What do Future Senators, Scientists, Social Workers, and Sales Clerks *Need to Learn* from Your Statistics Class?

Professor Erik Erhardt
Department of Mathematics and Statistics
University of New Mexico

April 26, 2017

Adapted by permission:
Professor Jessica Utts
Department of Statistics
University of California, Irvine

Basic Premise

- Most people will take at most one Statistics class in their lives.
- That includes future senators to sales clerks, … as well as presidents, CEOs, jurors, doctors, other decision makers.
- That one class might be yours!
- It's our job to teach them **how to make informed decisions**!

Why Are Students in Your Class?

- STEM majors:
  - To prepare for other courses that use statistics
  - To prepare for the rest of their lives!
- Most students:
  - To fulfill a General Education requirement
  - To prepare for the rest of their lives!

This Reason is Important!

- STEM majors:

  To prepare for the rest of their lives!

- Most students:

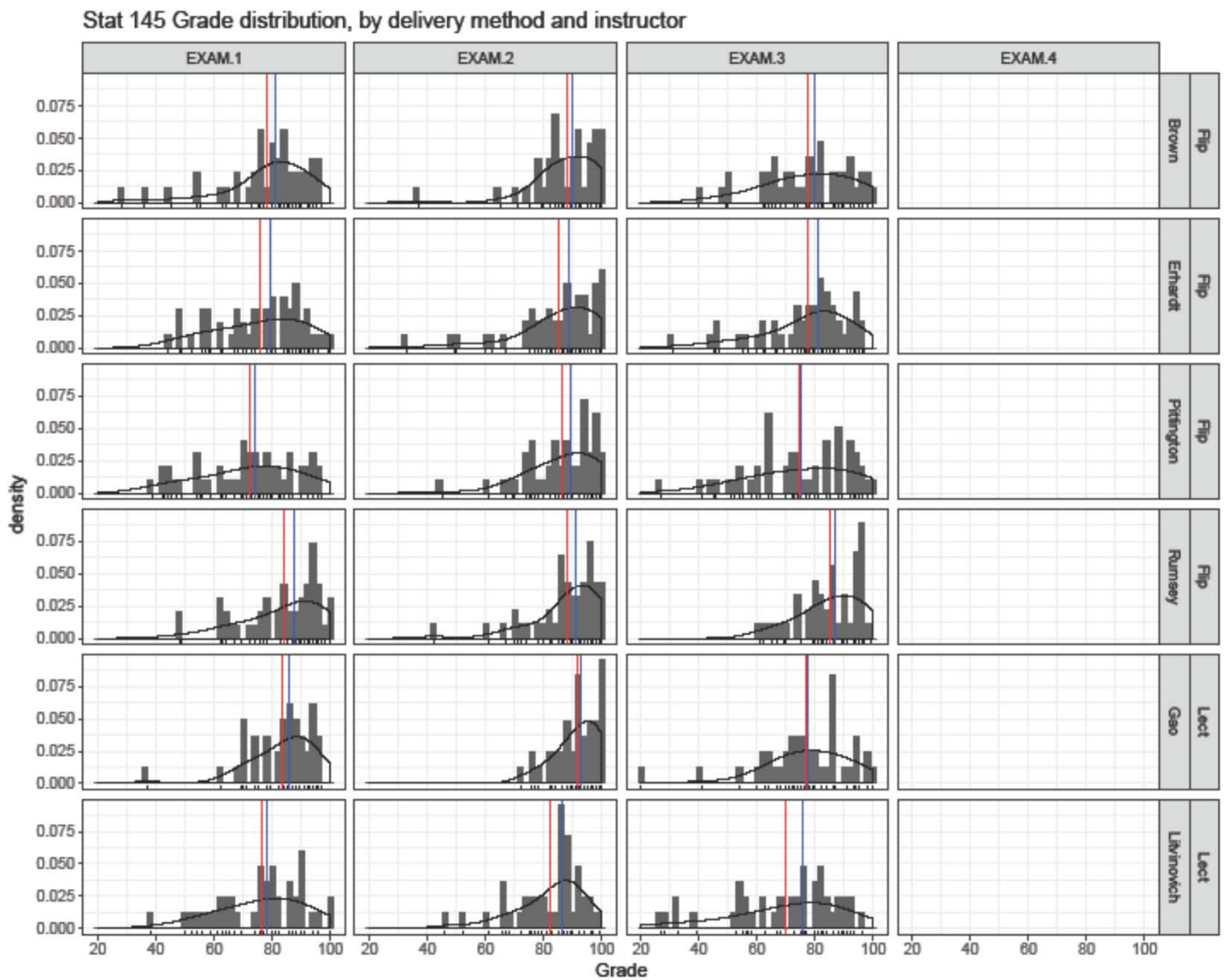
  To prepare for the rest of their lives!

Teaching Fellow intervention

- Teaching Fellow this year
- Coordinated 6 50-student sections of the 25 sections of Stat 145
- 4 hybrid/flipped active-learning classes and 2 experienced TA lecture controls

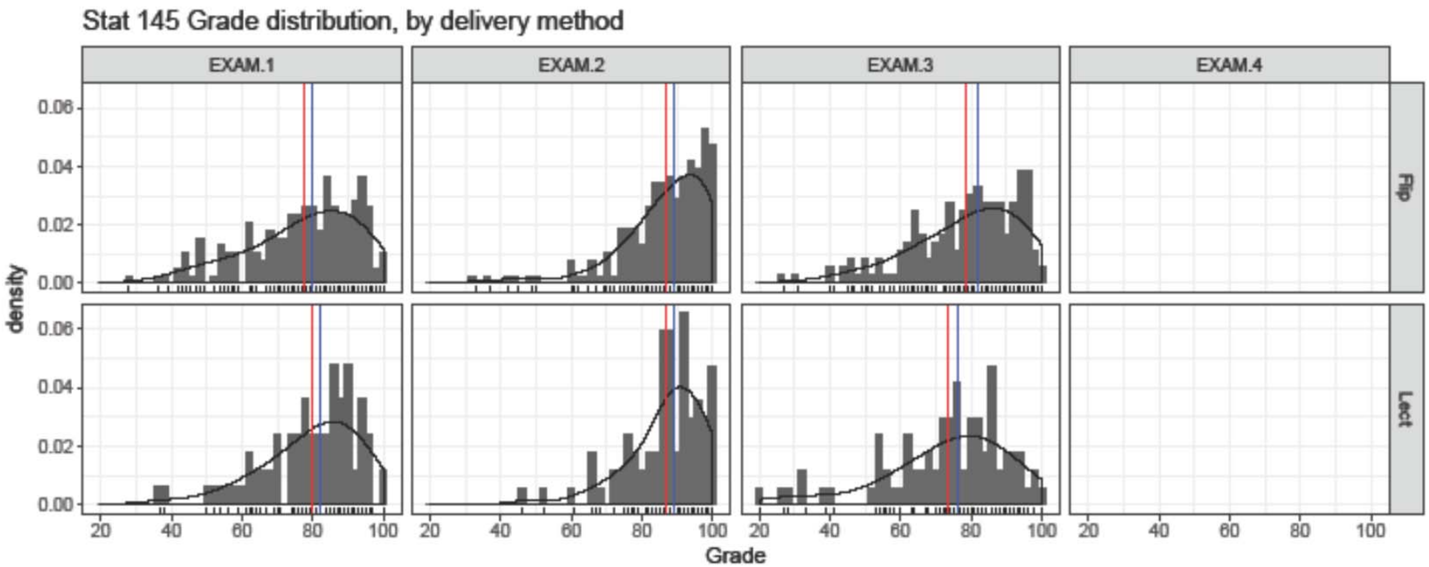
- Does teaching delivery method make a difference?

Exam scores by Method and Instructor



red = mean, blue = median

Exam scores by Method



red = mean, blue = median

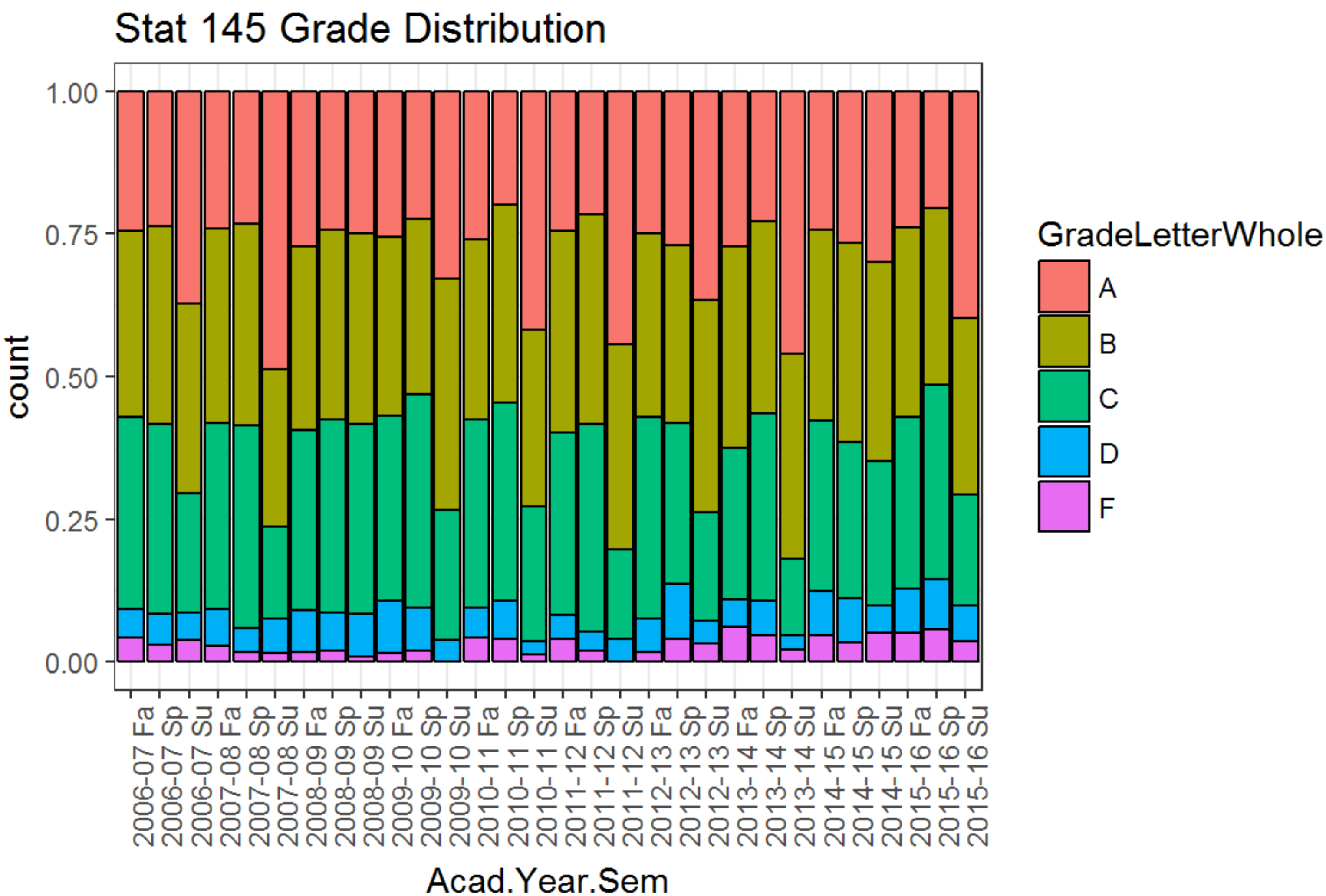
Means

	Delivery	EXAM.1	EXAM.2	EXAM.3	EXAM.4
1	Flip	77.4	86.8	78.6	NaN
2	Lect	79.8	86.9	73.5	NaN

Medians

	Delivery	EXAM.1	EXAM.2	EXAM.3	EXAM.4
1	Flip	80	89	82.0	NA
2	Lect	82	89	76.5	NA

Stat 145 Grade Distribution



Current curriculum of Stat 145

- Summaries, graphical and numerical (1.5 weeks)
- Normal distribution (1.5w)
- EXAM 1
- Correlation and regression (2w)
- Sampling and experiments (2w)
- EXAM 2
- Probability (0.5w)
- Sampling distributions (1w)
- CIs and Hypothesis test basics, known σ (2w)
- EXAM 3
- 1-sample t-test (0.5w)
- 2-sample t-test (1w)
- 1-sample p-test (1w)
- Chi-square test (0.5w)
- Review (1w)
- EXAM 4

Current curriculum of Stat 145

- Summaries, graphical and numerical (1.5 weeks)
- Normal distribution (1.5w)
- EXAM 1
- ~~Correlation and regression (2w)~~
- Sampling and experiments (2w)
- EXAM 2
- Probability (0.5w)
- ~~Sampling distributions (1w)~~
- CIs and Hypothesis test basics, known σ (2w)
- EXAM 3
- ~~1-sample t-test (0.5w)~~
- ~~2-sample t-test (1w)~~
- ~~1-sample p-test (1w)~~
- ~~Chi-square test (0.5w)~~
- Review (1w)
- EXAM 4

NM HED Statistical Competencies

Area II: Statistics Competencies		
Core Competency *Students will:*	**Rationale/Elaboration** *Students should:*	**Suggestions for Assessment**
Construct and analyze graphs and/or data sets.	• Organize data and display in frequency distribution and find percentile points and ranks for the distribution • Graph data distributions using the correct format for graphs, to include: histograms, frequency polygons, box plots and scatter plots and draw appropriate inferences	• Pre/post test • Test/quiz questions • Routine use of an accepted Classroom Assessment Technique (CAT) • Oral presentation by student • Written presentation by student • Student-created portfolio

Note how the Core Competency column reflects mathematical rather than statistical competencies for 3 of the 4 areas.

New Mexico Higher Education Department

http://www.hed.state.nm.us/uploads/files/Policy%20and%20Programs/HED%20Gen%20Ed%20Competencies-All%20Areas.pdf

NM HED Statistical Competencies

Use and solve various kinds of equations.	•Compute mean, median, mode, and standard deviation •Calculate the least squares regression equation and the correlation coefficient •Determine basic probabilities and probabilities associated with the standard normal curve • Understand the binomial distribution and its properties • Compute sampling distributions of sample means •Compute the mean and standard deviation of sample means •Calculate margin of error given sample size and sample size given margin of error •Construct confidence intervals for population means and proportions • Calculate test statistics	• Capstone project • Peer review • Student self-assessment • Group research and presentation on a real-life problem analyzed/solved by using statistics

NM HED Statistical Competencies

Understand and write mathematical explanations using appropriate definitions and symbols.	•Use Z-scores appropriately •Construct probability distributions •Write confidence intervals •Understand the Central Limit Theorem and when to apply it •Write null and alternate hypotheses •Understand the concept of significance level and P values •Apply the steps for inference/hypothesis testing •Describe the basic elements of sampling and experimental design • Define parameters and statistic

NM HED Statistical Competencies

Demonstrate problem solving skills within the context of mathematical applications.	• Determine appropriate methods to display data •Compare measures using Z-scores •Identify and analyze outliers •Use least-square regression equations to predict values •Select appropriate sampling techniques •Determine if random variables are continuous or discrete •Choose and construct appropriate hypothesis tests for population means and proportions	

NM HED Statistical Competencies

- Are these the skills that you use in your daily life?
- Are these the skills that our math-requirement one-and-done Intro Stat students need for their daily lives?
- What is the purpose of Intro Stats?

- It's our job to teach them **how to make informed decisions**!

My Top 10 Important Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Why many studies fail to replicate
4. Sample size and statistical significance
5. Does decreasing risk actually increase risk?
6. Personalized risk
7. Poor intuition about probability/expected value
8. The prevalence of coincidences
9. Surveys and polls
10. Average versus normal

My Top 10 Important Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Why many studies fail to replicate
4. Sample size and statistical significance
5. Does decreasing risk actually increase risk?
6. Personalized risk
7. Poor intuition about probability/expected value
8. The prevalence of coincidences
9. Surveys and polls
10. Average versus normal

My vision of UNM Intro Stats

Three pillars of change:

Two new versions of Stat 145 and TA training.

1. Statistical literacy version covering the Top 10 Important Topics for those students who will take one statistics course in their life and who need basic everyday skills to evaluate information.

My vision of UNM Intro Stats

Three pillars of change:

Two new versions of Stat 145 and TA training.

1. **Statistical literacy version.**
2. Computational version that covers the traditional material, but in a modern evidence-based way by integrating real data in the context of case studies, fostering active learning, and using technology to explore concepts and analyze data.

My vision of UNM Intro Stats

Three pillars of change:

Two new versions of Stat 145 and TA training.

1. **Statistical literacy version.**
2. **Computational version.**
3. We need to train our teaching assistants in the content of Stat 145 and in teaching methods before they are responsible for a class of 50 freshmen.

My vision of UNM Intro Stats

Three pillars of change:

Two new versions of Stat 145 and TA training.

1. **Statistical literacy version.**
2. **Computational version.**
3. **Train our teaching assistants.**

A (Partially True) Story

- Senator Chance, who took statistics from you, sees this (real!) headline:

  "Breakfast Cereals Prevent Overweight in Children"

- The article continues:

"Regularly eating cereal for breakfast is tied to healthy weight for kids, according to a new study that endorses making breakfast cereal accessible to low-income kids to help fight childhood obesity."

Hmm, Senator Chance Thinks…

- Maybe I should introduce the Chance Cereal Bill to make breakfast cereal available to low-income children throughout the United States! They would all lose weight! I would be a hero!
- But Senator Chance remembers some cautions from your class and decides to investigate a bit more.
- What is revealed?

Some Details

- This was an observational study
- 1024 children, 411 with usable data
- Mostly low-income Hispanic children in Austin
- Control group for a larger study on diabetes
- Asked what foods they ate for 3 days, in each of grades 4, 5, 6 (same children for 3 years)
- Study looked at number of days they ate cereal = 0 to 3 each year (Frosted flakes #1!)

More Details: The analysis

- Response variable = BMI percentile each year (BMI = body mass index)
- Explanatory variable = days of eating cereal in each year (0 to 3)
- Did not differentiate between other breakfast or no breakfast.
- Multivariate regression, forced "days of cereal" variable to be linearly related to response
- Also included ("adjusted for") age, sex, ethnicity and some nutritional variables

Uh-oh, Some Problems!
Problem #1: Confounding variables

- Observational study – no cause/effect.
- Obvious possible confounding variable is general quality of nutrition in the home
  - Unhealthy eating for breakfast (non-cereal breakfast or no breakfast), probably unhealthy for other meals too.
- High metabolism could cause low BMI and the need to eat breakfast. Those with high metabolism require more frequent meals.

Senator Chance Knew to Ask:

- Who did the study?
  - Lead author = Vice President of Dairy MAX, a regional dairy council. (Fair disclosure: Study funded by NIH, not Dairy MAX)
- What was the size of the effect?
  - Reduction of just under 2% in BMI percentile for each extra day (up to 3) of consuming cereal (regression coefficient was -1.97)
- So the Chance Cereal Bill died before it left Senator Chance's desk!

Who Else Needs to Know How to Evaluate This Study?

- Scientist – understand how to conduct study and report results.
- Social worker – if the program had been mandated for low income kids, how important is compliance?
- Sales clerk – does it matter if her/his kids eat cereal for breakfast?
- In other words, everyone!

More of my Favorite Headlines

- “6 cups a day? Coffee lovers less likely to die, study finds”
- “Oranges, grapefruits lower women's stroke risk”
- “Yogurt Reduces High Blood Pressure, says a New Study”
- “Walk faster and you just might live longer”
  - “Researchers find that walking speed can help predict longevity”
  - “The numbers were especially accurate for those older than 75”

Assessing possible causation

Some features that make causation *plausible* even with observational studies:

- There is a reasonable explanation for how the cause and effect would work.
- The association is consistent across a variety of studies, with varying conditions.
- Potential confounding variables are measured and ruled out as explanations.
- There is a “dose-response” relationship.

Another Story (also partially true)

- Mr. Rossman is a sales clerk
- At the Elite Togs Shop (ETS) in San Luis Obispo, California
- They specialize in Hawaiian shirts
- And **M**ens **Q**uirky **C**lothing
- Mr. Rossman has 3 daughters
- He would like to have a son
- So he asks his wife if she would please eat cereal for breakfast. Not because she's fat…

More about Cereal: Does it Produce Boys?

- Headline in *New Scientist*: "Breakfast cereal boosts chances of conceiving boys" Numerous other media stories of this study.
- Study in *Proc. of Royal Soc. B* showed of pregnant women who ate cereal, 59% had boys, of women who didn't, 43% had boys.
- Problem #1 revisited:

  Headline implies eating cereal *causes* change in probability, but this was an observational study. (Confounding variables???)

Problem #2: Multiple Testing

- The study investigated 132 foods the women ate, at 2 time periods for each food = 264 possible tests!
- By chance alone, *some* food would show a difference in birth rates for boys and girls.
- Main issue: Selective reporting of results when many relationships are examined, not adjusted for multiple testing. Quite likely that there are "false positive" results.

Common Multiple Testing Situations

- *Genomics*: “Needle in haystack” – looking for genes related to specific disease, testing many thousands.
- *Diet and disease*: For instance, ask cancer patients and controls about many different dietary habits.
- *Interventions* (*e.g. Abecedarian Project**): Randomized study put low-income infants into a kindergarten kids educational program (or not). Kids in program were almost 4 times as likely to graduate from college. (Many other differences; too many to all be multiple testing.)

Multiple Testing: What to do?

- There are statistical methods for handling multiple testing. See if the research report mentions that they were used.
- See if you can figure out how many different relationships were examined.
- If <u>many</u> significant findings are reported (relative to those studied), it's <u>less likely</u> that the significant findings are false positives.

Yet Another Story

- There is planet similar to earth, Planet PV, where *p*-values reign supreme.
- On that planet, babies are only allowed to be born in the spring.
- No one knows about the beneficial effects of taking aspirin to prevent heart attacks.
- Lots of other false notions from statistical studies (even more than here!).

On Planet PV, They Read This Headline

Spring Birthday Confers Height Advantage

Austrian study of heights of 507,125 military recruits.

- Results were highly statistically significant (tiny p-value), test of difference in means for men born in spring versus fall
- Men born in spring were, on average, about 0.6 cm taller than men born in fall, i.e. about 1/4 inch (Weber et al., *Nature,* 1998, 391:754–755).
- **Sample size so large** that **even a very small difference** was ***highly statistically significant.***

Does Aspirin Prevent Heart Attacks?

Physicians' Health Study (1988)

5-year randomized experiment

22,071 male physicians (40 to 84 years old).

$\chi^2 = 25.4$, p-value ≈ 0

Condition	Heart Attack	No Heart Attack	Attacks per 1000
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

But on Planet PV, $n = 2207$ instead, same rates

So $\chi^2 = 2.54$, p-value = .111, not significant!

Problem #3:
Role of sample size in statistical significance

- The *p*-value does ***not*** provide information about the ***magnitude/importance*** of the effect.
- If sample size **large** enough, almost **any null hypothesis can be rejected**.
- If the sample size is **too small** it is very hard to achieve statistical significance (low power)
- Don’t equate statistical significance with whether or not there is a real, important effect.
- If possible, get a confidence interval.

Problem #4:
Avoiding Risk May Put You in Danger

- In 1995, UK Committee on Safety of Medicines issued warning that new oral contraceptive pills "increased the risk of potentially life-threatening blood clots in the legs or lungs by twofold – that is, by 100%" over the old pills
- Letters to 190,000 medical practitioners; emergency announcement to the media
- Many women stopped taking pills.

Clearly there is increased risk, so what's the problem with women stopping pills?

Probable consequences:

- Increase of 13,000 abortions the following year
- Similar increase in births, especially large for teens
- Additional $70 million cost to National Health Service for abortions alone
- Additional deaths and complications probably *far exceeded* pill risk.

Actual Risk versus Relative Risk

- "Twofold" risk of blood clots:
  - 1/7000 to 2/7000, not a big change in *absolute* risk, and still a *small risk*.
- *Absolute* risk is what is important:
  - 2/7000 likely to have a blood clot
  - Compare to other risks of pregnancy
- But *Relative* risk (2 in this case) is what makes news!

Reported Risk versus Your Risk

"Older cars stolen more often than new ones"
Davis (CA) Enterprise, 15 April 1994, p. C3

- Of the 20 most popular auto models stolen in California the previous year, 17 were at least 10 years old.
- Many factors determine which cars stolen:
  - Type of neighborhood.
  - Locked garages.
  - Cars not locked and/or don't have alarms.
- *If I were to buy a new car, would my risk of having it stolen increase or decrease over my old car?*
- Article gives no information about that question.

Considerations about Risk

- Changing a behavior based on relative risk may *increase* overall risk of a problem. Trade-offs!
- Find out what the *absolute* risk is, and consider relative risk in terms of additional *number* at risk

  Example: Suppose a behavior doubles risk of cancer

  Brain tumor: About 7 in 100,000 new cases per year, so adds about 7 cases per 100,000 per year.

  Lung cancer: About 75 in 100,000 new cases per year, so adds 75 per 100,000, more than 10 times as many!
- Does the reported risk apply to you?
- Over what time period? (Risk per year? Per lifetime?)

Problem #5: Poor intuition about probability, chance, and expected value

- William James was first to suggest that we have an *intuitive* mind and an *analytical* mind, and that they process information differently.
- Example: People feel safer driving than flying, when probability suggests otherwise.
- Psychologists have studied many ways in which we have poor intuition about probability assessments.

Example: Confusion of the Inverse

Gigerenzer gave 160 gynecologists this scenario:

- About 1% of the women who come to you for mammograms have breast cancer (bc)
- If a woman has bc, 90% chance of positive test
- If she does not have bc, there is only a 9% chance of positive test (false positive)

A woman tests positive. What should you tell her about the chances that she has breast cancer?

Answer choices: Which is best?

- The probability that she has breast cancer is about 81%.
- Out of 10 women with a positive mammogram, about 9 have breast cancer.
- Out of 10 women with a positive mammogram, about 1 has breast cancer.
- The probability that she has breast cancer is about 1%.

Answer choices and % who chose them

- The probability that she has breast cancer is about 81%.” 13% chose this
- Out of 10 women with a positive mammogram, about 9 have breast cancer. [i.e. 90% have it] **47%** chose this
- Out of 10 women with a positive mammogram, about 1 has breast cancer. [i.e. 10% have it] 21% chose this
- The probability that she has breast cancer is about 1%. 19% chose this

What is the Correct Answer?

Let's look at a hypothetical 100,000 women.

Only 1% have cancer, 99% do not.

	Test positive	Test negative	Total
Cancer			1,000 (1%)
No cancer			99,000
Total			100,000

Let’s see how many test positive

90% who have cancer test positive.

9% of those who don’t have it test positive.

	Test positive	Test negative	Total
Cancer	900 (90%)		1,000
No cancer	8910 (9%)		99,000
Total	9810		100,000

Let’s complete the table for 100,000 women:

	Test positive	Test negative	Total
Cancer	**900**	100	1,000
No cancer	8910	90,090	99,000
Total	**9810**	90,190	100,000

Correct answer is 900/9810, just under 10%!

Physicians confused two probabilities:

P(positive test | cancer) = .9 or 90%

P(cancer | positive test) = 900/9810 = .092 or 9.2%

Confusion of the inverse: DNA Example

- DAN is accused of crime because his DNA matches DNA at a crime scene (found through database of DNA). Only 1 in a million people have this specific DNA. Is Dan surely guilty??
- Suppose there are 6 million people in the local area, so about 6 have this DNA. Only one is guilty!

Then:

- P(DNA match | innocent) ≈ only 5 out of 6 million, very low! (*Prosecutor would emphasize this*)
- But... P(innocent | DNA match) ≈ 5 out of 6, very high! (*Defense lawyer should emphasize this)*
- Jury needs to understand this difference!

The Conjunction Fallacy: Survey Question

Plous (1993) presented readers with the following test:

Place a check mark beside the alternative that **seems most likely to occur within the next 10 years**:

- An all-out nuclear war between the United States and Russia
- An all-out nuclear war between the United States and Russia in which neither country intends to use nuclear weapons, but both sides are drawn into the conflict by the actions of a country such as Iraq, Libya, Israel, or Pakistan.

Survey in my class: Using your intuition, pick the more likely event at that time.

44/138 = **32%** chose first option – CORRECT!

94/138 = **68%** chose second option – Incorrect!

The Representativeness Heuristic and the Conjunction Fallacy

- **Representativeness heuristic:** People assign higher probabilities than warranted to scenarios that are *representative* of how we *imagine* things would happen.
- This leads to the **conjunction fallacy** … when detailed scenarios involving the conjunction of events are given, people assign *higher* probability assessments to the *combined event* than to statements of one of the simple events alone.
- Remember that *P(A* <u>and</u> *B) = can't* <u>*exceed*</u> *P*(*A*)

Other Probability Distortions

- Coincidences have higher probability than people think, because there are so many of us and so many ways they can occur.
- Low risk, scary events in the news are perceived to have *higher* probability than they have (readily brought to mind).
- High risk events where we think we have control are perceived to have *lower* probability than they have.
- People place less credence on data that conflict with their beliefs than on data that support them.

Understanding Expected Value: Survey Question (my class)

Which one would you choose in each set? (Choose either A or B and either C or D.)

A. A gift of $240, guaranteed
B. A 25% chance to win $1000 and a 75% chance of getting nothing.

C. A sure loss of $740
D. A 75% chance to lose $1000 and a 25% chance to lose nothing

Survey Question Results

Which one would you choose in each set?
(Choose either A or B and either C or D.)

85%	**A.** A gift of $240, guaranteed
15%	**B.** A 25% chance to win $1000 and a 75% chance of getting nothing.

30%	**C.** A sure loss of $740
70%	**D.** A 75% chance to lose $1000 and a 25% chance to lose nothing

The Amount Makes a Big Difference

Which one would you choose in each set?

A. A gift of $5, guaranteed

B. A 1/1000 chance to win $4000

Now 75% chose B.

This is like buying lottery tickets.

C. A sure loss of $5

D. A 1/1000 chance of losing $4000

Now 80% chose C.

Like buying insurance or extended warranty.

Probability and Intuition Lessons

Examples of Consequences in daily life:

- Assessing probability when on a jury

  Lawyers provide detailed scenarios – people give higher probabilities, even though *less* likely.

- Extended warranties and other insurance

  "Expected value" favors the seller

- Gambling and lotteries

  Again, average "gain" per ticket is negative

- Poor decisions (e.g. driving versus flying)

Summary: What Future "Everyones" Need from Your Class!

1. Don't make cause/effect conclusions based on observational studies. (Understand confounding.)
2. Watch out for "multiple testing."
3. Don't confuse statistical and practical significance. Find out the size of the effect.
4. Consider *absolute risk* instead of *relative risk.*
5. Think carefully about probability, chance, and expected values.

My vision of UNM Intro Stats

Three pillars of change:

Two new versions of Stat 145 and TA training.

1. **Statistical literacy version.**
2. **Computational version.**
3. **Train our teaching assistants.**