Lab 7 - Math 58B: Relative Risk & Odds Ratios

your name here

due Tuesday March 8, 2022

Goals:

Just like any statistic which is measured on a dataset, the variability of the sample relative risk ($\widehat{RR}$) and the sample odds ratio ($\widehat{OR}$) is described by a sampling distribution. The goals for today include being able to:

- describe the null sampling distribution of ($\widehat{RR}$ and) $\widehat{OR}$ under the null hypothesis using a randomization test
- describe the sampling distribution of ($\widehat{RR}$ and) $\widehat{OR}$ for general setting using bootstrapping
- describe the sampling distribution of ($\widehat{RR}$ and) $\widehat{OR}$ (under the null or the alternative) using the Central Limit Theorem (normal distribution!)
- use the randomization distribution to run a complete hypothesis test
- use the bootstrap distribution to create a confidence interval
- use the normal distribution to run a complete hypothesis test

Advice for turning in the assignment

- knit early and often. In fact, go ahead and knit your .Rmd file right now. Maybe set a timer so that you knit every 5 minutes. Do **not** wait until you are done with the assignment to knit.

- The **assignment** part of the lab is **ONLY** the last six questions at the very bottom. However, the commands in the first half of the assignment are key to doing the second half.

- Save the .Rmd file somewhere you can find it. Don't keep everything in your downloads folder. Maybe make a folder called `StatsHW` or something. That folder could live on your Desktop. Or maybe in your Dropbox.

Getting started

- the infer package will be used for the R analysis
- the following two applets may help with forming intuition
  - Randomization test applet: http://www.rossmanchance.com/applets/2021/chisqshuffle/ChiSqShuffle.htm
  - Bootstrapping applet: http://www.rossmanchance.com/applets/2021/twopopprop/twopopprop.html

Load packages / data

In this lab we will continue to use the **infer** package, along with the applets. The infer syntax is meant to focus understanding on the randomization and bootstrapping processes. For each line, pay attention to what the code is doing.

> A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group to determine whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-Dichlorophenoxyacetic acid (`2,4-D`).

> Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. Case- Control Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4- Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.

The dog data are in **openintro**.

```
data("cancer_in_dogs")

cancer_in_dogs %>%
  table()
```

```
##           response
## order      cancer no cancer
##   2,4-D       191       304
##   no 2,4-D    300       641
```

Variablity of $\widehat{OR}$ when H_0 is true

The randomization process (which leads to the randomization test) creates different possible values of $\widehat{RR}$ and $\widehat{OR}$ assuming the null hypothesis is true.

See: Randomization test applet: http://www.rossmanchance.com/applets/2021/chisqshuffle/ChiSqShuffle.htm

Note that the variability of the statistics (the sampling distribution) can be displayed by applying the same ideas that were used on the difference in sample proportions.

What are the steps needed to create a sampling distribution?

```
cancer_in_dogs %>%
  ___(response ~ order, success = "cancer") %>%
  ___(null = "independence") %>%
  ___(reps = 1000, type = "permute") %>%
  ___(stat = "odds ratio", order = c("2,4-D", "no 2,4-D")) %>%
  ___() +
  xlab("statistic = odds ratio")
```

QUESTION

- What is the computer doing to generate the null sampling distribution of $\widehat{OR}$?
- Can you describe a tactile method that could mimic what the computer code does?

Hypothesis test of OR when H_0 is true

Because the data come from a case-control study, we'll choose to investigate the OR (instead of the RR). Note that the test is one-sided because the researchers suspected that the herbicide was associated with higher cancer rates.

$H_0 : OR = 1$
$H_A : OR > 1$

The p-value is 0.042 (you might have gotten a slightly differnt answer if you set a different seed). We can reject the null hypothesis and claim that the true probability ("risk") of earning more than $25,000 is higher for those who have a college degree than those who don't.

```
set.seed(4774)
obs_or <- cancer_in_dogs %>%
  specify(___ ~ ___, ___ = "___") %>%
  calculate(stat = "___", order = c("___", "___"))

obs_or

null_or <- cancer_in_dogs %>%
  specify(___ ~ ___, ___ = "___") %>%
  hypothesize(___ = "___") %>%
  generate(___ = ___, ___ = "___") %>%
  calculate(stat = "___", order = c("___", "___"))

null_or %>%
  ___() +
  xlab("statistic = odds ratio") +
  shade_p_value(obs = ___, direction = "___")

null_or %>%
  ___(obs = obs_or, direction = "___")
```

Variablity of $\widehat{OR}$ with no hypothesis

The bootstrap process (which leads to a bootstrap confidence interval) creates different possible values of $\widehat{OR}$ **without** assuming the null hypothesis is true.

See: Bootstrapping applet: http://www.rossmanchance.com/applets/2021/twopopprop/twopopprop.html

Note that the variability of the statistics (the sampling distribution) can be displayed by applying the same ideas that were used on the difference in sample proportions. It is important to observe that the distribution (for $\widehat{OR}$) is not centered at 1.

```
set.seed(5)
cancer_in_dogs %>%
  ___(response ~ order, success = "cancer") %>%
  ___(reps = 1000, type = "bootstrap") %>%  # no hypothesize step!!!
  ___(stat = "odds ratio", order = c("2,4-D", "no 2,4-D")) %>%
  ___() +
  xlab("statistic = odds ratio")
```

QUESTION

- What is the difference in the **code** from the H_0 curve vs the no hypothesis curve?
- What is the difference in the **look of the histograms** from the H_0 curve vs the no hypothesis curve?

Bootstrap Confidence Interval for OR

Using the histogram above, the CI for the OR can be taken directly from the bootstrap sampling distributions.

We are 93% confident that the true odds of lymphoma is between 1.09 and 1.64 times higher for those dogs who have been exposed to 2, 4-D than those dogs who have not.

```
set.seed(314)
boot_or <- ___ %>%
  specify(___ ~ ___, success = "___") %>%
  generate(___ = ___, type = "___") %>%  # no hypothesize step!!!
  calculate(stat = "___", order = c("___", "___"))

ci_or_93 <- get_ci(boot_or, level = ___)
```

```
ci_or_93

visualize(boot_or) +
  shade_confidence_interval(endpoints = ci_or_93, color = "red", fill = "pink") +
  xlab("statistic = odds ratio")
```

Variability of $\widehat{OR}$ assuming H_0 is true.

As with other statistics we've seen, the central limit theorem tells us about the distribution of $\ln(\widehat{OR})$, if the sample size is big enough. Indeed, the sampling distribution of $\ln(\widehat{OR})$ can be written as:

$$\ln(\widehat{OR}) \sim N\left(\ln(OR), \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \right)$$

It is hard to compare the Normal curve to the (computational) curves above because the computational curves describe the (slightly skewed) distribution of $\widehat{OR}$ and the normal curve describes the distribution of $\ln(\widehat{OR})$. However, you should be able to sketch (with a pencil) the theoretical sampling distribution of $\ln(\widehat{OR})$, given the information above.

Write the null and alternative hypotheses in terms of the true parameter ln(OR).

Find a Z-score

Using the Z-score, find the p-value

Conclude Using the p-value, conclude the test using words like **odds**, herbicide, dogs, cancer.

QUESTION

- What is the difference in the computational sampling distribution vs. the normal theory sampling distribution?
- What is the difference in the hypothesis test conclusions for the computational sampling distribution vs. the normal theory sampling distribution?

To Turn In

The data

Consider the article (and data therein) on sleepy driving: (Connor et al. "Driver sleepiness and risk of serious injury to car occupants: population based case control study", British Medical Journal, 2002, https://www.bmj.com/content/bmj/324/7346/1125.1.full.pdf)

Although the researchers look at many variables, consider the two following two variables: (1) driver sleepiness score of 1-3 vs 4-7, (2) driver involved in an "injury crash" or not. Note that there seems to be some missing information with respect to the sleepiness score. [See Table 1 in the paper.]

You will need to enter the data as you did in Lab 6 (after looking at Table 1 in the paper).

Q1. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Variables

In the study on sleepy driving, which is the explanatory and which is the response variable? What is an observational unit?

Q3. Study

How were the data selected, as a case-control study or a cohort study?

What aspect of the population is it impossible to know given how the data were sampled:

- is it impossible to estimate the proportion of people who crash (of those who are sleepy)?
- Or is it impossible to estimate the proportion of people who are sleepy (of those who crash)?

Note: in a case-control study, the participants are selected based on the value of the response variable. In a cohort study, the participants are selected based on the value of the explanatory variable. Ask yourself, how were the people in this study selected?

Q4. OR computational hypothesis test

Using the **infer** syntax and the odds ratio, complete a full hypothesis test: state null and alternative hypotheses with parameter values, the p-value from the randomization test, and the conclusion in words of the problem.

Q5. OR conmputational confidence interval

Using the infer syntax, find and interpret a 90% bootstrap percentile CI for the true OR of the variables above. [Note: to complete this question you will need to write out the words which describe the true OR. Do not use words like "probability" "rate" or "risk". Instead, use the word "odds".]

Q6. OR normal theory hypothesis test

Using central limit theorem, complete a full hypothesis test: state null and alternative hypotheses with parameter values, the Z-score, the p-value, and the conclusion in words of the problem.

Q7. "strong" and "evidence"

- Which part(s) of the analysis reveals evidence of a strong association? Explain.
- Which part(s) of the analysis reveals strong evidence of an association? Explain.

```r
praise()
```

```
## [1] "You are rad!"
```

HW & Lab assignments will be graded out of 5 points, which are based on a combination of accuracy and effort. Below are rough guidelines for grading.

Score & Description

5 points: All problems completed with detailed solutions provided and 75% or more of the problems are fully correct.

4 points: All problems completed with detailed solutions and 50-75% correct; OR close to all problems completed and 75%-100% correct. An assignment will earn a 4 if there is superfluous information printed out on the assignment.

3 points: Close to all problems completed with less than 75% correct

2 points: More than half but fewer than all problems completed and $>$ 75% correct

1 point: More than half but fewer than all problems completed and $<$ 75% correct; OR less than half of problems completed

0 points: No work submitted, OR half or less than half of the problems submitted and without any detail/work shown to explain the solutions.

General notes on homework assignments (also see syllabus for policies and suggestions):

- please be neat and organized, this will help me, the grader, and you (in the future) to follow your work.
- be sure to include your name on the assignment
- please include at least the number of the problem, or a summary of this question (this will also be helpful to you in the future to prepare for exams).
- for R problems, it is required to use R Markdown. You can write out other problems with pencil and combine pdf as appropriate.
- please do not print errors, messages, warnings, or anything else that makes your homework unwieldy. You will be graded down for superfluous printouts.
- in case of questions, or if you get stuck please don't hesitate to email me or DM on Discord! The sooner (and more often) questions get asked, the better for everyone.