Midterm 1 Review

Material up to Confidence Intervals

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

Announcments

- Lab 3 is due tomorrow at 12:45 pm

Midterm 1

Exam Details:

- When: Tomorrow, Thursday May 31
- What you can use:
  - Scientific calculator
  - Cheat sheet (8.5 x 11)
- A link to a probability table will be provided

Exam Format

- Sakai

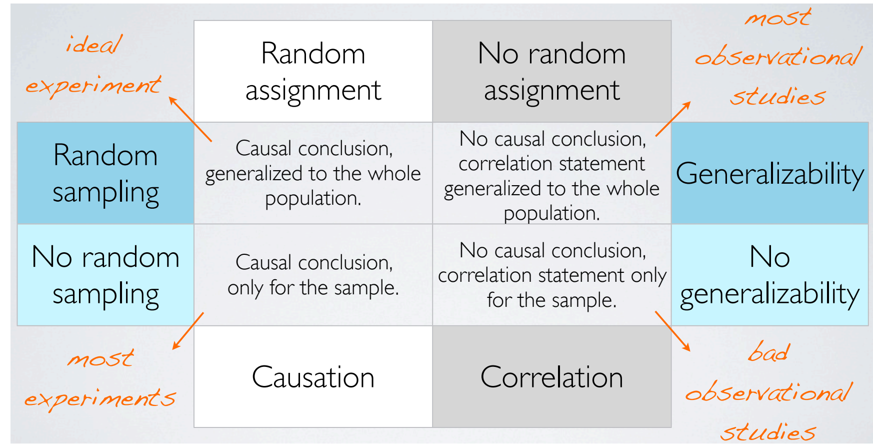
Exam Format

- Covers everything up to and including yesterday's lecture
- There are written, multiple choice, and rue/false questions: Very similar to the format of the practice midterm
  1. First 3 'open' questions have multiple parts, including some multiple choice (46 points total)
  2. 5 true/false questions (1 point each)
  3. 8 multiple choice questions (2 points each)
- Sakai quiz: Many of the the problems will require scratch work. When submitting answers, include the relevant finalized formula and then the final answer.

Key Terms

- Population
- Parameter
- Statistic
- Simple Random Sample
- Stratified Sample
- Cluster Sample
- Multistage Sample
- Experiment
- Observational Study
- Control
- Placebo
- Confounding Variable
- Sampling Distribution

Data Collection, Observational Studies & Experiments



Exploratory Data Analysis

Describing Distributions of Numerical Variables:

- *Shape*: skewness, modality
- *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
  - Notation: μ: population mean, $\bar{x}$: sample mean
- *Spread*: measure of variability in the distribution (standard deviation, IQR, range, etc.)
- *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

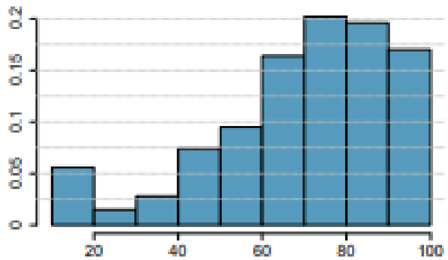
Exploratory Data Analysis

Robust statistics

- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- Therefore we choose median & IQR (over mean & SD) when describing skewed distributions.

Clicker question

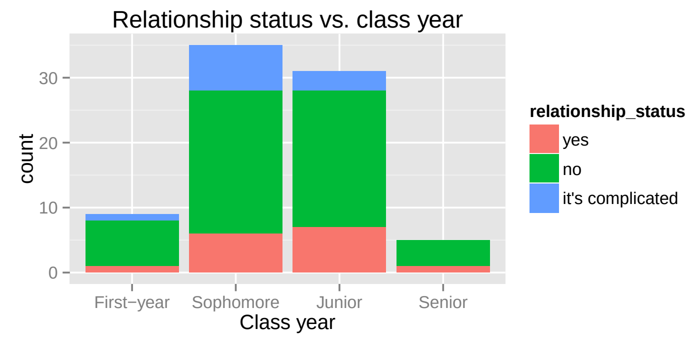
Which of the following is **false**?



- (a) The box plot would have outliers on the lower end.
- (b) The median is between 70 and 80
- (c) More than 25% of the data is above 90.
- (d) More thn 50% of the data have positive Z-scores
- (e) The mean is likely to be smaller than the median.

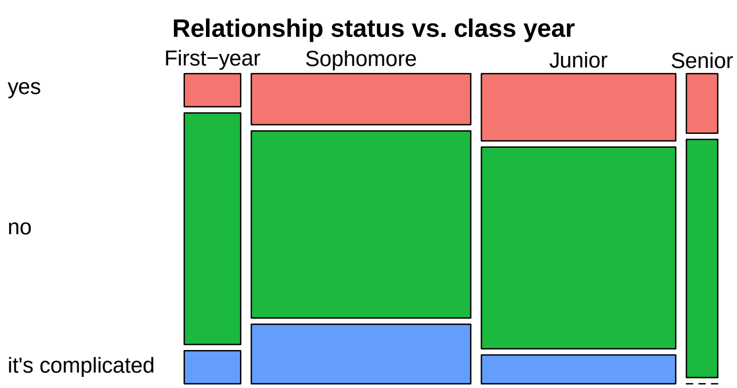
Use segmented bar plots for visualized relationships between 2 categorical variables

What do the heights of the segments represent? Is there a relationship between class year and relationship status? What descriptive statistics can we use to summarize these data? Do the widths of the bars represent anything?



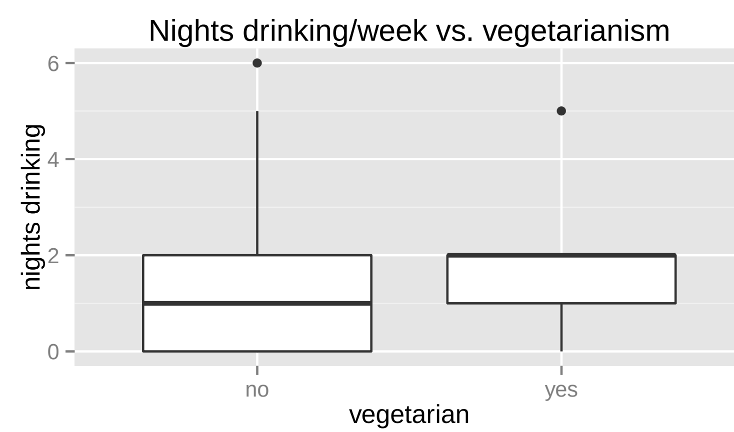
... or use a mosaic plot

What do the widths of the bars represent? What about the heights of the boxes? Is there a relationship between class year and relationship status? What other tools could we use to summarize these data?



2. Use side-by-side box plots to visualize relationships between a numerical and categorical variable

How do drinking habits of vegetarian vs. non-vegetarian students compare?



Key Ideas:

- Observed differences may be due to random chance
- Test whether difference is significant using simulations

CPR is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. The chest compressions involved with this procedure can also cause internal injuries. Blood thinners that are often given to help release a clot that is causing the heart attack may also negatively affect such internal injuries. An experiment was designed to evaluate if blood thinners have an impact on survival after a heart attack. Patients were randomly divided into a treatment group (received a blood thinner) or the control group (no blood thinner). The outcome variable of interest was whether the patients survived for at least 24 hours.

For hypotheses for this study in plain and statistical language. Let p_c represent the true survival population in the control group and p_r represent the survival proportion for the treatment group

- H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group.
  - $p_t - p_c = 0$
- H_A: Blood thinners do have an impact on survival.
  - $p_t - p_c \neq 0$

Clicker question

Given these hypotheses, what is the sample statistic?

$$H_0 : p_t - p_c = 0 \quad H_A : p_t - p_c \neq 0$$

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

(a) (11/25) - (39/65) = -0.16

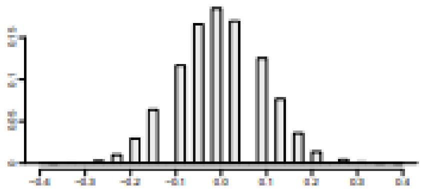
(b) (14/40) - (11/50) = 0.13

(c) (14/90) - (11/90) = 0.033

(d) (40/90) - (50/90) = -0.111

Clicker question

A randomization test was conducted to evaluate these hypotheses.
Based on the randomization distribution below, what is the conclusion?



These data

- (a) provide convincing evidence that blood thinners
- (b) provide convincing evidence that blood thinners do not
- (c) do not provide convincing evidence that blood thinners
- (d) do not provide convincing evidence that blood thinners do not

have an impact on survival.

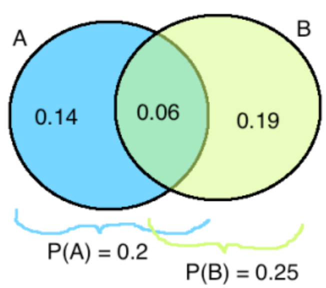
Probability and Conditional Probability

- *Disjoint (mutually exclusive) events* cannot happen at the same time.
  - For disjoint A and B: $P(A \text{ and } B) = 0$
- If A and B are *independent events*, having information on A does not tell us anything about B (and vice versa)
  - If A and B are independent:
    - $P(A|B) = P(A)$
    - $P(A \text{and} B) = P(A) \times P(B)$
- *General addition rule:* $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- *Bayes' theorem:* $P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

Clicker question

Which of the following is true?

- (a) A and B are Independent.
- (b) P(A but not B) = 0.2
- (c) P(A | B) = 0.06 / 0.14
- (d) P(A or B) = 0.14 + 0.19 + .06
- (e) P(neither A nor B) = 1 - 0.06



Bayes' Theorem and Bayesian Inference

- Probability trees are useful for organizing information in conditional probability calculations
- They're especially useful in cases where you know P(A | B) along with some other information, and you are asked for P(B | A)
- Using Bayes' Theorem:

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis and data})}{\text{data}} = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

- Two types of probability distributions: discrete and continuous
- Normal distribution is unimodal, symmetric, and follows the 68-95-99.7 rule
- Z scores serve as a ruler for any distribution

$$Z = \frac{obs - mean}{SD}$$

- Z score: number of standard deviations the observation falls above or below the mean

- The *Binomial distribution* describes the probability of having exactly k successes in n independent trials with probability of success p.

$$P(k \text{ successes in n trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

- *Expected Value*: np
- *Standard Deviation:* $\sqrt{np(1-p)}$
- Shape of the binomial distribution approaches normal when the S-F rule is met

Clicker question

Which of the following probabilities should be calculated using the Binomial Distribution?

Probability that

(a) a basketball players misses 3 times in 5 shots
(b) train arrives on the time on the third day for the first time
(c) height of a randomly chosen 5 year old is greater than 4 feet
(d) a randomly chosen individual likes chocolate ice cream best

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shtos, and makes the last two. The probability of this scenario is:

$$0.4^3 0.6^2 \approx$$

- But this isn't the only possible scenario

1. *MMM*HH
2. *MM*H*M*H
3. *M*H*MM*H
4. H*MMM*H
5. H *MM*H*M*
6. H*M*H*MM*
7. HH*MMM*
8. *M*H*M*H*M*
9. *M*HH*MM*
10. *MM*HH*M*

- Each one of these scenarios has 3 *M*'s and 2 H's, therefore the probability of each scenario is 0.023.
- Then, the total probability is $10 \times 0.23 = 0.23$

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$\begin{aligned} P(k \text{ successes in n trials}) &= \binom{5}{3} p^3 (1-p)^{5-3} \\ &= 10 \times 0.023 \\ &= 0.23 \end{aligned}$$

Clicker question

Which of the following highlights the correct outcomes for "at most 3 misses in 5 shots?"

a) {*0*,*1*,*2*,3,4,5}
b) {*0*,*1*,*2*,*3*,4,5}
c) {0,1,2,*3*,*4*,*5*}
d) {0,1,2,*3*,4,5}
e) {0,1,2,*3*,*4*,5}

Clicker question

Which of the following is the correct calculation for "P(at most 3 misses in 5 shots)?"
Note: P(k) means P(k misses in 5 shots), calculated using the binomial distribution

a) P(0) + P(1) + P(2)
b) P(3) + P(4) + P(5)
c) 1 - P(0)
d) 1 - [P(0) + P(1) + P(2)]
e) 1 - [P(4) + P(5)]

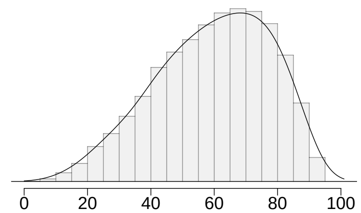
- Sample statistics vary from sample to sample
- CLT describes the shape, center, and spread of sampling distributions

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

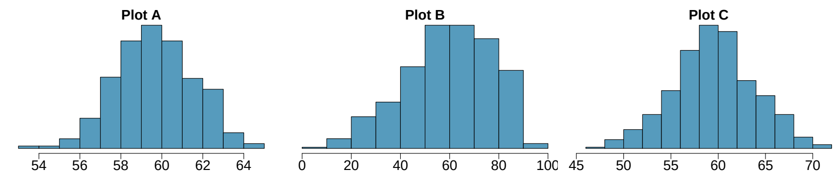
- CLT only applies when independence and sample size/skew conditions are met.

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.



(a) (2) - B; (3) - A; (4) - C
(b) (2) - A; (3) - B; (4) - C
(c) (2) - C; (3) - A; (4) - D
(d) (2) - B; (3) - C; (4) - A



A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

(a) yes
(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 = 0.0049 \end{aligned}$$

- Statistical inference methods based on the CLT require the same conditions as the CLT
- *CI*: point estimate ± margin of error
- Calculate the sample size a priori to achieve desired margin of error

Solve for n:

$$ME = z^* \frac{s}{\sqrt{n}}$$