STA130 (Winter 2022) Midterm Examination

Professor Scott Lee Schwartz

Thursday October 28^{th}, 2022

- **All answers you provide must be your own.**
- **A pencil rather than a pen is recommended for this examination.**
- **You may have ONE "any size font front+back"** $8\frac{1}{2} \times 11\frac{1}{2}$ **"cheat sheet".**
- ***You may not use any other resources during the duration of the exam.***
- **You may not use your phone or a calculator or a computer, etc.**
- **All such items (phone, etc.) must remain stored at all times.**
- **You may not take any items with you if you leave the room.**

0. To keep things fair, no questions about the exam will be answered during the duration of the exam. If you think there is a problem with a question, note the question and briefly describe the problem in the space below and your concern will be evaluated during marking.

1. *In the box below*, explain what the code below does, explaining explicitly what `%>%` and `object` are.

   ```
   "file.csv" %>% read_csv() -> object
   ```

   (1) '%>%' passes or "pipes"
   (1) the string "file.csv" into 'read_csv()'
   (1) which stores the data stored in file "file.csv"
   (1) in the 'object' object as a 'tibble'

2. *Use three of the four options below to fill in the blanks and complete the following analogy sentence.*

   "If functions in R are like light, then D or restarting Jupyterhub breaks a lot of lightbulbs, but B or `install.packages()` screws in a new lightbulb, and C or `library()` turns on the light switch."
   1 point if ALL are right; 0.5 if TWO are right; 0 points otherwise

   A. cloud GUI IDE B. `install.packages()` C. `library()` D. restarting Jupyterhub

Broken Off On

3. *In the follow blank space*, put the value of `(TRUE | (TRUE | FALSE)) & (TURE & FALSE)`: (1) FALSE

4. What is the most important thing that Rstudio does? *Select ONE of the following.* (1) A

   A. Facilitates R analysis reproducibility
   B. Saves files and manages R packages
   C. Allows you to code and program in R
   D. Makes R easily accessible in the cloud

5. *Fill in the blanks below.*

   - A colour measured scientifically as a wave frequency number is
     (0.2) continuous or numeric or quantitative variable represented as a (0.2) double or float or numeric data type and visualized using (0.2) geom_histogram() or geom_boxplot() in R.
   - A word describing a colour is
     (0.2) nominal categorical or nominal qualitative variable represented as a (0.2) string or factor or integer data type and visualized using (0.2) geom_bar() in R.
   - A day of the week starting on Monday is
     (0.2) ordinal categorical or ordinal qualitative variable represented as a (0.2) string or factor or integer data type and visualized using (0.2) geom_bar() in R.
   - An either/or variable is
     (0.2) logical or binary or boolean or TRUE/FALSE variable represented as a (0.2) logical data type and visualized using (0.2) geom_bar() in R.
   - The number of questions on an exam is
     (0.2) discrete or numeric or quantitative variable represented as a (0.2) double or float or numeric data type and visualized using (0.2) geom_histogram() or geom_boxplot() or geom_bar() in R.

6. *Indicate in the box below* what important parameter you should consider when using `geom_histogram()` that you don't need to consider when using `geom_boxplot()` and `geom_bar()`.

   (1) bins or bin or binwidth

7. For the *tibble* called `people` with the column `handedness`, with values that are either `"left"` and `"right"`, *use the three boxes below to indicate the three things that are wrong with the code below.*

```
people %>%
  ggplot(x=handedness, y="") %>%
  geom_boxplot()
```

   Problem # 1:

   (1) should not use geom_boxplot() – should use geom_bar() – for this kind of data

   Problem # 2:

   (1) missing 'aes()' wrapper

   Problem # 3:

   (1) ggplot should use + not %>%

8. The encoding of the `handedness` variable in the `people` *tibble* above is not the only way this variable could be encoded. *Indicate in the boxes below* what the encoding of the `handedness` variable in the `people` *tibble* is expected by the `mutate()` function for each of the two code chunks.

   Hint: the code comments in the NEXT problem may be helpful for understanding the "`TRUE ~`" code.

```
# left code chunk
people %>% mutate(
  case_when(handedness ~ "right",
                  TRUE ~ "left"))
```

```
# right code chunk
people %>% mutate(
  case_when(handedness == "right" ~ 1,
                                TRUE ~ 0'))
```

   `# left code chunk`

   (1) "right" is encoded as TRUE, and "left" as FALSE

   `# right code chunk`

   (1) handedness is coded as in the previous problem OR "right" is encoded as "right", and "left" as "left"

9. Assume the `people` *tibble* doesn't have columns named `row_id` and `even_odd` before this code is run.

   Hint: the comments in the code truthfully below explain what each part of the code does.

```
people %>%
  mutate(row_id = row_number(), # column row_id now stores the row number
                        # if the row_id value divided by 2 has remainder 0
                        # then row_id %% 2 == 0 is TRUE
         even_odd = case_when(row_id %% 2 == 0 ~ "even", # if this line is TRUE
                                               TRUE ~ "odd")) # this case is not run
```

   What columns are added to the `people` *tibble* after this code is run? *Select ONE of the following.*

   (1) D or Neither `row_id` nor `even_odd`

   A. `row_id` B. `even_odd` C. Both `row_id` and `even_odd` D. Neither `row_id` nor `even_odd`

10. *Fill in the blanks in the following sentences and in the code below.* "Possible (1) missing values or NAs can be handled with either of the completed code constructions below, but the (1) first or filter() version is preferable because then the (1) n() and mean() functions operate on the same input."

    Code blanks are (1) filter(!is.na(handedness)) and (1) na.rm=TRUE

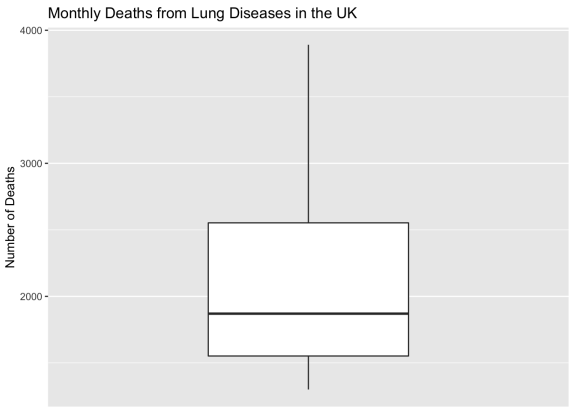
```
people %>% ______________________________  %>%
  summarise(n=n(), `% right handed` = paste(100*mean(handedness, __________), "%", sep=""))
                                    # paste() uses coercion to change the number to
                                    # a character string onto which it appends "%"
```

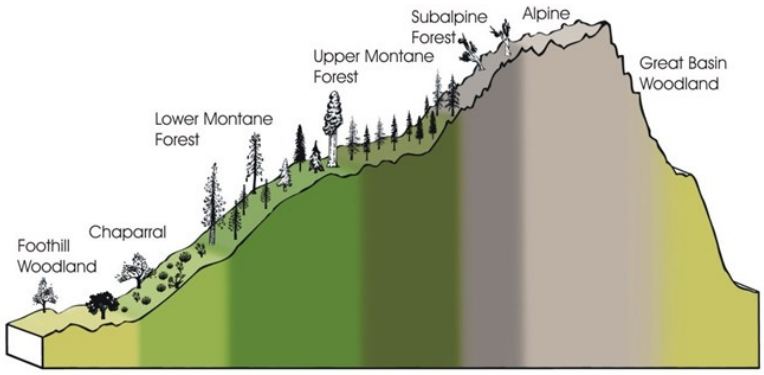
    "Still, this code above will only work if `handedness` is encoded as (1) 1's and 0's or "right" as 1 and "left" as 0 or (due to (1) coercion) if `handedness` is encoded as (1) TRUE's and FALSE's or "right" as TRUE and "left" as FALSE."

11. What is the modality of this distribution given in the boxplot below (left figure)? (1) E

A. Unimodal B. Bimodal C. Multimodal D. Uniform E. Can't tell

12. *Fill in the blanks to complete the following sentence.* "Describe the mountain below (right figure) as if it was a data distribution. This mountain is unimodal and left skewed (which means that as a data distribution its *median* will be greater than the *mean*.)" 2 points if ALL are right; 1 point if TWO are right; 1 if ONE is right; 0 otherwise.





13. *In the box below* write the *sample mean* $\bar{x}$ and *sample standard deviation* $s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ for the sample x which has the values `c(1, 2, 3)`?

$\bar{x} = 1$ and $s = s^2 = 1$

1 points if BOTH are right.

14. For a *tibble* called `olympics` (which includes the `year`, `athlete`, `age`, and `event` variables), which of the following R code chunks results in a *tibble* containing only the `athlete`, `age`, and `event` information for the 10 youngest `athletes` for `year` $\geq$ 2000 (greater than or equal to 2000)?
Select ONE of the following choices. (1) B

A.
```
olympics %>% select(athlete, age, event) %>%
  filter(year >= 2000) %>% arrange(desc(age)) %>% head(10)
```

B.
```
olympics %>% filter(year >= 2000) %>%
  select(athlete, age, event) %>% arrange(age) %>% head(10)
```

C.
```
olympics %>% filter(year > 2000) %>% head(10) %>%
   select(athlete, age, event) %>% arrange(age)
```

D.
```
olympics %>% filter(year > 2000) %>% arrange(age) %>%
  head() %>% select(athlete, age, event)
```

15. For the `olympics` *tibble* noted above, which of the following R code chunks gives the average `age` of *athletes* for each `event` for `year` < 2000 and `year` ≥ 2000 sorted by oldest to youngest average ages? *Select ONE of the following choices.* (1) A

A.
```
olympics %>% mutate(pre2000 = case_when(year<2000~TRUE, TRUE~FALSE)) %>%
  group_by(event, pre2000) %>% summarize(ave_age = mean(age, na.rm=TRUE)) %>%
  arrange(desc(ave_age))
```

B.
```
olympics %>% arrange(age, year<2000) %>%
  group_by(event, year<2000) %>% summarize(mean(age, na.rm=TRUE))
```

C.
```
olympics %>% mutate(case_when(pre2000 = year<2000~TRUE,
                              post2000 = year>=2000~TRUE) %>%
  group_by(event, pre2000, post2000) %>% arrange(desc(ave_age)) %>%
  summarize(ave_age = mean(age, na.rm=TRUE))
```

D.
```
olympics %>% group_by(event & year<2000 | event & year>=2000) %>%
  summarize(ave_age = mean(age, na.rm=TRUE))) %>% arrange(desc(ave_age))
```

16. Which of the following best describes the *test statistic* of the *data generating mechanism* given in the following code? *Select ONE of the following choices.* (1) C

```
phat <- mean(sample(c(0,1), size=100, prob=c(1/3,2/3), replace=TRUE))
```

A. The proportion of heads for coin flips

B. The proportion of heads for 100 coin flips where the chance of heads is 1/3

C. The proportion of heads for 100 coin flips where the chance of heads is 2/3

D. This code won't run as is with `replace=TRUE`

17. *In the two boxes below*, describe the values in the `test_stats` object when the following code is run for `n=2` and `n=3`.

```
test_stats = 1:N; for(i in 1:N){set.seed(123)
  x <- sample(c(0,1), size=n, replace=FALSE)
  test_stats[i] <- mean(x, na.rm=TRUE)
}
```

`n=2`

(1) just 0.5's or a whole vector or list or sequence or 0.5

`n=3`

(1) the code will produce an error or it will be a vector or list or sequence of numbers 1 to N

18. Which TWO of the following return a "shuffled" version of `x`? *Choose TWO that apply.*
Hint: `rep(1,2)` returns `c(1,1)`. (1) A and C no partial credit

A. `sample(x, size=length(x), replace=FALSE)`
B. `sample(x, size=length(x), prob=rep(1/length(x), length(x)), replace=TRUE)`
C. `sample(x, size=length(x), prob=rep(1/length(x), length(x)), replace=FALSE)`
D. `sample(x, size=length(x), prob=c(0.5,0.5), replace=TRUE)`

19. What assumption does mixing and shuffling two samples up implicitly make about the nature of the two samples? *In the box below, describe what you're assuming about two populations generating two samples when you're willing to shuffle the two samples, as in a two-sample permutation test.*
(2) It assumes that the populations are exchangeable or the same
OR It assumes that there's no difference between the groups, so they can be mixed
OR $p_1 = p_2$ or $H_0 : p_1 = p_2$ or some variant of this

20. *Explain in the box below*, why the code below would no longer simulate a sampling distribution for a *two-sample permutation test* for `data` if `size=n1+n2` (in the first function in the `for` loop) was changed to `size=n1`? Address the *test statistic* compared to `permutation_test_stats` if `size=n1`.

```
# synthetic data created as an example data set
n1 <- 100; n2 <- 100; groups <- c(rep("one", n1), rep("two", n2))
x1 <- rnorm(mean=0, n=n1) # sample(c(0,1), size=n1, replace=TRUE)
x2 <- rnorm(mean=1, n=n1) # sample(c(0,1), size=n2, p=c(1/3,2/3), replace=TRUE)
data <- tibble(group = groups, outcome = c(x1, x2))
N <- 10000; permutation_test_stats <- 1:N; set.seed(130)
for(i in 1:N){
  shuffled_data <- data %>% mutate(group = sample(group, size=n1+n2, replace=FALSE))
  # shuffling the groups above, instead of shuffling the outcomes as done below
  # shuffled_data <- data %>% mutate(outcome = sample(outcome), size=n1+n2, replace=FALSE)
  # shuffled_xs <- sample(c(x1,x2), size=n1+n2, replace=FALSE) # sample(c(x1,x2))
  permutation_test_stats[i] <- shuffled_data %>% group_by(group) %>%
    summarise(means = mean(outcome), .groups="drop") %>%
    summarise(value = diff(means)) %>% as.numeric()
  # the above is equivalent to the following if we were instead using `shuffled_xs`
  # permutation_test_stats[i] <- mean(shuffled_xs[1:n1])-mean(shuffled_xs[(n1+1):(n1+n2)])
}
```

(3) we wish to judge the test statistic, so we need the sampling distribution of the test statistic under the null hypothesis; but, if the sample size of the permutation test statistics comprising the sampling distribution is different than the actual sample size of the observed test statistic then these are not comparable. – partial credit may be awarded for answers beginning to address this issue.

21. Which code below is the most general statement of a *p-value*? *Select ONE of the following choices.* (1) D

A. `mean(abs(sim_teststats)<=abs(obs_teststat))`

B. `mean(abs(sim_teststats)>=abs(obs_teststat))`

C. `mean(abs(sim_teststats-H0_parameter)<=abs(obs_teststat-H0_parameter))`

D. `mean(abs(sim_teststats-H0_parameter)>=abs(obs_teststat-H0_parameter))`

22. *Write in the box below* the definition of a *p-value*.

(0.5) The probability of observing a test statistic that is as or more extreme than the one we actually observed if the null hypothesis was true. – no partial credit

23. Which of the following is a true description of the *p-value*? *Select ONE of the following choices.* (0.5) D

A. The probability the *NULL hypothesis* H_0 is true

B. The probability the parameter p of the *NULL hypothesis* $H_0 : p = p_0$ equals p_0

C. The probability of a *Type-I* error.

D. None of the above

24. For an $\alpha = 0.01$ *significance* (formal hypothesis) test, if a *p-value* of 0.02 has been calculated, what type of error might we make? *Select ONE of the following choices.* (1) B

A. A Type I Error B. A Type II Error C. A Type III Error D. None of the above

25. *Explain in the box below* what happens to the *p-value* as `N` is increased in the following procedure.

```
for(i in 1:N){
  # sample data generating mechanism
  # calculate and save simulated test statistic
}
# compare observed test statistic to simulated test statistics
# to produce the estimated p-value of the observed test statistic
# with respect to the hypothesized data generating mechanism
```

(2) The p-value [estimated by the simulation] becomes more and more accurate.

OR the p-value changes [randomly] less and less as N is increased. – no partial credit

26. From "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine" published in the New England Journal of Medicine by Polack et al. of the C4591001 Clinical Trial Group:

> A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of [laboratory-confirmed] Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo. ... Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient.

and the study found a vaccine efficacy of $\tilde{x} = 1 - (8/21720)/(162/21728) \approx 95\%$ with an associated *p-value* that resulted in the rejection of the *NULL hypothesis* at the $\alpha = 0.025$ level.

(a) What does the test statistic $\tilde{x} \approx 95\%$ measure? *Select ONE of the following choices.* (1) C

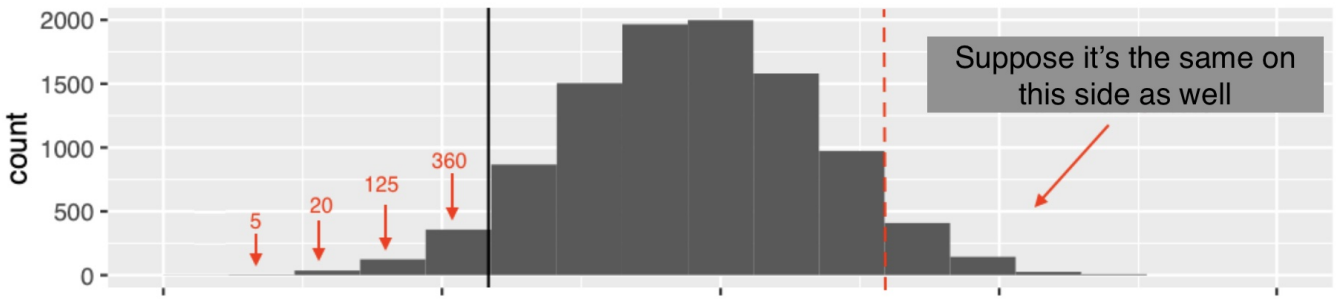
A. The rate of Covid-19 in the placebo group ($162/21728 \approx 0.0075$)
B. The rate of Covid-19 in the BNT162b2 group ($8/21720 \approx 0.0004$)
C. A ratio of Covid-19 rates between BNT162b2 and placebo groups
D. The chance of death from Covid-19 between BNT162b2 and placebo groups

(b) *State in the box below* the *NULL hypothesis* of the hypothesis in terms of p_1 and p_2, the chances of laboratory-confirmed Covid-19 cases in the BNT162b2 and placebo groups?
(1) $H_0 : p_1 = p_2$ – half mark if "H_0" isn't included

(c) *Fill in the following blanks.* "The *test statistic*'s *p-value* is (1) less than or equal to $\alpha = 0.025$."
1 point if BOTH are right only – no partial credit

(d) Suppose the following *simulated sampling distribution* of the *test statistic* based on $10,000$ *simulated test statistics* created assuming the *NULL hypothesis* is true was used to produce the *p-value* for this study. Assuming the *p-value* is based on a *two-sided hypothesis test* (where "as or more extreme" is symmetric), what is the "tallest" bin the *test statistic* could have fallen into while still definitely producing the result of this study? *Select ONE of the following choices.*
(2) B 2*(5+20+125) = 300 means the p-value could have been 0.03 if it fell into the "125 bin".

A. The bin with 5 counts
B. The bin 20 counts
C. The bin with 125 counts
D. The bin with 360 counts



27. *Fill in the blanks to complete the following sentence.* "A (0.5) statistics estimates a (0.5) parameter; but, when bootstrapping a (0.5) sample approximates a (0.5) population."

28. *Fill in the blanks to complete the following bootstrapping code and, as asked by the prompts below, explain in the boxes below the code* why your choices for `size` and `replace` parameters respectively makes `bootstrap_stats` approximate a *sampling distribution* that is relevant for the *test statistic*.

```
bootstrap_statistics <- 1:B
for(b in 1:B){
```

(1) length(x) or (0.5) partial mark for "n" and (1) TRUE and

```
  bootstrap_sample <- sample(x, size=_________, replace=_______)
  bootstrap_stats[b] <- statistic(bootstrap_sample) # a sample statistic
}
```

(1) quantile or (0.5) partial mark for "percentile" and (1) 0.025, 0.0975

```
_________(bootstrap_stats, c(_________)) # 95% bootstrap confidence interval
```

Why would the opposite choice of `replace` *not work for your choice of* `size`?

(1) every bootstrap sample would be the same!

For your choice of `replace` *why must you also use your choice of* `size`?

(1) So the bootstrap test statistics are based on the same sample size as the observed test statistic!

29. *Indicate in the box below* what, for the given *fixed confidence level* of 95% above, could be increased in the context of the *bootstrapping* code above to reduce the width of the *confidence interval*?

(1) The sample size n of the original sample x!

30. Which of the following correctly represents a 90% confidence interval $I = [\hat{\theta}_{\texttt{lower bound}}, \hat{\theta}_{\texttt{upper bound}}]$ for parameter θ? *Select ONE of the following choices.* (1) B

A. A probability about θ; namely, $\Pr_\theta(\hat{\theta}_{\texttt{lower bound}} \leq \theta \leq \hat{\theta}_{\texttt{upper bound}}) = 0.90$

B. A probability about the statistic I; namely, $\Pr_I(I \texttt{ bounds the true parameter value}) = 0.90$

C. Exactly one out of every ten 90% *confidence intervals* will capture the true *parameter* value

D. None of these correctly represent 90% *confidence intervals*