SOC6302 Winter 2023

Midterm Exam

3 March 2023

Details

- There are 25 questions.
- You will have 2 hour and 50 mins to complete the exam (10:10am – 1pm).
- For multiple choice questions, circle your answer(s).
- For short answer questions, write your answer in the space provided.

1

I wish to study how long people at UofT spend watching shows on Netflix. To do so, I decide to undertake the following:

- Contact a random sample of 100 UofT staff and students
- Ask them the following question: "Approximately how many minutes did you spend watching Netflix yesterday?"
- Calculate a quantity called $\hat{\theta}$ by sorting the 100 responses of minutes from smallest to largest and reporting the number of minutes which falls in the middle.

The quantity $\hat{\theta}$ is a

a) Mean
b) Median
c) Interquartile range
d) Impossible to say without seeing the data

2

It seems like a lot of businesses have closed in downtown Toronto since the pandemic. To investigate this, I decide to walk along some blocks downtown and count the number of businesses that are closed and open. To decide which blocks to walk, I open a map of Toronto, randomly choose 3 zip codes, and then randomly sample 10 blocks to walk along within each zipcode. This type of sampling is (circle all that apply):

a) Cluster sampling
b) Stratified sampling
c) Simple random sampling
d) Convenience sampling

3

I have a dataset that contains measurements of height (in cm) for a sample of 300 penguins, who are either the Adeline or Emperor species. I am interesting in visualizing the distribution of heights by species in a graphical way. What is an appropriate type of graph to use? Circle all that apply.

a) Line graph
b) Scatter plot
c) Box plot
d) Histogram

4

I have a dataset of 1000 observations of 23 different variables. There are no missing values for any of the variables. Do I still need to carry out exploratory data analysis on this dataset? Why or why not?

5

I have two datasets that report individual's swimming times for 100 metres (variable X) and 400 metres (variable Y). Both datasets contain observations for 50 people. Based on calculating summary statistics, I conclude that $\bar{X}$, $\bar{Y}$, and the correlation between X and Y is the same across the two datasets. Do I need to plot my data? Why or why not?

6

Suppose all my shirts are either blue or white, and I have 5 blue shirts and 5 white shirts. Every day I randomly choose a shirt to wear. Each day I note down whether or not the chosen shirt was blue.

a) The number of times I wear a blue shirt in a week is a (circle one)

- random variable in this experiment
- probability of an outcome
- probability distribution for this experiment
- constant in this experiment

b) The probability of wearing a blue shirt all five days in a week is a (circle one)

- random variable in this experiment
- probability of an outcome
- probability distribution for this experiment
- constant in this experiment

c) If we found the probability of wearing a blue shirt one time, two times, three times, four times, and five times in a week, the set of six probabilities would be the (circle one)

- random variable in this experiment
- probability of an outcome
- probability distribution for this experiment
- constant in this experiment

7

I am analyzing a dataset called `penguins`, which contains measurements of bill length, bill depth, flipper length, and body mass, for 344 penguins from 3 different species (Adelie, Chinstrap and Gentoo)

a) Describe the output of the R code below

```
penguins |>
  group_by(species) |>
  tally() |>
  mutate(proportion = n/sum(n))
```

8

This question also relates to the `penguins` dataset, as described above. Consider the following R code and output:

```
penguins |>
  drop_na() |>
  group_by(species) |>
  summarize(mean_flipper = mean(flipper_length_mm),
            corr_flipper_bill = cor(flipper_length_mm, bill_length_mm))
```

```
## # A tibble: 3 x 3
##   species   mean_flipper corr_flipper_bill
##   <fct>            <dbl>             <dbl>
## 1 Adelie            190.             0.332
## 2 Chinstrap         196.             0.472
## 3 Gentoo            217.             0.664
```

a) Interpret the meaning of the number '190'

b) Interpret the meaning of the number '0.472'.

9

I have a batch of 12 cookies. A total of 4 of the cookies are chocolate flavored, and the remaining cookies are oatmeal flavored. One of the chocolate-flavored cookies has chocolate chips, and 6 of the oatmeal-flavored cookies have chocolate chips. I put all the cookies into a paper bag and pull out one at random.

- Let A be the event that the cookie is chocolate flavored.
- Let B be the event that the cookie has chocolate chips.

Calculate the following probabilities (expressed as fractions):

a) $P(A)$

b) $P(B|A)$

c) $P(A \text{ and } B^c)$

d) $P(A^c|B)$

10

The mean for a population is

a) A parameter referred to as μ
b) A sample statistic referred to as μ
c) A sample statistic referred to as $\bar{Y}$
d) A parameter referred to as $\bar{Y}$

11

In a box plot, what percent of the observations fall in the box?

a) 25%
b) 50%
c) 75%
d) Depends on the shape of the distribution of data

12

What would happen to the standard deviation of a distribution of test scores if a teacher adds 10 points to each score?

a) It stays the same
b) It increases by 10 points
c) It becomes 10 times as large
d) It will increase, but the amount depends on the shape of the distribution

13

What would happen to the mean of a distribution of test scores if a teacher adds 10 points to each score?

a) It stays the same
b) It increases by 10 points
c) It becomes 10 times as large
d) It will increase, but the amount depends on the shape of the distribution

14

On a holiday in New Zealand, I considered bungee jumping off a 100 meter cliff. Before jumping, I considered two variables: (1) the distance below the cliff that I am (0 meters when I first jump), and (2) the distance above the ground that I am (100 meters when I first jump). What would the correlation between these two variables be?

15

If the distribution of sample means has the same variance as the population distribution, the sample size is:

a) 1
b) 0
c) 100
d) Impossible to say without seeing the data

16

In hypothesis testing, researchers may gain support for their research hypothesis by showing that the data are (circle all that apply):

a) Inconsistent with the null hypothesis
b) Consistent with the null hypothesis
c) Inconsistent with the alternate hypothesis
d) Consistent with the alternate hypothesis

17

Suppose we wish to test the null hypothesis that the population mean is 50 and that we have observed a sample mean of 55. We also note there is an outlying score of 35. After further checking we find that this score was entered incorrectly and that it should have been a 53. If we correct this data entry mistake, the t-value:

a) Will decrease
b) Will increase
c) Will stay the same
d) Impossible to say without seeing the rest of the sample

18

Compared to a 95% confidence interval for a proportion, an 80% confidence interval will be

a) Wider
b) Narrower
c) Be the same width 80% of the time
d) Impossible to say without seeing the data

19

I take a sample of 100 people who work full-time and record their income. The mean income of the sample is $55,000. The median income of the sample will be

a) Greater than $55,000
b) Smaller than $55,000
c) Equal to $55,000
d) Impossible to say without seeing the data

20

In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions", and the standard error for this estimate is 1.2%, and so the 95% confidence interval is (43%,47%). **Identify each of the following statements as true or false.**

a) We can say with certainty that the confidence interval contains the true percentage of U.S. adults who suffer from a chronic illness.

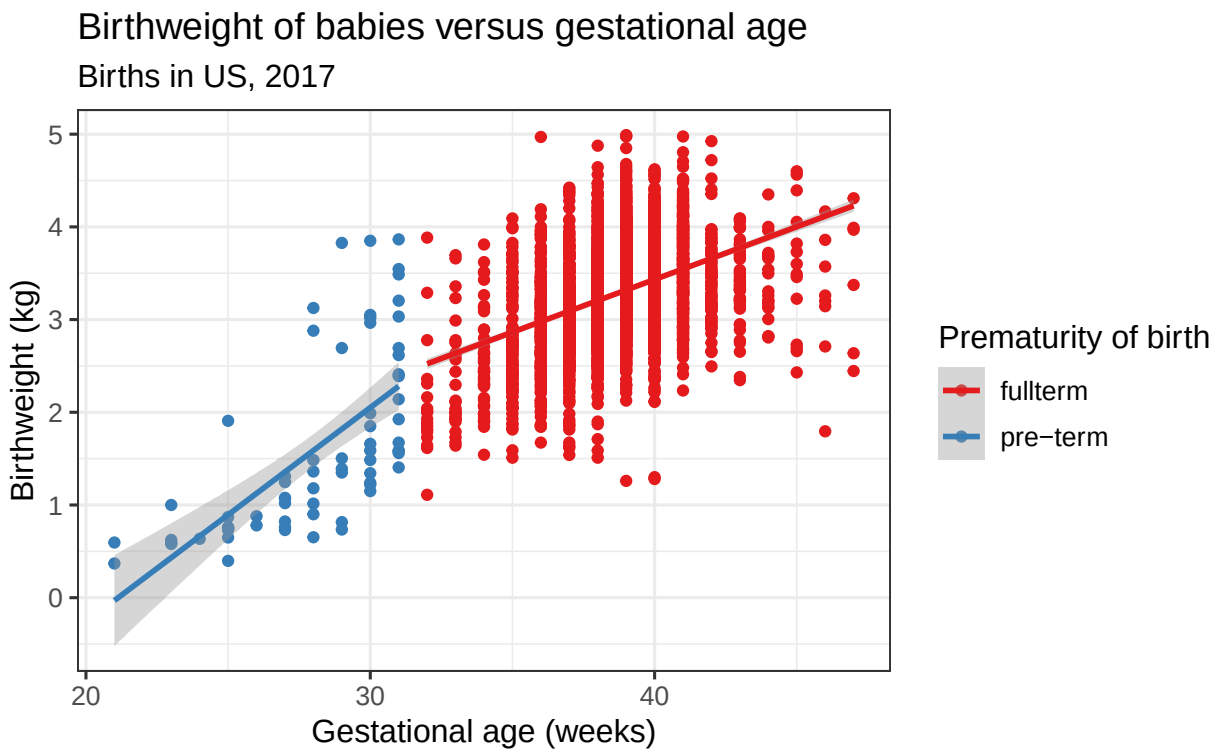
b) If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

c) Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

d) If we want to reduce the standard error of the estimate, we should collect less data.

21

Consider the chart below. Write down three main observations about the patterns you observe.



22

In 2017 the average time to complete the New York Marathon was 4.5 hours. Earlier this year, 12 runners ran the New York Marathon course, with an average time of 4 hours, with a standard deviation of 0.25 hours. I am interested to see whether runners are faster than they were in 2017.

a) Calculate the T-score for these data, given the formula

$$T = \frac{\bar{X} - \mu}{s}$$

b) More formally, I am interested in testing H_0: marathon times have stayed the same versus H_A: marathon times are different. Under the null hypothesis, how many degrees of freedom does the distribution of possible T-scores have?

c) The p-value for the T-score is 0.071. Assuming a significance level of $\alpha = 0.05$, what should I conclude?

23

Consider the following information:

- The mean nap length for children aged 9 months is 1 hour, with a standard deviation of 0.25 hours
- The mean nap length for children aged 18 months is 1.5 hours, with a standard deviation of 1 hour

We observe two nap times:

- Hugo (aged 18 months): 1.25 hours
- Lucio (aged 9 months): 1.25 hours

a) Calculate Hugo's and Lucio's Z scores. Recall that the formula for a Z score is

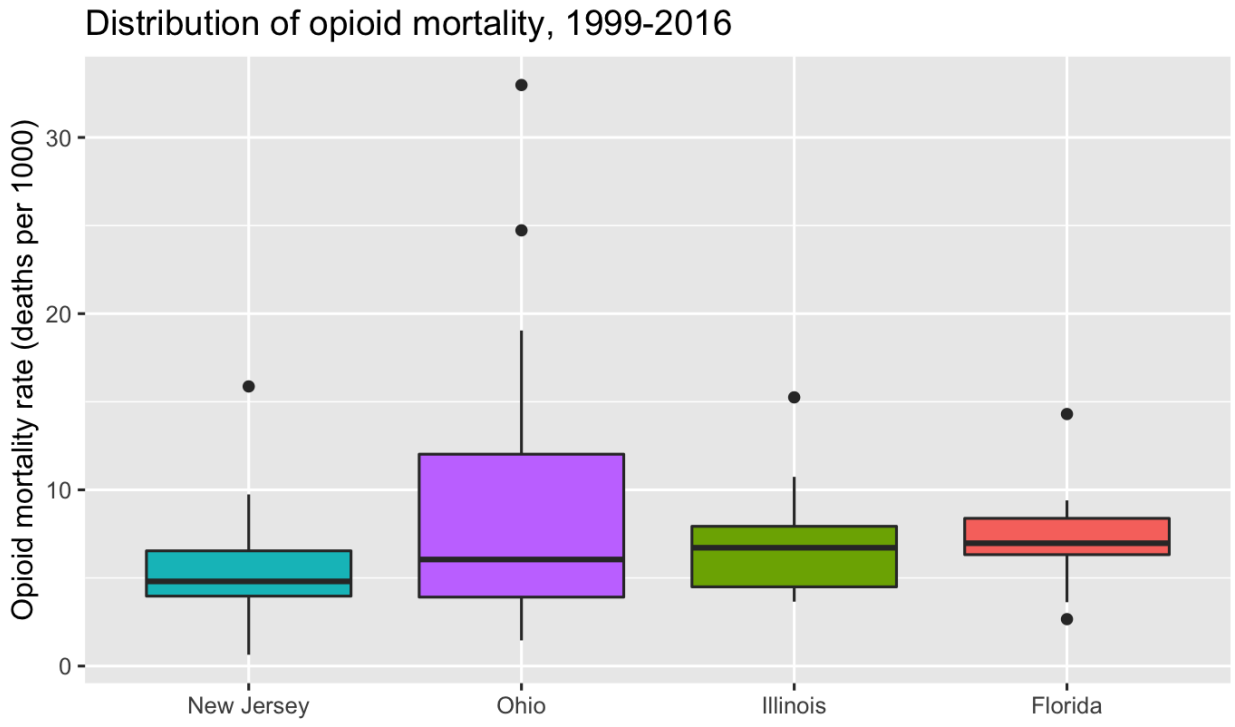
$$Z = \frac{X - \mu}{\sigma}$$

b) Which child's has a nap time that is further away from their population mean?

c) **True or False**: The probability of observing an 18 month old with a nap time shorter than Hugo's is greater than the probability of observing a 9 month old with a nap time shorter than Lucio's. If you cannot answer this question based on the information provided, explain what further information you would need.

24

Consider the chart below. Write down three main observations about the patterns you observe.



25

In R, I simulate 1000 flips of a biased coin, where the probability of flipping a head is $p = 0.4$. Based on a particular set of simulations, I calculate the sample proportion of heads to be 0.375.

a) I calculate a confidence interval for the proportion of heads based on the data using the code below. What is the level of the confidence interval? Report and interpret the confidence interval calculated below.

```
# calculate standard error
se <- sqrt(prop_heads*(1-prop_heads)/n)
# sample proportion
prop_heads
```

```
## [1] 0.375
```

```
# lower confidence interval bound
prop_heads-abs(qnorm(0.1))*se
```

```
## [1] 0.3553803
```

```
# upper confidence interval bound
prop_heads+abs(qnorm(0.1))*se
```

```
## [1] 0.3946197
```

b) Notice that the confidence interval above does not include the true probability of heads (0.4). Does this mean that these observations could not have been simulated using a value of $p = 0.4$? Why or why not?