Practice Exam #2 (Sta-209, S24)

Ryan Miller

The following information will appear verbatim on the first page of Exam 2. You do not need to memorize this information, but you should be familiar with it.

Directions

- Answer each question using *no more than specified number of sentences* and not attempt to avoid these guidelines by using run-on sentences. Answers that are unnecessarily verbose may result in point loss.
- Do not include superfluous information in your answers, you may be penalized if you make an inaccurate statement even if you go on to provide a correct answer. Your answers should be clear, concise, and include only what is needed to answer the question that was asked.

Formula Sheet

Definitions:

- **Risk**: relative frequency of an event/outcome
- **Relative Risk**: ratio of the risks across two groups
- **Odds**: ratio of how often an event/outcome is observed relative to how often it is not observed
- **Odds ratio**: ratio of odds across two groups

Formulas:

$$\bar{x} = \tfrac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$r = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(\tfrac{x_i-\bar{x}}{s_x}\right)\left(\tfrac{y_i-\bar{y}}{s_y}\right)$$

Statistic	Standard Error	Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1, 2\}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$

Question #1 (conceptual questions)

Part A: In 1-2 sentences, explain the meaning of "95% confidence" in the statistical term "95% confidence interval". Avoid using the words "confident" or "confidence" as core components of your explanation.

Part B: Suppose your friend wants to use a statistical significance threshold of $\alpha = 0.01$ in a hypothesis test instead of the "conventional" threshold of $\alpha = 0.05$. In 1-2 sentences, briefly explain the trade-offs involved using your friend's proposed threshold (relative to the conventional threshold) in terms of Type 1 and Type 2 errors.

Part C: Consider a genomics experiment that involves 12,000 hypothesis tests each corresponding to a different genetic marker. Briefly explain why using a statistical significance threshold of $\alpha = 0.05$ could be problematic in this experiment.

Part D: Briefly explain the difference between a *sample* and a *population*. Why is this distinction important when describing the relationships found when analyzing sample data? Limit your response to at most 3 sentences.

Question #2 (confidence intervals)

Earlier in the semester you encountered the American Community Survey data, a random sample of $n = 1287$ United States residents that is administered by the US Census Bureau on a rolling basis.

Shown below is some `R` output involving data. You may need to use some, all, or none of this output in the questions that follow.

```
acs = read.csv("https://remiller1450.github.io/data/EmployedACS.csv")

acs %>% group_by(Race) %>% summarize(Mean_Income = mean(Income),
                                     SD_Income = sd(Income),
                                     Mean_Age = mean(Age),
                                     SD_Age = sd(Age),
                                     n_HealthInsurance = sum(HealthInsurance == 1),
                                     n_Race = n())
```

```
## # A tibble: 4 x 7
##   Race  Mean_Income SD_Income Mean_Age SD_Age n_HealthInsurance n_Race
##   <chr>       <dbl>     <dbl>    <dbl>  <dbl>             <int>  <int>
## 1 asian        61.2      77.5     41.2   14.1                86     92
## 2 black        34.5      29.9     43.5   14.2               106    116
## 3 other        31.2      42.4     38.9   13.6                80    102
## 4 white        45.5      55.5     43.7   15.4               907    977
```

```
cor(acs$Age, acs$Income)
```

```
## [1] 0.1649631
```

```
cor_results = cor.test(acs$Age, acs$Income, conf.level = 0.99)
cor_results$conf.int
```

```
## [1] 0.09431882 0.23395441
## attr(,"conf.level")
## [1] 0.99
```

Part A: Calculate a 95% confidence interval estimate for the difference in the proportions of white individuals and black individuals with health insurance in the United States. Show all of your work. You should use the value $c = 1.96$ to calibrate your interval.

Part B: Does the interval you calculated in Part A support the claim that white individuals are more likely to have health insurance than black individuals in the United States? Briefly explain, limiting your response to no more than 2 sentences.

Part C: Suppose your friend decides that a 99% confidence level is more appropriate for the interval estimate you calculated in Part A. Would this interval suggest a *wider range* or a *narrower range* of plausible differences in the population? Briefly explain your answer, limiting your response to no more than 2 sentences.

Part D: Consider the task of estimating the mean income of all individuals in United States belonging to each of the four racial categories used by the ACS researchers. This task would involve 4 different confidence intervals, one for each racial category. Briefly explain which of these interval estimates would have the smallest margin of error. You should assume all of the intervals use the same confidence level.

Part E: Without considering any other factors, interpret the relationship between age and income in the United States population based upon an analysis of the ACS data.

Part F: Now consider the third variable "race" in regard to the interpretation you provided in Part E. Do you believe it is necessary to perform a stratified analysis by race to have a reasonable understanding of how age and income are related in the United States population? Briefly explain.

Question #3 (hypothesis testing)

An experiment conducted at the University of Sydney in Australia investigated whether electrical stimulation to the brain could help participants successfully solve problems that required non-routine approaches.

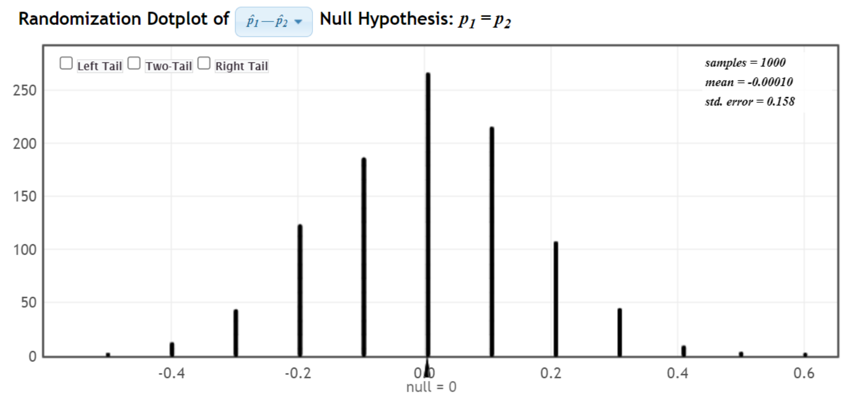
The experiment trained 40 participants to solve problems in a particular way and then asked them solve an unfamiliar problem that required a creative solution, 20 of the participants were randomly assigned to receive electrical stimulation to the brain, and the other 20 received a placebo condition (the same apparatus without any electricity).

In the electrical stimulation group 60% successfully solved the problem, while only 20% of the placebo group solved the problem.

Part A: Perform the initial steps of an appropriate z or t test to evaluate the hypothesis that electrical stimulation improves the ability to solve non-routine problems, as summarized by a difference in proportions. Your answer should clearly state your hypotheses, and properly calculate a test statistic. *You do not need to find a p-value or make a conclusion.*

Part B: Considering the assumptions of the hypothesis test you began in Part A, would you trust that the p-value that would arise from this test would be accurate? Briefly explain.

Part C: The figure below shows a randomization distribution for these data (generated under the hypothesis that an equal proportion of each group successfully solve the problem). Use this randomization distribution to *estimate* the p-value. Your estimate does not need to be exact; any estimate within reason will be scored as a correct answer.



Original Sample

Group	Count	Sample Size	Proportion
Group 1	12	20	0.600
Group 2	4	20	0.200
Group 1-Group 2	8	n/a	0.400

Randomization Sample

Group	Count	Sample Size	Proportion
Group 1	10	20	0.500
Group 2	6	20	0.300
Group 1-Group 2	4	n/a	0.200

Figure 1: StatKey output used in Question Q3-C

Part D: Provide a 1-2 sentence conclusion using the p-value you estimated in Part C and the context of the experiment.

Part E: Suppose you used bootstrapping to construct a 95% confidence interval estimate for the difference in the proportion of problems solved. Would you expect this confidence interval estimate to suggest that a difference of zero is plausible? Briefly explain.