Midterm 2

2023-05-24

Instruction: Please read the problems carefully. Write down your choice next to each problem, and then copy it to the final page chart.

Write your Student name and ID below:

Part 1: T/F Problems (3 * 12 = 36)

1. A p-value of 0.05 means there is a 5% chance that the null hypothesis is true.
2. For a dataset listing the students' ID and students' letter grade, we can use `geom_histogram` in `ggplot2` package to visualize the grade distribution of the students.
3. In the dataset `nycflights` in package `nycflights13`, command `dim(flights)` give you the result: `[1] 336776 19`, based on the result we should use `geom_hex` to visualize the relationship between departure delay and arrival delay.
4. For a sample $x_1, x_2, ..., x_n$ draw from an unknown distribution X, $\frac{1}{n}\sum_{i=1}^{n} x_i$, $\min(x)$, x_1 are all sample statistics.
5. A Type I error occurs when we reject an alternative hypothesis that is actually true.
6. The null hypothesis is always the hypothesis we want to prove.
7. You calculate the sample correlation on the variable of study hours per week and the variable of final exam score. The result correlation is 0.7653. So you proved study hard will cause students to get higher final exam score.
8. In homework 5 we used the Monte Carlo simulation to calculate complex function integration. Due to the central limit theorem (CLT), the more points we simulate, the more precise the integration result will be.
9. `for` loop, `sapply`, and `replicate` can be used to repetitively execute some functions.
10. We use pipes to add additional layers (such as labels, texts, lines) when we using `ggplot2` to generate plots.
11. The Poisson distribution can be used to model the number of defects in a manufacturing process. For this Poisson distribution, its mean is equal to its variance.
12. When we make inference about the binomial distribution $Y \sim Binomial(n, p)$, there is only 1 unknown parameter.

Part 2: Multiple Choice Problems (4 * 12 = 48)

13. Two dice are rolled. What is the probability of getting a sum of 7 or 10?

A. 1/6

B. 1/3

C. 1/4

D. 1/12

14. For the following 4 statements, which one is **incorrect**?

A. The Bayes formula states that $P(A \mid B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$

B. For two events A and B are independent, where $P(A) > 0$ and $P(B) > 0$, then $P(A \cup B) = P(A) + P(B)(1 - P(A))$

C. For two events A and B where $P(A) > 0$ and $P(B) > 0$, $P(A \cap B^C) = P(A) - P(A \cap B)$.

D. For any random variable X, the variance of X is larger than the variance of Y where $Y = aX$ and $a < 0$.

15. A weather forecaster predicts whether or not it will rain each day. Out of the rainy days, they correctly predict the weather 80% of the time. Out of the non-rainy days, they correctly predict the weather 90% of the time. Assume it rains 10% of the time. Every day is independent. What is the probability of a correct prediction?

A. 0.87

B. 0.88

C. 0.89

D. 0.90

16. For the following 4 statements, which one is **incorrect**?

A. The Central Limit Theorem states that if you take a random sample from a population and your sample size is at least 30, then the distribution of your population is normally distributed.

B. For a normal random variable, the z-score of a value x that is below the mean will always be negative.

C. You flip a coin until you get 4 heads. The distribution of the number of tail is **not** a binomial distribution.

D. We can use Poisson distribution to model the highway speed tickets per month on I-80 in Yolo County.

17. Suppose the voltage of batteries follow a Normal Distribution. Batteries of type A have a mean voltage of 6.0 and a standard deviation of 0.3. Type B have a mean voltage of 12 and a standard deviation of 0.4. Assume battery voltage is independent. What's the probability that the combined voltage exceeds 17.7?

A. `pnorm(17.7,18,0.5) = 0.2743`

B. `1 - pnorm(17.7,18,0.5) = 0.7257`

C. `pnorm(17.7,18,0.25) = 0.1151`

D. `1 - pnorm(17.7,18,0.25) = 0.8849`

18. According to the Law of Large Numbers, as the sample size increases:

A. The probability of an event occurring approaches zero.

B. The variability of the sample decreases.

C. The sample mean approaches the population mean.

D. The sample becomes less representative of the population.

19. A 95% confidence interval is constructed for a population mean. Which of the following statements correctly interprets this interval?

A. There is a 95% probability that the true population mean falls within the interval.

B. The sample mean has a 95% chance of being within the interval.

C. When we repeatedly construct 100 confidence intervals (including sample from the population), on average 95 intervals will contain the true population mean.

D. The confidence interval contains 95% of the sample data.

20. Jane has just begun her new job as part of the sales force of a very competitive company. In a sample of 16 sales calls it was found that she closed the contract for an average value of 108 dollars with a standard deviation of 12 dollars. Company policy requires that new members of the sales force must exceed an average of $100 per contract during the trial employment period. Test at 5% significance that whether Jane satisfies the company policy, assuming the population of the contract is normally distributed. What is the correct code and conclusion?

A. `pt((108 - 100) / (12 / sqrt(16)), df = 16) > 0.05`, so we fail to reject the null hypothesis.

B. `1 - pt((108 - 100) / (12 / sqrt(16)), df = 16) < 0.05`, so we reject the null hypothesis.

C. `pt((108 - 100) / (12 / sqrt(16)), df = 15) > 0.05`, so we fail to reject the null hypothesis.

D. `1 - pt((108 - 100) / (12 / sqrt(16)), df = 15) < 0.05`, so we reject the null hypothesis.

21. Monte Carlo simulation is a technique used to:

A. Estimate probabilities through repeated random sampling.

B. Study the relationship between two random variables.

C. Predict future events with high accuracy.

D. Perform statistical analysis on large datasets.

22. Which of the following statements is **incorrect**?

A. For hypothesis test problems, we assume population parameter unknown and sample statistics known.

B. When the sample size is big, by the central limit theorem we can use the Z score even though the population distribution is right skewed.

C. In a hypothesis test, fail to reject the null hypothesis means that the alternative hypothesis is proven to be false.

D. Type II error occurs when the null hypothesis is rejected, but it is actually true.

23. Which function we should use when we want to generate a map in midterm 1?

A. geom_bar()

B. geom_polygon()

C. geom_line()

D. geom_point()

24. Which type of the problem is not part of midterm 1?

A. Use binomial distribution to simulate a chicken farm.

B. Use `tidyverse` package functions to manipulate datasets.

C. Calculate the "robust" sample statistics by excluding the outliers

D. Problem using Combination about salad toppings at Pluto's

Part 3: Code sorting problem (16)

1. Sort the code that analysis covid data in midterm 1.

A.

```
cv_states_20210428 %>%
  ggplot(aes(x = pop_density, y = per100k,
             label = abb)) +
  geom_point(aes(size = population, color = naive_CFR))
```

B.

```
cv_states = as.data.frame(read.csv(
"https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv"))
# Assume the data is already processed
```

C.

```
cv_states = cv_states %>% mutate(per100k = cases / 100000) %>%
mutate(naive_CFR = (deaths*100/cases) )
```

D.

```
library(tidyverse); library(ggplot2)
```

E.

```
cv_states_20210428 = filter(cv_states, date=="2021-04-28")
```

Copy your solution here for grading

1-12											
13-24											

Code sorting: