Stat 260 Lecture Notes
Set 2 - Basic Descriptive Statistics

Knowing what type of data we have affects what can be done to analyze the data. **Univariate data** is collected from single measurements on subjects.

Suppose we take a sample with n observations $x_1, x_2, x_3, \ldots, x_n$. Each x_i represents the single value that was measured in that observation. This is univariate data (i.e. we measured only one number per each observation). We can display univariate data with a **frequency table** or a **histogram** or a **boxplot**.

Example 1: Measurements of lengths of lizards captured in months of August and October.

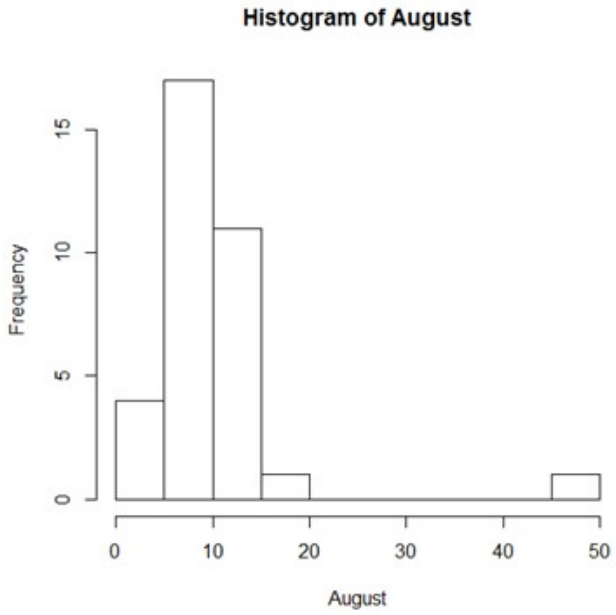
```
August measurements:
[1] 7.5 7.2 3.0 12.1 15.1 12.1 11.5 11.8 7.2 13.2 13.6 8.2 9.5 8.4 13.3
[16] 12.5 12.4 2.1 10.7 9.4 6.7 6.8 6.1 8.3 7.9 6.0 7.6 13.2 4.5 9.3
[31] 8.1 3.5 9.0 50.0


October measurements:
[1] 43.7 37.2 29.0 31.6 47.5 48.3 38.3 19.7 32.5 45.2 36.1 30.5 37.2 50.5 36.9
[16] 44.5 35.9 28.7 37.5 30.2 36.9 43.2 27.0 26.2 41.8 26.4 34.3 28.6 35.9 22.0
[31] 45.4 30.3 29.8 46.1 42.7 31.5 37.4 25.1 27.2 45.0
```

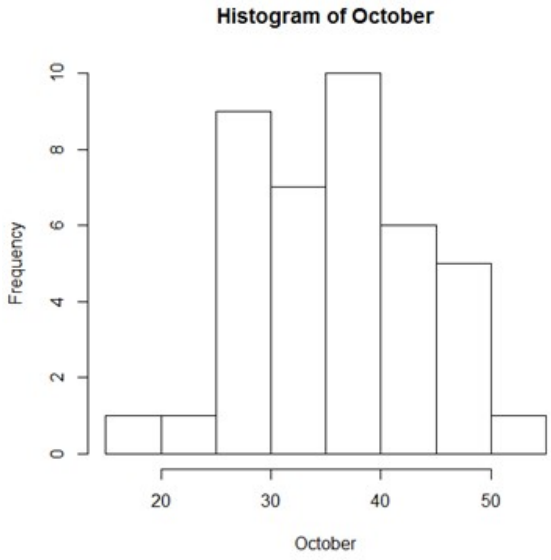
Look at the August measurements. Here we can create a frequency table.

Interval	Frequency	Relative Frequency
$[0, 5)$		
$[5, 10)$		
$[10, 15)$		
$[15, 20)$		
$[20, 25)$		
$[25, 30)$		
$[30, 35)$		
$[35, 40)$		
$[40, 45)$		
$[45, 50]$		

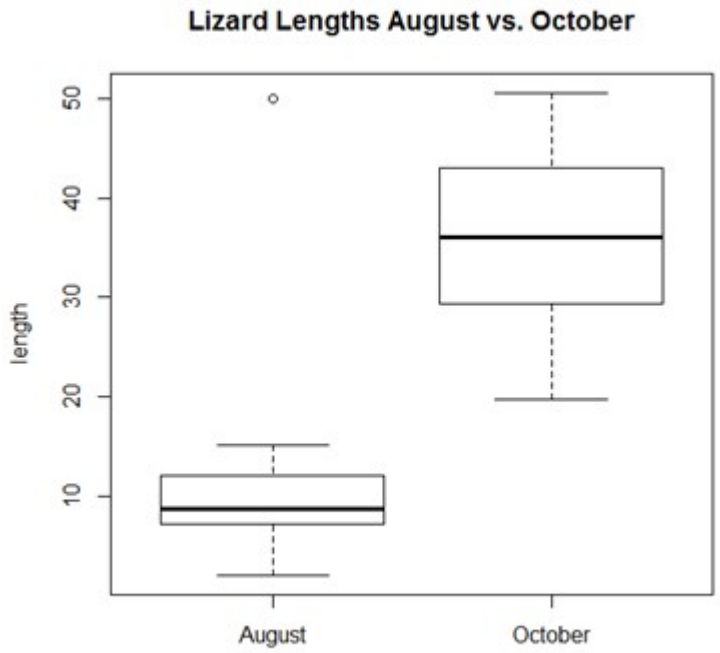
Using this frequency table we can create a histogram. The histogram for the August data looks like this:



Doing a similar thing for the October data gives a histogram that looks like this:



It can be difficult to compare two data sets like this, so we could also use a boxplot. The boxplots for the August and October data look like this:



With this boxplot it is easy to see that October lengths are larger than August lengths.

Categorical data occurs when our recorded data falls into categories. (e.g. eye colour, program major, customer satisfaction). While we cannot do many of the calculations with categorical data that we can do with univariate data, we can display categorical data with a **bar chart**.

Suppose we have a sample with observations $x_1, x_2, \ldots, x_n$.

- *sample mean*, $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$. The mean is the same thing as the average.
- *sample median*, $\widetilde{x}$, is the the middle observation. If there is an even number of observations then there are two middle observations so the median is the midpoint (here, the same as the average) of these two values. Remember that the data must be sorted first!
- *mode*, is the most common value. There may be many modes if many values are tied for occurring the most often.

Example 2: Suppose we have the sample:

$$0, 0, 2, 3, 6, 7, 10, 11, 20.$$

Calculate the mean, median, and mode.

Example 3: Suppose we have the sample:

$$1, 1, 2, 8, 15, 15, 25, 100.$$

Calculate the mean, median, and mode.

Notes:

- $\overline{x}$ doesn't have to be a value actually observed in the sample. Neither does $\widetilde{x}$. (So don't round your final answers for average or median, even if they physically aren't possible for a single observation.)
- $\overline{x}$ is sensitive to outliers (extreme values), $\widetilde{x}$ is not. Look at what happens in Example 3 if the value of 100 is removed. The mean of the new data list changes quite a bit, the median of the new data list changes very little.
- The median splits the data in two: 50% of the observations are larger than $\widetilde{x}$, and 50% of the observations are smaller than $\widetilde{x}$.
- The mean and median tell us where the "center" of our sample data is.
- The mode tells us which observation is most common. The mode can be used for categorical data, whereas $\overline{x}$ and $\widetilde{x}$ cannot be used for categorical data.
- In a histogram, the mean (average) occurs at the "balance point" of the picture.

Example 4: Look at the following two samples. How could we describe the difference in these samples?

sample 1:	**sample 2:**
10	10
20	48
49	49
50	50
51	51
80	52
90	90

Both samples have $\overline{x} = \widetilde{x} = 50$. Which sample is more spread out?

We look at *measures of variability* to describe differences between these samples.

- *sample variance*, $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
- *sample standard deviation*, $s = \sqrt{\text{variance}}$
- *shortcut formula for sample variance*, $s^2 = \dfrac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$

Note: The standard deviation s and the mean $\overline{x}$ have the same units as your original data. The variance s^2 uses (units)^2.

Example 5: Calculate the sample variance for sample 1 and sample 2 in Example 4. Calculate the sample standard deviation for sample 1 in Example 4.

Note: The **population variance** σ^2 is calculated using a slightly different formula than the sample variance. Suppose our population has N items and we take a sample of n items from it.

- *sample variance*, $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
- *population variance*, $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

In the sample, we don't know the whole picture so the sample variance s^2 is just an estimate of the population variance σ^2. As it turns out (due to technical reasons and a definition we won't cover here) that dividing by $n-1$ gives a better estimate of the population when we perform the sample variance calculation. Dividing by n gives an estimate that underestimates.

Lastly, the quickest way to calculate standard deviation, variance, and the mean is by using the stats functions on your calculator. On tests and assignments it is preferred that you use the calculator stat functions (so you do not need the formulas from Example 5).