Quiz Week 1 - Sample Statistics - Questions

FIT5197 teaching team

Unit context in brief

Statistical inference refers to inferring the values of statistics (e.g. the mean) that describe data, the parameters of statistical models, or the values of variables given the values of other variables. Statistical inference enables us to make predictions using statistical models.

In this unit we consider the Frequentist (i.e. counting) approach to statistical inference and computing the probability/likelihood of the data y given the parameters θ, $P(y|\theta)$. Under this approach we use Maximum Likelihood Estimation (MLE) to estimate the parameters of probabilistic/likelihood models, i.e. $\hat{\theta} = argmax_{\theta} P(y|\theta)$. If we are dealing with a known probability distribution of the data then we use $P(y|\theta)$, but in general we are interested in the probability $P(y|x, u, \theta)$ where u can be 'predictor' variables and $x = x(u, \theta)$ can be intermediate variables dependent on u and θ. In this unit we primarily focus on the cases of $P(y|\theta)$ and $P(y|u = x, \theta) = P(y|x, \theta)$.

The above will become clearer as we go through the unit.

An alternative view, not covered in the unit, to the Frequentist inference approach is the Bayesian probability approach where we start with a distribution of $P(\theta)$, then apply Bayes theorem to obtain maximum a posteriori (MAP) estimates, i.e. $\hat{\theta} = argmax_{\theta} P(\theta|y)$.

What you need to know about sample statistics

For this Week 1 lecture you need to learn how to compute sample statistics using the formula in section 1 Sample Statistics in the unit formula sheet. This sheet also shows the Tukey boxplot defintion. Applying these formulas and the Tukey boxplot is examinable! Use the formula sheet to answer the questions below.

Question 1

Consider this sample of heights: Height $\in \{173, 160, 162, 172\}$

What is the Median height?

Question 2

What is the sample variance and sample mean?

Question 3

Consider a new set of heights: heights $\in \{160, 162, 172, 173, 200\}$.

What are the outliers? Use the Tukey boxplot definitions of upper and lower inner fence, where

$$\begin{aligned}
upper\ inner\ fence &= Q_3 + 1.5 \times IQR \\
lower\ inner\ fence &= Q_1 - 1.5 \times IQR \\
Q_3 &= 75th\ percentile \\
Q_1 &= 25th\ percentile \\
Q_2 &= Median \\
IQR &= Q_3 - Q_1
\end{aligned}$$

R hackers all or nothing challenge

Create a function to generate a scatter plot of weight versus height colored by group gender for the given data. Label axes clearly and use a legend.

In [1]:

```
df <- data.frame(height = c(173,160,161,160,188,170,162,179,165,172,159,190),
                 weight = c(66,55,56,54,97,60,59,79,64,77,52,102),
                 job = as.factor(c("construction", "construction", "police", "announcer", "announcer","announcer", "studer
                 hand = as.factor(c("R", "R", "L","R", "R", "R", "L","R", "L","R", "R", "R")),
                 gender = as.factor(c("male","female","female","female","female","male","female",
                                      "male","female","female","female","male"))
                 )
```

Compute sample mean, sample standard deviation, correlation, and box plot summary statistics (Minimum,1st Quartile, Median, Mean, 3rd Quartile and Maximum) for height and weight.

Plot boxplots of height and weight in the same figure. Label axes clearly and provide a title.