Capella University

Research:
Statistical Thinking

Session Objectives

- After attending this session, learners 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 and the dissertation
  - Approach a statistics course 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.**

The Mind, Heart, & Soul

"Critical thinking, statistical reasoning, and the ability to draw appropriate conclusions based on the data collected and analyses conducted: This is what lies at the core of sound social science research."

The Grand Poobah (date unknown)

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?
- 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?
- Poor measurements leads to poor analysis (e.g., turnover, unemployment, NCAA Football BCS, etc.)
- In 1996 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

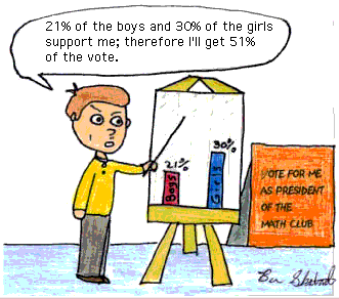
- Tire shredding cars
- Gun Shows
- Government definitions:
  - Unemployment
  - Homelessness
  - Hunger
  - Living in Poverty
- Healthy Infants
- Arm-wrestling for income

There are Lies, and Then…

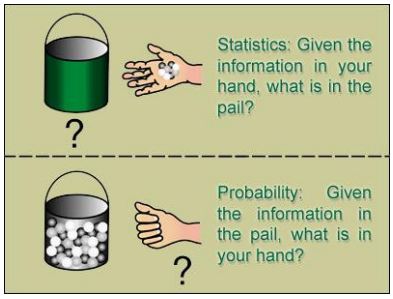
Common Mistakes New Users of Statistics Make

- 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



The Difference between Probability & Statistics



Basic Terminology I

- 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

Why Sample?

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

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

Non-Probability Sampling

Advantages	Disadvantages
• Lower cost	• Greater opportunity for bias
• Less time	• Results not generalizable
• May be the only feasible alternative	• Lack of objectivity

Probability Sampling

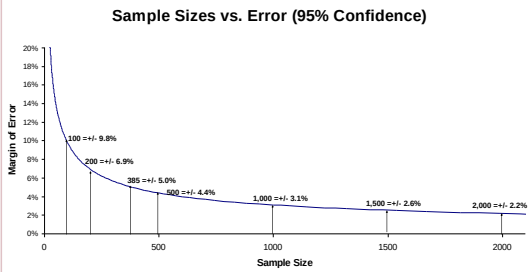
Advantages	Disadvantages
• Minimization of bias	• More costly
• Generalizability of results	• More time consuming

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



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
- **K**eep **I**t **S**tatistically **S**imple **A**nd **S**tatistically **S**ound (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 <u>not</u> 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, etc.

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.

Data Processes

- Operationalize variables
  - Determine level / scale of measurement
- Collect data
- Organize data
- Visually Inspect the data
- Descriptive statistics
- Inferential statistics

Conclusions: Statistical Significance, Theoretical Significance, & Practical Significance

Graphical Tools

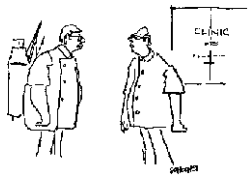
Visualizing & inspecting the data / graphing data appropriately

- 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

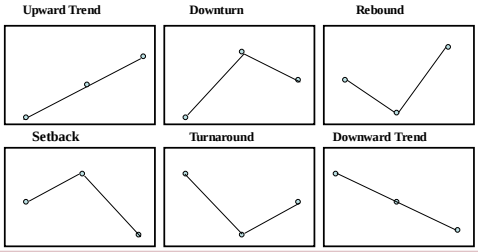


"Sure, your patients have 50% fewer cavities. That's because they have 50% fewer teeth."

Jumping to Conclusions

Defining a Trend

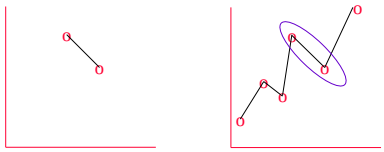
- 3 data points are typically used (is it correct?)
- Given any 3 numbers, there are 6 possible patterns



Defining a Trend

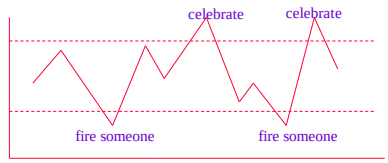
- Notice the typical conclusions that are drawn from the patterns of 3 random data points
- Only 2 of 6 patterns might be beginning of a trend
- In 4 of 6 patterns we jump to conclusions that the trend has shifted
- Declaring 3 points a trend = treating random fluctuation as a special cause of variation
- Typically takes 6 - 7 points to declare a trend with low level of risk
- Something to consider on *quarterly reports* or reports showing *Current Month / Last Month / Year Ago*

Jumping to Graphical Conclusions



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

Jumping to Graphical 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

Summary Statistics

	Uses / Advantages	Disadvantages
Mean	• Average • Uses all data in the computation • Use with scale data (height, weight, age, etc.) • Can be used for estimating projected totals	• 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	• 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	• 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

Mean vs. Median (You Decide)

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

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.

Probability

- Gambler's Fallacy
- Let's Make a Deal
- Birthdays vs. Deathdays – is there a connection?
- False Positives

Probability (Let's Make a Deal)

Pick a door. Suppose you pick #1.

I reveal #3 has a donkey.

Are you better off keeping #1 or switching to #2?

Probability (Let's Make a Deal)

There are 3 possible scenarios. The money is behind 1, 2 or 3.
If you pick 1 and it's behind 1, you win; I will open 2 or 3.
If you pick 1 and it's behind 2, I will open 3.
If you pick 1 and it's behind 3, I will open 2.

In 1 of the 3 scenarios, you have the correct door.

In 2 of the 3 scenarios, you have the wrong door, but there is only one other door.

Thus, there is a 2 / 3 probability of winning if you switch.

False Positives

If a medical test for cancer is 98% accurate and you test positive, what is the probability you have cancer?

a) 98%
b) 50 – 97%
c) Less than 50%

False Positives

The answer is LESS THAN 50% thanks to FALSE POSITIVES.

Assume there are 10,000 people being tested.
Assume that 1% (i.e., 100) actually have cancer.

98 of the 100 with cancer will test + properly.
198 of the 9900 without cancer will test + incorrectly.

Of the 296 people who tested positive,
about 1/3 actually have cancer.

Hypothesis Testing

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

Hypothesis Testing Tools

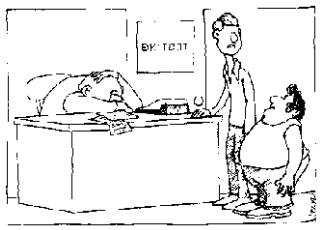
- t-Tests
  - ➢Test if two MEANS differ significantly
  - ➢Test if two PROPORTIONS differ significantly
- Analysis of Variance (ANOVA)
  - ➢Test if 3+ MEANS differ significantly
  - ➢Test for interaction among factors
- Regression Analysis
  - ➢Test the relationship between 2 scale variables
- Chi Square 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

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 an important 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…

Lessons Learned

- Use statistical tools with care and caution
- Must have data intelligence (collect meaningful data that you understand)
- 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