4 — Variability

4.1 Box Plots and the IQR

A box plot is a great way to show the 5 number summary of a data set in a visually appealing way. The 5 number summary consists of the minimum, first quartile, median, third quartile, and the maximum

Definition 4.1 — Interquartile range. The Interquartile range (IQR) is the distance between the 1st quartile and 3rd quartile and gives us the range of the middle 50% of our data. The IQR is easily found by computing: $Q3 - Q1$

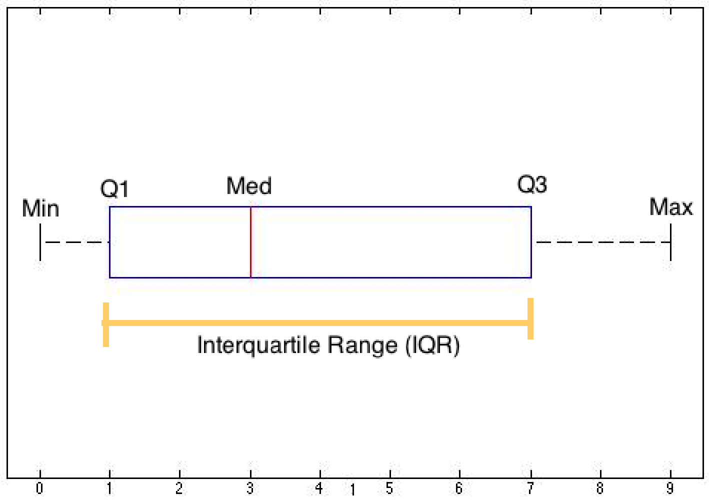

Figure 4.1: A simple boxplot

4.1.1 Finding outliers

Definition 4.2 — How to identify outliers. You can use the IQR to identify outliers:

1. Upper outliers: $Q3 + 1.5 \cdot IQR$
2. Lower outliers: $Q1 - 1.5 \cdot IQR$

4.2 Variance and Standard Deviation

Definition 4.3 — Variance. The variance is the average of the squared differences from the mean. The formula for computing variance is:

$$\sigma^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}$$

Definition 4.4 — Standard Deviation. The standard deviation is the square root of the variance and is used to measure distance from the mean.

R In a normal distribution 65% of the data lies within 1 standard deviation from the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.

4.2.1 Bessel's Correction

Definition 4.5 — Bessel's Correction. Corrects the bias in the estimation of the population variance, and some (but not all) of the bias in the estimation of the population standard deviation. To apply Bessel's correction we multiply the variance by $\frac{n}{n-1}$.

R Use Bessel's correction primarily to estimate the population standard deviation.

4.3 Practice Problems

Problem 4.1 Make a box plot of the following monthly incomes

\$2500	\$3000	\$2900
\$2650	\$3225	\$2700
\$2740	\$3000	\$3400
\$2500	\$3100	\$2700

Table 4.1: Incomes

Problem 4.2 Find the standard deviation of the incomes.

Problem 4.3 What is a better descriptor of the distribution the box plot, or the mean and standard deviation? Why?