STAT 305: Lecture 5

Chapter 3: Elementary Descriptive Statistics

Course page: imouzon.github.io/stat305

Leftovers from Lecture 4

Frequency Tables

```
      Group 1              Group 2
    74 79 77 81          65 77 78 74
    68 79 81 76          76 73 71 71
    81 80 80 78          86 81 76 89
    88 83 79 91          79 78 77 76
    79 75 74 73          72 76 75 79
```

- **Class**: A grouping of the observations
- **Frequency**: The number of observations in a class
- **Relative Frequency**: The proportion of the observations in the class
- **Cumulative Relative Frequency**: The proportion of observations in the current class or a previous class.

Summarizing

Purpose

Simple Graphs

Freq Tables

Histograms

Histograms

```
   Group 1           Group 2
74 79 77 81       65 77 78 74
68 79 81 76       76 73 71 71
81 80 80 78       86 81 76 89
88 83 79 91       79 78 77 76
79 75 74 73       72 76 75 79
```

A **histogram** is essentially a graphical representation of a frequency table.

Tips for useful frequency tables

1. Use equal class intervals
2. When the goal is to compare multiple groups, use uniform scales on each graph (i.e., keep lengths consistent)
3. Show the entire vertical axis (especially for relative frequency histograms)

Continuing

Summaries of Location and Central Tendency

Motivated by asking what is *normal/common/expected* for this data. There are three main types used:

Mean: A "fair" center value. The symbol used differs depending on whether we are dealing with a sample or population:

	Mean
Population	$\mu = \sum_1^n x_i$
Sample	$\bar{x} = \sum_1^n x_i$

Mode: The most commonly occurring data value in set.

Quantiles: The number that divides our data values so that the proportion, p, of the data values are below the number and the proportion $1 - p$ are above the number.

Median: The value dividing the data values in half (the middle of the values). The median is just the 50th quantile.

Summaries of Location and Central Tendency

```
       Group 1              Group 2
     74 79 77 81          65 77 78 74
     68 79 81 76          76 73 71 71
     81 80 80 78          86 81 76 89
     88 83 79 91          79 78 77 76
     79 75 74 73          72 76 75 79
```

Calculating Mean Think of it as an equal division of the total

- each value in the data is an "x_i" (i is a **subscript**)
- Group 1: $x_1 = 74, x_2 = 79, \ldots, x_{20} = 73$
- The sum: $x_1 + x_2 + x_3 + \ldots + x_{20}$
- divides : $(x_1 + x_2 + x_3 + \ldots + x_{20})/20$
- Or using summation notation: $\frac{1}{20} \sum_{i=1}^{20} x_i$

Summaries of Location and Central Tendency

The Quantile Function

Two useful pieces of notation:

floor: $\lfloor x \rfloor$ is the largest integer smaller than or equal to x

ceiling: $\lceil x \rceil$ is the smallest integer larger than or equal to x

Examples

- $\lfloor 55.2 \rfloor = 55$
- $\lceil 55.2 \rceil = 56$
- $\lfloor 19 \rfloor = 19$
- $\lceil 19 \rceil = 19$
- $\lceil -3.2 \rceil = -3$
- $\lfloor -3.2 \rfloor = -4$

Summaries of Location and Central Tendency

The Quantile Function

For a data set consisting of n values that when ordered are $x_1 \leq x_2 \leq \ldots \leq x_n$ and $0 \leq p \leq 1$. We define the **quantile function** $Q(p)$ as:

$$Q(p) = x_i + (n \cdot p + 0.5 - i)\,(x_{i+1} - x_i)$$

where

$$i = \lfloor n \cdot p + 0.5 \rfloor$$

(note: this is the definition used in the book - it's just written using *floor* and *ceiling* instead of in words)

Summaries of Location and Central Tendency

Example: Find the median, first quartile, 17th quantile and 65th quantile for the following set of data values:

58, 76, 66, 61, 50, 77, 67, 64, 41, 61

Summarizing

Simple Graphs

Freq Tables

Histograms

Center Stats

Spread Stats

Summaries of Variablity (or "Spread")

Motivated by asking what kind of *variability* is seen in our data or *how spread out* our observed values are.

- **Range**: the total distance the data values are spread across
- **Interquartile Range (IQR)**: the distance *the middle of data values* are spread across.
- **Variance** and **standard deviation**: measures for average distance from the center. Calculation differs depending on whether we have a population or sample:

	Variance	Standard Deviation
Population	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$
Sample	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Summarizing

Simple Graphs

Freq Tables

Histograms

Center Stats

Spread Stats

Boxplots

Boxplots

```
    Group 1            Group 2
74 79 77 81        65 77 78 74
68 79 81 76        76 73 71 71
81 80 80 78        86 81 76 89
88 83 79 91        79 78 77 76
79 75 74 73        72 76 75 79
```

A boxplot can be used to summarize the values of a single quantitative variable. It does this by making use of both many of the statistics we have discussed up to this point. It depicts both

- spread (with IQR, range, etc.)

and

- location statistics (min, median, max, etc.)

Quantile Plots:

Scatterplots using quatiles and their corresponding values

For each x_i in the data set, we plot $\left(\frac{i-.5}{n}, x_i\right)$ - meaning we are plotting $(p, Q(p))$. We connect the points with a straight line, which follows the values of $Q(p)$ exactly.

Consider the sample: 13, 15, 18, 19, 21, 34, 35, 35, 36, 39. The following table which helps create the plot:

	1	2	3	4	5	6	7	8	9	10
p										
$Q(p)$										

Quantile-Quantile Plots:

QQ plots are created by plotting the values of $Q(p)$ for a data set against values of $Q(p)$ coming from some other source.

- Empirical QQ plots: the other source are quantiles from another actual data set.
- Theoretical QQ plots: the other source are quantiles from a theoretical set - we know the quantiles without having any data.

Example

- Set 1: 36, 15, 35, 34, 18, 13, 19, 21, 39, 35
- Set 2: 37, 39, 79, 31, 69, 71, 43, 27, 73, 71

	1	2	3	4	5	6	7	8	9	10
p										
Set 1 $Q(p)$										
Set 2 $Q(p)$										

Recap

Plots and Quantiles

Boxplots

Quantile Plots

QQ Plots

Quantile-Quantile Plots:

The resulting plot shows some kind of linear pattern - this means that the quantiles increase at the same rate, even if the sizes of the values themselves are very different.