Extended Qustions

1.(a) Consider two lab classes of the same size for DATA1001. In class A, the passing rate for Project 1 is 20% and 10% for women and men respectively, and in class B, the rates are 50% and 40% respectively. It is claimed:“The combined passing rate across the two classes must be higher for women than it is for men.” Comment.

Answer:

Simpson’s Paradox - The claim may be right or wrong. - The word ‘must’ is the mistake. - It is possible that the trend reverses when the data from the 2 Lab classes is pooled together (Simpson’s Paradox).

1.(b) Given quantitative data on air quality from 2 measuring stations (in the Central Coast and the Illawarra), without code propose one way that you could construct a clustered barchart. Sketch an example.

Answer:

Clustered Barchart on quantitative data - As AQI is currently quantitative data , it would need to be summarised as qualitative data, eg by dividing into “high” and “low” days of air quality. - Sketch: Eg divide by the 2 locations (CE and NW) on the x-axis, and then within the plot, divide each into 2 colours for “high” and “low”.

1.(c) Explain the significance of Anscombe’s Quartet.

Answer:

Anscombe’s Quartet Anscombe’s Quartet is informative for warning against mistakes in linear regression interpretation. Provide example: A high value of r might not indicate a linear trend.

1.(d) A standard pack of 52 cards has one queen of spades.The pack is shuffled, and then five cards are dealt off the top of the pack. Find the chance that the 5th card dealt is the queen of spades. Justify your answer.

Answer:

Chance - In the shuffle, the queen of spades has to land in one of the 52 card positions. - So the chance that it lands in the 5th position (hence 5th card dealt) is 1/52.

1.(e) A company finds that on average their employees have 10 ‘sick days’ per year. They hope to reduce the number of sick days, by introducing more flexible working arrangements. They select a simple random sample of 100 employees and find after introducing the new arrangements, that those employees had on average 9 ‘sick days’ that year, with a sample SD of 5. Formulate a hypothesis and test using a box model.

```
pt(-2,99)
## [1] 0.02411985
```

Answer: Hypothesis Test Hypotheses: Ho: The mean number of sick days = 10 (μ=10) H1: The mean number of sick days is reduced (μ<10) [1 sided] A box model represents the null hypothesis. Mean of Box = 10; SD of box is unknown; OV of sample = 9. After 100 draws with replacement, EV = 10 ; SE = 5/sqrt(100) = 0.5 (approximation as use sample SD) Hence, the test statistic = (9-10)/0.5 = -2 [compare to t with 99 df] The p-value is 0.025 (approx) Hence, there is evidence against Ho (assuming 0.05 significance level). Hence, the number of sicks days seems to have reduced, giving evidence that flexible working arrangements are helpful.

2. Spotify is a popular music streaming platform that allows users to listen to music on their devices. Kahn is having a 21st party and wants to investigate what music he should play for his guests. He downloads the data set spotify from Kaggle.com, which is a public data platform that is owned by Google. The data was scraped from the Spotify API wrapper in November 2018.

```
dim(spotify)
## [1] 116372 17
head(spotify,2)
## artist_name track_id
## 1 YG 2RM4jf1Xa9zPgMGRDiht8O
## 2 YG 1tHDG53xJNGsItRA3vfVgs
## track_name acousticness danceability
## 1 Big Bank feat. 2 Chainz, Big Sean, Nicki Minaj 0.00582 0.743
## 2 BAND DRUM (feat. A$AP Rocky) 0.02440 0.846
## duration_ms energy instrumentalness key liveness loudness mode
## 1 238373 0.339 0 1 0.0812 -7.678 1
## 2 214800 0.557 0 8 0.2860 -7.259 1
## speechiness tempo time_signature valence popularity
## 1 0.409 203.927 4 0.118 44
## 2 0.457 159.009 4 0.371 10
```

(a)

(i) How many songs are in the data set? **Answer**: 116372

(ii) Outline one possible limitation with using this data. *Answer_*:

Examples Conflict of interest for Spotify releasing data (eg is any data missing or changed?). Undefined data dictionary: What are the algorithms that determine ‘valance’ etc?

(b) Kahn is interested in the average length of songs on Spotify.

(i) What type of variable is duration ?: **Answer**: 116372

```
class(spotify$duration)
## [1] "integer"
spotify$duration = spotify$duration/(60*1000) # convert to minutes
```

(ii) Give 3 observations from the following summaries.

summary(spotify$duration)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.05338 2.73415 3.36288 3.54244 4.00448 93.50033

boxplot(spotify$duration, horizontal =T)

Answer:

Quantitative variable (or integer, numeric). Examples of observations in context: The boxplot is highly right skewed, indicating lots of anomalies in terms of long songs. The mean/median are similar (3.4/3.5 minutes) with small IQR indicating many songs are a similar length. There is a huge range between the minimum song of 0.05 minutes & maximum of 93.5mins.

(iii)
Kahn wonders whether the mode of the song (major or minor) affects the average length of songs on Spotify. What does he discover?

```
boxplot(spotify$duration ~ spotify$mode, horizontal =T) # 0 = minor; 1= major
```

4.(a) It is claimed that songs in a minor key sound more ‘sad’. Test this claim using the spotify1 data, where a high score of valence indicates a ‘happier’ song. Use α = 0.

```
##
## Welch Two Sample t-test
##
## data: spotify1$valence by spotify1$mode
## t = 0.2404, df = 71.292, p-value = 0.8107
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.09065477 0.11551293
## sample estimates:
## mean in group 0 mean in group 1
## 0.4522581 0.4398290
```

Answer: Welch 2 sample t-test Ho:Mean(Major) = Mean(Minor) vs 2 sided alternative

A: The 2 populations (modes) are approximately Normally distributed (as suggested by boxplots; minor boxplot is a bit skewed) and the 2 samples are independent. So both assumptions seems reasonable.

[Note to marker: Welch 2 sample t-test does not require equal SD.]

T: Using the R Output, the observed value of the Welch 2 sample t-test is 0.2404. This is compared to a t distribution with df = 71.292 df.

P: The p-value is very large (0.8107, from R Output).

C:

Hence, given the significance level of 0.05, the data is consistent with Ho. Hence, we conclude that songs in a minor key do not sound more `sad'. Alternative method: Correct explanation involving the CI.

(b) It is claimed that there are an equal number of Spotify songs in each of the 12 keys. Test this claim using the spotify1 data. Use α = 0.01.

Answer:

Ho: Proportions of songs in each of the 12 keys is 1/12 vs not Ho.

A: We assume there are more than 5 songs in each of the 12 keys (as suggested by the boxplots in 2(e) for valance, and large sample size 2(a)).

T: Using the R Output, the observed value of the chi-squared test is 29.84. This is compared to a chi-squared distribution with df = 12-1 = 11 df.

P: The p-value is very small (0.001679, from R Output).

C:

Hence, given the significance level of 0.01, there is strong evidence against Ho. Hence, we conclude that there is not an equal amount of songs in the 12 keys.

(c) In your own words, explain what a p-value is, its role in hypothesis testing, and any potential issues.

Answer:

A p-value is the chance of observing the test statistic (or something more extreme relative to H1), if H0 is true. It’s role in hypothesis testing is to see how the data fits with the null hypothesis. A possible issue is p-hackingm, with the significance level is not decided prior to the hypothesis test.