Stat 381: Applied Statistical Methods I **Spring 2015**

January 12, 14

Instructor: Brian Powers

1.1 Statistical Inference

Definition 1.1. A **population** is a collection of all individuals or individual items of a particular type. A **sample** is a subset of a population. The **sample size**, often given the notation n, is simply the number of elements, individuals or observations in a sample.

Definition 1.2. A **simple random sample** is any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.

Definition 1.3. Statistical inference is the process of making conclusions about a population through sampling.

Relationship between Probability and Statistical Inference:
The sample along with inferential statistics allows us to draw conclusions about the population, with inferential statistics making clear use of elements of probability.

Elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population.

1.2 Representation of Data

1.2.1 Data Table

Example 1.4. Two samples of seedlings of red oaks are planted, some treated with nitrogen and others without. The stem weights in grams are given in the following table.

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

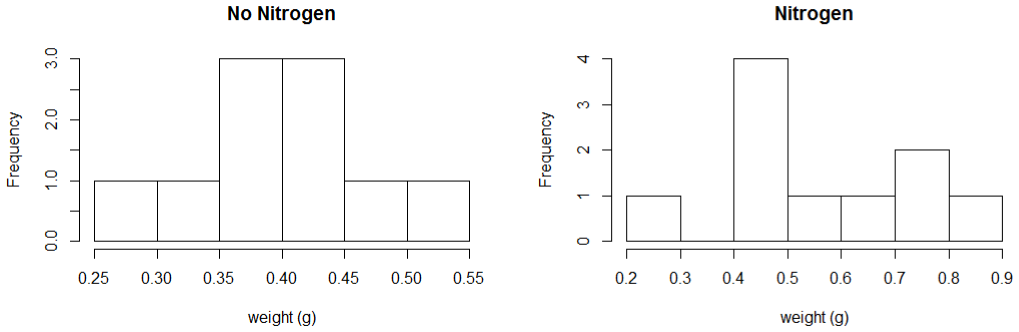
1.2.2 Stem and Leaf Plot

No Nitrogen	
0.2	8
0.3	2
0.3	678
0.4	233
0.4	7
0.5	3

Nitrogen	
0.2	6
0.3	
0.4	3679
0.5	2
0.6	2
0.7	59
0.8	6

1.2.3 Histogram

Not the same as a bar chart. The number of **bins** or **classes** should be roughly $\sqrt{n}$ as a rule of thumb, but in general should be big enough and small enough to give a good sense of the shape of the data. The y-axis may be the **frequency** or **relative frequency** depending, depending on which is more useful.



1.3 Statistics

Definition 1.5. A **statistic** is a single measure of a sample, a function of sample data.

1.3.1 Descriptive Statistics

Measures of the location or **central tendency** of a sample of data, $x_1, x_2, \ldots, x_n$, where data values are assumed to be sorted in increasing order:

Definition 1.6. The **sample mean**

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The sample mean can also be thought of as the centroid of the sample.

Example 1.7. For the no nitrogen data sample, the sample mean is

$$\bar{x} = \frac{.32 + .53 + .28 + .37 + .47 + .43 + .36 + .42 + .38 + .43}{10} = .399$$

Definition 1.8. The **sample median**

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{x/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

Example 1.9. The median for the no nitrogen data is $(.38 + .42)/2 = .4$

The sample mean is sensitive to extreme data values in the sample, whereas the median is not affected.

Definition 1.10. The **mode** of a sample of data is the most common value.

Definition 1.11. The X% **trimmed mean** is calculated after removing the top and bottom X% of the sample.

Measures of the variability of a sample include

Definition 1.12. Quantiles of a data sample are values which create intervals which contain equal (or roughly equal) number of observations. The **quartiles** are the values which divide the data into four subgroups of roughly equal size.

The median is the second quartile (Q_2). The first quartile (Q_1) is the median of the lower half of the data (all observations less than the median). The third quartile (Q_3) is the median of the upper half of the data.

Definition 1.13. The **minimum** ($X_{(1)}$ or X_{min}) is the lowest value among a sample, the **maximum** ($X_{(n)}$ or X_{max}) is the highest value in the sample. The minimum is sometimes referred to as Q_0 and the maximum as Q_5.

Definition 1.14. 5-number summary is the ordered set of statistics: $X_{(1)}, Q_1, \tilde{X}, Q_3, X_{(n)}$.

Example 1.15. The 5-number summary for the no nitrogen sample data is $(0.28, 0.3625, 0.4, 0.43, 0.53)$

Definition 1.16. The **range** of the data is $X_{(n)} - X_{(1)}$. The **inter-quartile range (IQR)** is $Q_3 - Q_1$.

Example 1.17. The range of the no nitrogen data is $.53 - .28 = .25$. The inter-quartile range is $.43 - .3625 = .0675$

Definition 1.18. The **upper-fence** is $Q_3 + 1.5(IQR)$, the **lower-fence** is $Q_1 - 1.5(IQR)$. Any data value which is greater than the upper-fence or less than the lower-fence is considered an **outlier**

Definition 1.19. The **sample variance** is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The **sample standard deviation** $s = \sqrt{s^2}$.

Example 1.20. Consider the following data sample: 4,5,2,6,8. The sample variance can be calculated either in one big calculation or in a table. First we need $\bar{x} = 5$.

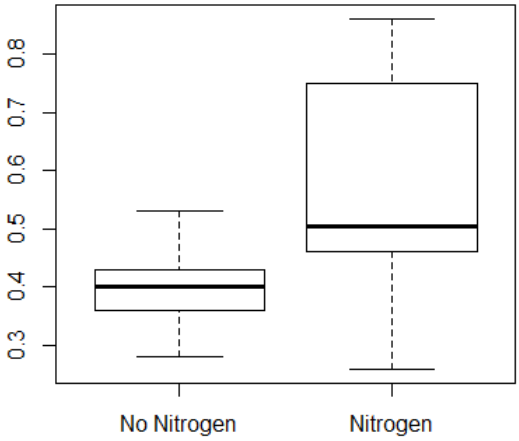
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
4	-1	1
5	0	0
2	-3	9
6	1	1
8	3	9
		20

Then $s = 20/4 = 5$.

Note that s has the same units as the data, but because the variance is the average squared deviation, its units are meaningless.

1.3.2 Boxplot

Box Q_1 to Q_3 with a line at the median. a whisker goes out to the maximum (and minimum) non-outlier, outliers marked individually (if any) with X or O.



1.4 TI-83 / 84 Tasks

1.4.1 Entering a List of Data

1.4.2 Descriptive Statistics: 1-Var Stats

1.4.3 Creating Histograms

1.4.4 Creating Box Plots