Sepehr Akbari

May 4th, 2024

Final Exam: Takehome

Problem 1

Loading the Data Set:

```
library(readxl)
carseats_sample <- read_excel("/Users/sepehrakbari/Documents/LFC/Semester 2/MATH 150/DSs/cars
#View(carseats_sample)
```

(a) Is the US variable quantitative or qualitative? What is the level of measurement? Briefly explain your answers.

Answer:

The `US` variable is a **qualitative** data point with a **nominal** level of measurement. This means it describes categories rather than quantities, with the values 0 or 1 acting as labels for those categories ("US" or "not US"). There's no inherent order or meaning to the values themselves; "1" doesn't represent a "greater US" than "0", they simply indicate membership in distinct groups).

(b) Compute the five-number summary and inter-quartile range of competitor prices at the stores in this set.

Answer:

```
fivenum_summary <- fivenum(carseats_sample$CompPrice)
cat("Five-Number summary:",fivenum_summary)
```

```
Five-Number summary: 77 115.5 127 136.5 161
```

```
IQR <- 136.5 - 115.5 # Q3 - Q1
cat("Inter-quartile range:",IQR)
```

```
Inter-quartile range: 21
```

```
bound <- IQR * 1.5

lower <- 115.5 - bound # Q1 - IQR*1.5
upper <- 136.5 + bound # Q3 + IQR*1.5

cat("Any value lower than",lower,"and higher than",upper,"can be considered an outlier.")
```

```
Any value lower than 84 and higher than 168 can be considered an outlier.
```

(c) Determine the 60th percentile of carseat prices (not competitor prices) in this data set.

Answer:

```
quantile(carseats_sample$Price, 0.6)
```

```
  60% 
122.4 
```

(d) Compute a level 95% confidence interval for the mean price of carseats (not competitor price) assuming that the population standard deviation is $25. Briefly explain your choice of method. Identify the point estimate, margin or error, and endpoints of the interval.

Answer:

As we are working with the population standard deviation, rather than the sample standard deviation, we will use z-scores instead of t-scores. Standard deviation = σ.

```
size <- length(carseats_sample$Price)
mean <- mean(carseats_sample$Price)
stndDev <- 25

z <- round(qnorm(0.975), 2) ## 1.960
margin_of_error <- (z * (stndDev / sqrt(size)))
lower_bound <- mean - margin_of_error
upper_bound <- mean + margin_of_error

# Point Estimate
cat("Point Estimate:",mean,"\n")
```

```
Point Estimate: 115.5067
```

```
# Margin of Error
cat("Margin of Error:",margin_of_error,"\n")
```

```
Margin of Error: 5.658033
```

```
# Endpoints of the Interval
cat("The price will be between",round(lower_bound,2),"and",
    round(upper_bound,2),"dollars.")
```

```
The price will be between 109.85 and 121.16 dollars.
```

Problem 2

Loading the Data Set:

```
library(readxl)
jumping <- read_excel("/Users/sepehrakbari/Documents/LFC/Semester 2/MATH 150/DSs/jumping.xls
#View(jumping)
```

(a) In just a sentence or two, explain the circumstances under which it would be appropriate to calculate the correlation coefficient of these two variables. No R code is needed for this part.

Answer:

It would be appropriate to calculate the correlation coefficient of height and jump distance in this experiment if we are looking to see if there is a linear relationship between one's height and how far they can jump. The correlation coefficient measures the strength and direction of a linear relationship between two variables. In this case, it would tell you whether taller people tend to jump further, shorter people tend to jump further, or there is no relationship at all.

It's important to note that correlation doesn't equal causation. Even if we find a correlation between height and distance, it doesn't necessarily mean that being tall causes us to jump further. There might be other factors at play, such as strength or jumping technique.

(b) Assume the conditions in part (b) are met and determine the correlation of these two variables.

Answer:

```r
cor(jumping$height, jumping$distance)
```

```
[1] 0.5902164
```

(c) Find the equation of the least-squares regression line. Use `height` as the explanatory variable.

Answer:

```r
model <- lm(distance ~ height, data = jumping)
slope <- coef(model)[2]
intercept <- coef(model)[1]

cat("The equation of the least-squares regression line is: \n jump_distance = "
    ,slope,"*","height",intercept,"\n\n")
```

```
The equation of the least-squares regression line is:
 jump_distance =  3.643357 * height -365.5874
```

```r
cat("or in standard form: 'y ="
    ,round(slope,2),"* x + (",round(intercept,2),")'\n")
```

```
or in standard form: 'y = 3.64 * x + ( -365.59 )'
```

(d) What is the predicted jump distance of a child with height 150cm? If this calculation isn't appropriate, briefly explain why.

Answer:

```r
(slope * 150) + intercept
```

```
  height
180.9161
```

This calculation is indeed not appropriate because of extrapolation! In the data set we have data from `height` from 130 to 138, so 150 would not be stretching the data (or our luck) too much in simpler words.

(e) What is the residual of the child with height 135cm? Briefly interpret this number in ordinary human language.

Answer:

```
# observed_value - expected_value
150 - (slope * 135 + intercept)
```

```
  height
23.73427
```

This means the observed value is about 23.7cm higher than what the model predicted. So we can say the model **underestimated** the actual value.

Problem 3

A hospital emergency room classifies incoming patients as either high, medium, or low priority, hopefully with equal proportions. As part of an internal audit, a random sample of 140 patients is collected. The results are as follows.

- 49 patients were classified as high priority
- 63 patients were classified as medium priority
- 28 patients were classified as low priority

Is the hospital in alignment with its own standard? Test at significance level $\alpha = 0.05$. Use **both** of the methods covered in class and follow all of the best practices we have established. Make sure your process is clear!

Answer:

1- Laying down the hypotheses:

Null Hypotheses (H_0): The number of incoming patients in different priority levels are equally destributed (1/3 each).

Alternative Hypotheses (H_1): The number of incoming patients in different priority levels are NOT equally distributed.

2- Calculating P-Value:

```
# Sample Size
size <- 140

# expected values based on H0
expected_value <- c(140/3, 140/3, 140/3)
# observed values based on data collection
observed_value <- c(49, 63, 28)
```

```
# degrees of freedom
df <- length(expected_value) - 1
cat("Degrees of Freedom (df):",df,"\n")
```

```
Degrees of Freedom (df): 2
```

```
# X-squared statistic
x2 <- sum((observed_value - expected_value) ^ 2 / expected_value)
cat("X-squared:",x2,"\n")
```

```
X-squared: 13.3
```

```
# P-Value
p_value <- 1 - pchisq(x2, df)
cat("p-value:",p_value,"\n")
```

```
p-value: 0.001294022
```

3- Confirming Result:

```
# Validation
chisq.test(observed_value)
```

```

	Chi-squared test for given probabilities

data:  observed_value
X-squared = 13.3, df = 2, p-value = 0.001294
```

4- Testing the significance level ($\alpha = 0.05$):

```
if (p_value < 0.05){
  cat("Reject Null Hypotheses")
} else {
  cat("Accept Null Hypotheses")
}
```

```
Reject Null Hypotheses
```

5- Conclusion:

As the p-value is less than alpha (our significance level), it means there's enough evidence to suggest the null hypothesis is unlikely to be true.