Class 5 notes and code

Part 1: z-scores

LeBron James is a one of the greatest basketball player of all time. The table below shows some of his statistics in the 2011 NBA season, along with the league average (mean) and standard deviation of these statistics.

Using the data in the table, calculate the z-score for each statistic to determine which statistic was most impressive relative to his peers.

statistic	LeBron James	League mean	League sd
Field goal %	0.510	0.464	0.053
Points scored	2111	994	414
Assists	554	220	170
Steals	124	68.2	31.5

Answer

Part 2: Quantiles

The P^{th} percentile is the value of a quantitative variable which is greater than P percent of the data.

We can calculate the percentiles of a sample of data using the `quantile()` function.

The code loads information about the winner for the best actor and best atress Oscars award winners from 1929 to 2019, into vectors called `best_actor_age` and `best_actress_age`.

Let's calculate the age at the 20th, 50th, and 80th percentiles for the best actor award winners.

```r
# Load the data
load("oscars.rda")

# create the vectors that have the ages of the best actor and best actress award winners
best_actor_age <- subset(oscars, award == "Best actor")$age
best_actress_age <- subset(oscars, award == "Best actress")$age


# Calculate the 20th, 50th, and 80th percentiles for best actor ages
```

How can we interpret these results?

Part 3: Five number summary and boxplots

A five number summary consists of the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum of a quantitative variable.

We can calculate the five number summary using the `fivenum()` function.

We can also visualize the five number summary using a boxplot. We can create boxplots using the `boxplot()` function.

Let's create side-by-side boxplots comparing the ages of best actor and best actress award winners.

If you look at the boxplot of the ages of best actor you will notice that there is an outlier in the data. Is this outlier due to an error in the data or is it a valid data point?

Part 4: If there is time - additional practice

The code below loads 4 vectors of data called `x1`, `x2`, `x3`, and `x4`.

Please do the following:

1. Create 4 histograms of showing the distribution of each of the 4 vectors of data.
2. Create a side-by-side boxplot comparing the 4 vectors of data.

Can you tell which boxplot corresponds to which histogram?

```
# Load the data
load("hist_vs_boxplot.Rda")

# Create the histograms
#par(mfrow = c(2, 2))  # Set up a 2x2 plotting area









# Reset plotting area

# Create the side-by-side boxplot
#par(mfrow = c(1, 1))  # Reset to a 1x1 plotting area
```