based on results of the wrong tool than not to use any tool.

Getting to Know Your Toolbox

Graphical tools (visualize the data)

- Bar charts, pie charts, trend lines, stem & leaf diagrams,...

Descriptive statistics (get the facts behind the picture)

- e.g., Frequencies, central tendency, variability, correlation

Inferential Statistics / Hypothesis testing
(draw inferences about the population)

- ANOVA, Chi-Square, t-tests, Correlation & Regression, Mann-Whitney, Kruskal-Wallis, etc.

Graphical Tools

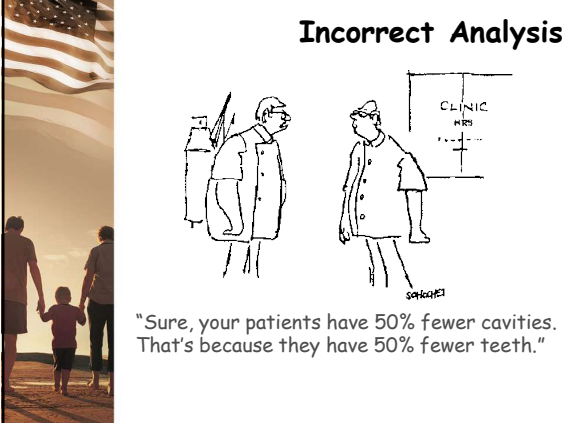
Visualizing & inspecting the data / graphing data appropriately

- qualitative / attribute data → bar charts & pie charts
- quantitative data → histograms & scatterplots
- mixed data → box-whisker plots
- time series data → control charts & trend charts

Looking at Data Intelligently

- What is the source of the data?
 - reports generated by different systems
- Does the data make sense?
 - Does tonic make you drunk?
- Is the information complete?
 - Do I have everything I need to draw a conclusion?
- Is the arithmetic faulty?
 - budget crisis (50% decrease vs. 50% increase)

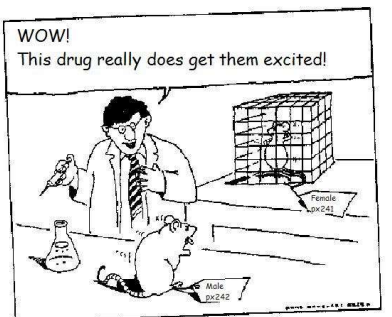
Incorrect Analysis



"Sure, your patients have 50% fewer cavities. That's because they have 50% fewer teeth."







Jumping to Conclusions

WOW!
This drug really does get them excited!

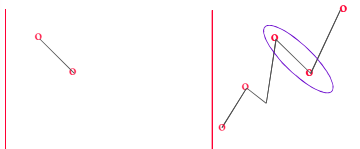


Defining a Trend

- 3 data points are typically used (is it correct?)
- Given any 3 numbers, there are 6 possible patterns

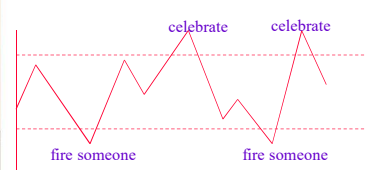
Upward Trend	Downturn	Rebound
		
Setback	Turnaround	Downward Trend
		

Jumping to Conclusions



If you only look at an isolated set of observations, you sometimes lose the big picture!

Jumping to Conclusions



Reacting to data that is out of your control can actually have damaging effects.

Summary Statistics

To truly assess a set of data, you must ask your two questions:

1. **What is the typical observation in the sample?**
 - Measures central tendency / location of the data
 - Mean vs. Median
2. **How spread out is the sample?**
 - Measures dispersion / spread of the data
 - How big is the about?
 - Range vs. Standard deviation vs. Interquartile range


© Powered Templates.com

Summary Statistics

	Uses / Advantages	Disadvantages
Mean	<ul style="list-style-type: none"> • Average • Uses all data in the computation • Use with scale data (height, weight, age, etc.) • Can be used for estimating projected totals 	<ul style="list-style-type: none"> • Sensitive to outliers • Meaningless without dispersion • Should not be used with other data types (e.g., rank-based) • The mean may be an impossible value
Median	<ul style="list-style-type: none"> • Middle observation (half above, half below) • Use with almost any distribution • Tells what a typical value is • Not affected by outliers • The median is an actual observation 	<ul style="list-style-type: none"> • Cannot be used for estimating projected totals (e.g., if you know the median salary for a company, you cannot budget a team of 8 by multiplying the median by 8) • Not used enough / not understood

© Powered Templates.com

Mean vs. Median (You Decide)



"Should we scare the opposition by announcing our mean height, or lull them by announcing our median height?"

© Powered Templates.com

Think About It

What's the catch in each of these?

1. Half of the partners are performing below average. We cannot afford to keep them on our payroll.
2. The CEO of Company X claims that the average salary of his employees has increased 7% in the last year, and yet the total payroll has not increased.
3. Despite our efforts to improve literacy in the past 3 years, half of our children in America are still reading below the median reading level.

© Powered Templates.com

Hypothesis Testing

- What is hypothesis testing?
- Reject vs. not reject (why not accept the null)
- Guilty vs. not guilty
- Are we really proving anything?

© Powered Templates.com

Hypothesis Testing Tools

- **t-Tests**
 - Test if two MEANS differ significantly
 - Test if two PROPORTIONS differ significantly
- **Analysis of Variance (ANOVA)**
 - Test if 3+ MEANS differ significantly
 - Test for interaction among factors
- **Regression Analysis**
 - Test the relationship between 2 scale variables
- **Chi Square Analysis**
 - Test the relationship between 2 nominal variables
- **Non-parametric Tests**
 - Can conduct the equivalent of t-tests, ANOVA or regression analysis when assumptions cannot be met

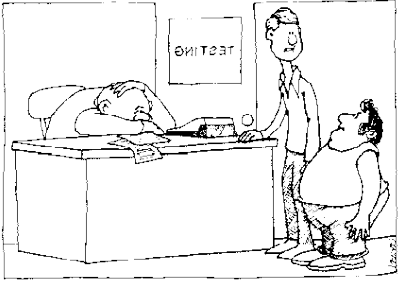
© Powered Templates.com

Why Correlation May Exist between 2 Variables

- **X directly causes Y**
 - more hours studying results in better grades
- **Y directly causes X (i.e., you got it backwards!)**
- **X contributes to changes in Y, but is not the sole cause**
 - exercise helps weight loss, but diet is a major factor too
- **X and Y result from a common cause**
 - spend more money to save more money???
- **Both variables change over time**
 - DJIA vs. # books read
- **Seasonality**
 - ice cream sales vs. toilet flushing
- **Just a coincidence**

Never assume that one variable directly *causes* the other - it is not often the case.

Correlation Mishaps



"He says we've ruined his positive correlation between height and weight."

Think About It

Correlation vs. Causation:

1. Why is there a negative correlation between the total sales of ice cream vs. the # of flu cases? Does ice cream cure the flu?
2. Since the sheriff added more cops on the street, crime has doubled? Do the cops cause the crime?
3. Training a flea to jump...

Pitfalls Throughout the Sampling Hierarchy

Statistical analysis begins with good data.

1. Start with Total Population
2. Select Sampling Frame
 - **Sampling Frame Error** - Certain elements of the population are not included in the sampling frame or unwanted elements are included
 - Using a telephone book to define the sample frame for residents of a particular neighborhood
 - 1936 election: FDR vs. Alf Landon

Pitfalls (continued)

3. Select Sample
 - **Random Sampling Error** - The difference between the result of a sample and the result of a census due solely to observations chosen
 - 75% of a selected sample might be male when only 40% of the population is male
 - Caused by bad luck
 - Caused by sampling bias (i.e., tendency to favor selection of certain data)

Pitfalls (continued)

4. Gather Responses
 - **Non-Response Error** - Errors that cause the sample to be less than representative of the population
 - A disproportionately large group of males responds to a questionnaire
 - Respondents unavailable OR refuse to cooperate
 - Most serious limitation of surveys
 - Don't confuse *response rate* with *sample size*

Sampling Bias: The 1936 Presidential Election



ALF M. LONDON
Republican Nominee for President

*"That leadership along the trail
Which we have loved long since,
And lost awhile,
Has come to us again!"*

Literary Digest predicts Landon win



Franklin Roosevelt
Democratic Nominee for President

Gallup poll predicts Roosevelt win

Sampling Bias: The 1936 Presidential Election

<p>Literary Digest Sample</p> <ul style="list-style-type: none"> • <u>Sample Size:</u> 2,400,000 • <u>Sampling Frame:</u> 10,000,000 <ul style="list-style-type: none"> - Magazine subscribers - Telephone directories - Club and association rosters • <u>Estimated Sampling Error:</u> +/- 0.06% 	<p>Gallup Sample</p> <ul style="list-style-type: none"> • <u>Sample Size:</u> 3,000 • <u>Sampling Frame:</u> N/A <ul style="list-style-type: none"> - Quota sample • <u>Estimated Sampling Error:</u> +/-1.8%
--	---

Sampling Bias: The 1936 Presidential Election

The Outcome ... Roosevelt Wins!

So, What Went Wrong?

What Went Wrong?

<p>Sampling Frame Error</p> <ul style="list-style-type: none"> • Sources of sample lists represented middle- and upper-income voters <ul style="list-style-type: none"> - They had phones - They subscribed to magazines - They belonged to clubs - They tended to vote Republican 	<p>Nonresponse Error</p> <ul style="list-style-type: none"> • Low response rate <ul style="list-style-type: none"> - 10,000,000 million ballots - 2,400,000 responses • Respondents tended to be: <ul style="list-style-type: none"> - Better educated - Higher income - Republican
---	---

Beware of Voluntary Samples

- 900 number surveys
- 800 number surveys
- Text message responses
- "Opinions" site at malls
- News / sports polls on Web
- Talk shows

➤ Voluntary surveys may bring large response totals (*not the same as response rate*), but don't be satisfied with a large sample size. If it is not representative of the population, size will not compensate.

Lessons Learned

- Use statistical tools with care and caution
- Must have data intelligence (collect meaningful data that you understand)
- Analysis is useless without good data
- Graph your data (with the correct tool)
- Central tendency is meaningless without dispersion
- Averages aren't the only statistics game in town
- Watch for false conclusions
- Correlation does not mean causation
- Hypothesis testing doesn't truly prove anything; beware of the conclusions you draw
- **THINK STATISTICALLY and HAVE FUN**