# STAT 305: Lecture 4

Amin Shirazi

2019-09-05

# STAT 305: Lecture 4

# Chapter 3

## Elementary Descriptive Statistics

Course page:
ashirazist.github.io/stat305.github.io

# Section 3.1

## Elementary Graphical and Tabular Treatment

## of

## Quantitative Data

# Summarizing Univariate Data



## Introduction: Creative Writing Workshops

Two methods of teaching a creative writing workshop are being studied for their effectiveness of improving writing skills. First, two groups of creative writing students who were randomly assigned to one of two different 3-hour workshops. At the end of the workshop, the students were given a standard creative writing test and their score on the test was recorded.

**Exam Scores for Two Groups of Students Following Different Courses**

| 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 |

## Intro

**Exam Scores for Two Groups of Students Following Different Courses**

| 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 |

We may have several questions we are interested in answering using this data. For instance,

- Which group did better on average?
- Which group has the most consistent scores?
- Were there any unusually low or high scores in either group?
- If we ignore unusual scores, which group is better?
- Which group had the most scores over 80?
- ...

However, none of these are immediately clear looking at the raw recorded data.

# Summarizing The Purpose of Summaries

## Intro

Certain questions can and should be asked across many types of experiments.

## Purpose

But seeing data in this kind of *flat* format presents lots of problems for a person trying to understand the relationship between the two groups.

**Summaries** are tools (mainly mathematical or graphical) which help researchers quickly understand the data they have collected.

The purpose of a summary is to faithfully present aspects of the data in such a way that

- we are capable of answering the types of core questions about the data asked on the previous page,
- we are able to identify more complicated aspects of the data that we may want to investigate further.

**Key Idea**: Good summaries should be quickly interpreted, provide clear insight into the data, and be widely applicable.

# Summarizing Simple Graphical Summaries



| 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 |

Simple graphical summaries aim to provide a better view of the entire set of data. The best graphs are able to make important points clearly and give valuable insights with closer study. Producing good graphs is an art.

**Two common graphical summaries**

- Dot Diagrams
- Stem and Leaf Diagrams

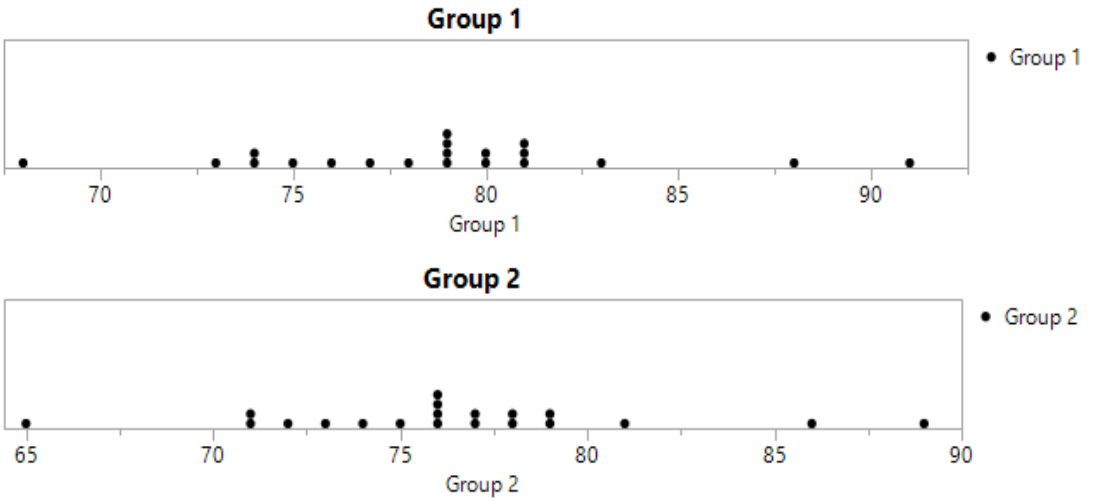
  Carries much the same visual information as a dot diagram while preserving the original values exactly

# Summarizing Simple Graphical Summaries



| 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 |

**Dot Diagrams**



# Summarizing Simple Graphical Summaries



| 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 |

**Stem and Leaf Diagrams**

**Stem and Leaf**

| Stem | Leaf | Count |
|---|---|---|
| 9 | 1 | 1 |
| 8 | 8 | 1 |
| 8 | | |
| 8 | | |
| 8 | 3 | 1 |
| 8 | 00111 | 5 |
| 7 | 89999 | 5 |
| 7 | 67 | 2 |
| 7 | 445 | 3 |
| 7 | 3 | 1 |
| 7 | | |
| 6 | 8 | 1 |

6|8 represents 68

**Stem and Leaf**

| Stem | Leaf | Count |
|---|---|---|
| 8 | 9 | 1 |
| 8 | 6 | 1 |
| 8 | | |
| 8 | | |
| 8 | 1 | 1 |
| 7 | 8899 | 4 |
| 7 | 666677 | 6 |
| 7 | 45 | 2 |
| 7 | 23 | 2 |
| 7 | 11 | 2 |
| 6 | | |
| 6 | | |
| 6 | 5 | 1 |

6|5 represents 65

# Summarizing Frequency Tables



Dot diagrams and stem-and-leaf plots are useful devices when analyazing a data set, but not commonly used in presentations and reports. In such more formal contexts, **frequency tables** and **histograms** are more often used.

A frequency table is made by

- First breaking an interval containing all the data into an apropriate number of smaller intervals of **equal length.**
- Then tally marks can be recorded to indicate the number of data points falling into each interval.
- Finally, add frequency, relative frequency and cumlative relative frequency can be added.

# Summarizing Frequency Tables



- **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.

Table 3.2
Frequency Table for Laid Gear Thrust Face Runouts

| Runout (.0001 in.) | Tally | Frequency | Relative Frequency | Cumulative Relative Frequency |
|---|---|---|---|---|
| 5–8 | III | 3 | .079 | .079 |
| 9–12 | 𝍸 𝍸 𝍸 III | 18 | .474 | .553 |
| 13–16 | 𝍸 𝍸 II | 12 | .316 | .868 |
| 17–20 | IIII | 4 | .105 | .974 |
| 21–24 | | 0 | 0 | .974 |
| 25–28 | I | 1 | .026 | 1.000 |
| | | 38 | 1.000 | |

# Summarizing Histograms



After making a frequency table, it is common to use the organization provided by the table to create a **histogram**.

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

**Tips for useful frequency tables**

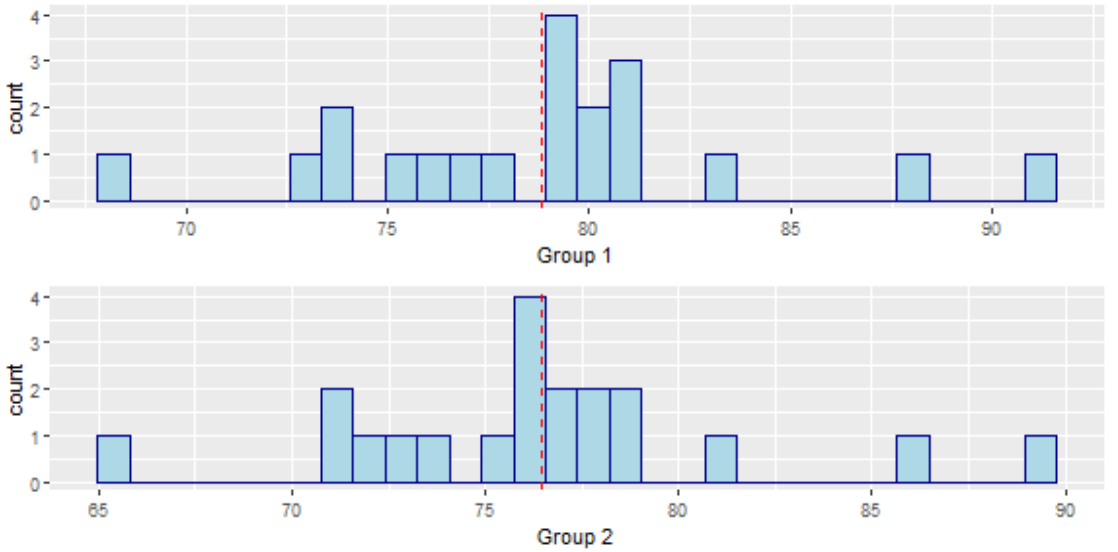
1. Use equal class intervals
2. When the goal is to compare multiple groups, use uniform scales on each graph (i.e., keep lengths consistent)
3. Show the entire vertical axis (especially for relative frequency histograms)

# Summarizing 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 |

**Unit interval**

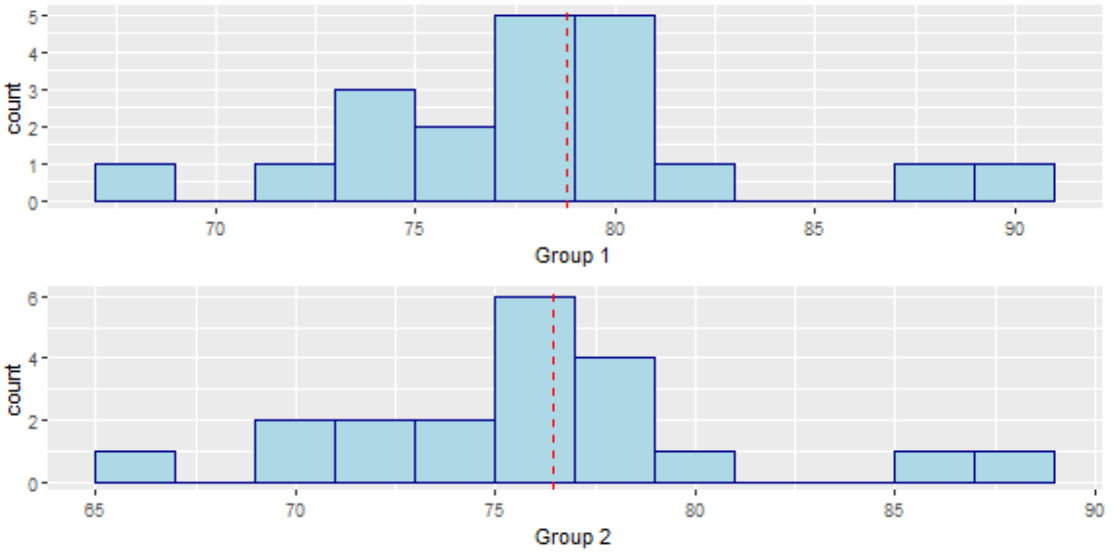


# Summarizing Histograms



| 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 |

**Interval of length two**



# Summarizing 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** | | $\mu = \frac{1}{N}\sum_1^N x_i$ |
| **Sample** | | $\bar{x} = \frac{1}{n}\sum_1^n x_i$ |

**N** is the population size and **n** is the sample size.

**Mode**: The most commonly occurring data value in set.

# Summarizing Summaries of Location and Central Tendency



**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.

**Range**: The difference between the highest and lowest values (Range = max - min)

**IQR**: The Interquartile Range, how spread out is the middle 50% (IQR = Q3 - Q1)

# Summarizing 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: $\frac{1}{20} \sum_{i=1}^{20} x_i$

# Quantiles

**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**

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

**Quantiles**

- Already familiar with the concept of "percentile".
    - e.g in the context of reporting scores on exams:

  If a person has scored at the 80th percentile, roughly 80% of those taking the exam had worse scores, and roghly 20% had better scores.

- It is more convenient to work in terms of fractions between 0 and 1 rather than percentages between 0 and 100. We then use terminology **Quantiles** rather than percentiles.

- For a number **p** between 0 and 1, the **p quantle** of a distribution is a number such that a fraction p of the distribution lies to the left of that value, and a fraction 1-p of the distribution lies to the right.

**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:

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)\,(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + .5 \end{cases}$$

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

**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

First notice that $n = 10$. It is possible helpful to set up the following table:

- **Step 1: sort the data**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| data | 41 | 50 | 58 | 61 | 61 | 64 | 66 | 67 | 76 | 77 |
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

**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

- **Step 2: find** $\frac{i-.5}{n}$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| data | 41 | 50 | 58 | 61 | 61 | 64 | 66 | 67 | 76 | 77 |
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $\frac{i-.5}{n}$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |

- **Step 3: find** $Q(p)$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| data | 41 | 50 | 58 | 61 | 61 | 64 | 66 | 67 | 76 | 77 |
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $\frac{i-.5}{n}$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)\,(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + .5 \end{cases}$$

Finding the first **quartile** ($Q(.25)$):

- $np + .5 = 10 \cdot .25 + .5 = 3.$
- since $\lfloor 3 \rfloor = 3$

  then $i = 3$ and

  $Q(.25) = x_3 = 58$

# Your turn

## Find the median

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| data | 41 | 50 | 58 | 61 | 61 | 64 | 66 | 67 | 76 | 77 |
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $\frac{i-.5}{n}$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)\,(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + .5 \end{cases}$$

Finding $Q(.17)$

- $np + .5 = 10 \cdot 0.17 + 0.5 = 2.2.$
- since $\lfloor 2.2 \rfloor = 2$ then $i = 2$ and

$$\begin{aligned} Q(.17) &= x_i + (n \cdot p - i + .5) \cdot (x_{i+1} - x_i) \\ &= x_2 + (10 \cdot 0.17 - 2 + .5) \cdot (x_{2+1} - x_2) \\ &= x_9 + (.2) \cdot (x_3 - x_2) \\ &= 50 + (.2) \cdot (58 - 50) \\ &= 51.6 \end{aligned}$$

Finding $Q(.65)$

- $np + .5 = 10 \cdot 0.65 + 0.5 = 7.$
- since $\lfloor 7 \rfloor = 7$ then $i = 7$ and

$$\begin{aligned} Q(.65) &= x_i + (n \cdot p - i + .5) \cdot (x_{i+1} - x_i) \\ &= x_7 + (10 \cdot 0.65 - 7 + .5) \cdot (x_{7+1} - x_7) \\ &= x_7 + (0) \cdot (x_8 - x_7) \\ &= x_7 + 0 \\ &= 66 \end{aligned}$$

# Section 3.2: Plots Using Quantiles

# Plots and Quantiles

## Quantile Plots

## 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** |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | | | | | | | | | | |
| $Q(p)$ | | | | | | | | | | |

# Plots and Quantiles

## Quantile Plots

## 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.

Notice that we have $n = 10$ observations which means that $Q(0.05) = x_1 = 13$. We can get the quantile for each of our observations and create this table:

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |
| $Q(p)$ | 13 | 15 | 18 | 19 | 21 | 34 | 35 | 35 | 36 | 39 |

# Plots and Quantiles

## Quantile Plots

## Quantile Plots:

**Quantile plots**

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |
| $Q(p)$ | 13 | 15 | 18 | 19 | 21 | 34 | 35 | 35 | 36 | 39 |

# Plots and Quantiles

## Quantile Plots

## Quantile Plots:

**Quantile plots**

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |
| $Q(p)$ | 13 | 15 | 18 | 19 | 21 | 34 | 35 | 35 | 36 | 39 |

# Plots and Quantiles

## Quantile Plots

## QQ Plots

### 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.

- Compare the shape of two data sets (distributions).
- Two data sets having "equal shape" is equivalent to say their quantile functions are "**linearly related**".
- If the two data sets have different sizes, the size of smaller set is used for both.
- A **QQ plot** that is linear indicates the two distributions have similar shape.
- If there are significant departures from linearity, the character of those departures reveals the ways in which the shapes differ.

## Quantile-Quantile Plots:

**Example**: How similar the two data sets are?

- 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 $Q(p)$ | | | | | | | | | | |

# Plots and Quantiles

## Quantile Plots

## QQ Plots

### Quantile-Quantile Plots:

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0.05 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 | 0.65 | 0.75 | 0.85 | 0.95 |
| Set 1 $Q(p)$ | 13 | 15 | 18 | 19 | 21 | 34 | 35 | 35 | 36 | 39 |
| Set 2 $Q(p)$ | 27 | 31 | 37 | 39 | 43 | 69 | 71 | 71 | 73 | 79 |

## Quantile-Quantile Plots:

**Interpretation**

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.

# Plots and Quantiles

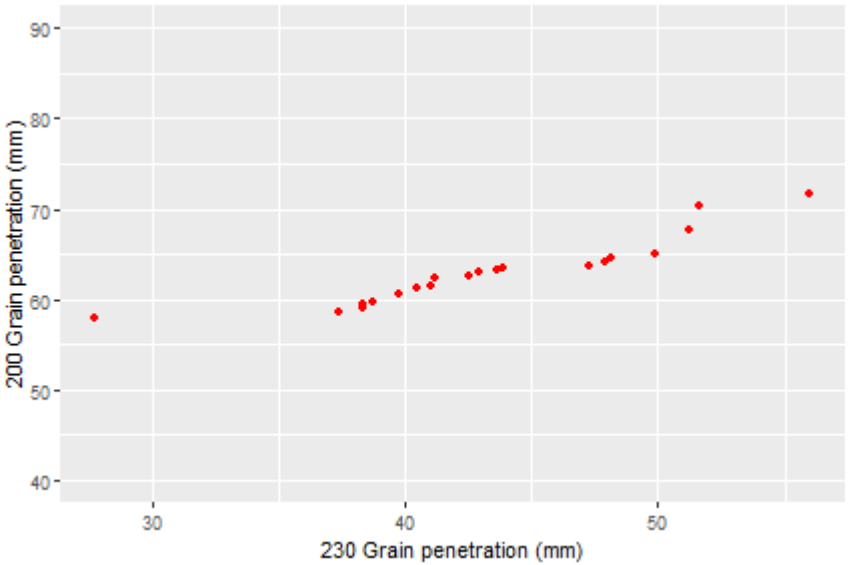
## Quantile Plots

## QQ Plots

## Quantile-Quantile Plots:

**Example 6 of chapter 3**: Bullet penetration depth

- 230 Grain penetration (mm)
- 200 Grain penetration (mm)




Except extreme lower values, it **seems** the two distributions have smiliar shapes; however, it still needs more attention to make a rough decision (consider boxplots). Might want to figure out what has caused the extereme value

## Quantile-Quantile Plots:

The idea of QQ plots is most useful when applied to one quantile function that represents data and a second that represents a **theoretical distribution**

- 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.

This allows to ask "Does the data set have a shape similar to the theoretical distribution?"

# Boxplots

# Boxplots

A simple plot making use of the first, second and third quartiles (i.e., $Q(.25)$, $Q(.5)$ and $Q(.75)$.

1. A box is drawn so that it covers the range from $Q(.25)$ up to $Q(.75)$ with a vertical line at the median.
2. Whiskers extend from the sides of the box to the furthest points within 1.5 IQR of the box edges
3. Any points beyond the whiskers are plotted on their own.

**Example**: Draw boxplots for the groups using quantile function

| Group 1 | | | |
|---|---|---|---|
| 74 | 79 | 77 | 81 |
| 68 | 79 | 81 | 76 |
| 81 | 80 | 80 | 78 |
| 88 | 83 | 79 | 91 |
| 79 | 75 | 74 | 73 |

| Group 2 | | | |
|---|---|---|---|
| 65 | 77 | 78 | 74 |
| 76 | 73 | 71 | 71 |
| 86 | 81 | 76 | 89 |
| 79 | 78 | 77 | 76 |
| 72 | 76 | 75 | 79 |

**solution**: First we need the quartile values:

| | $Q(.25)$ | $Q(.5)$ | $Q(.75)$ |
|---|---|---|---|
| Group 1 | 75.5 | 79 | 81 |
| Group 2 | 73.5 | 76 | 78.5 |

This means that Group 1 has IQR = 5.5 and

- 1.5*IQR = 8.25

while Group 2 has IQR = 5 and

- 1.5*IQR = 7.5

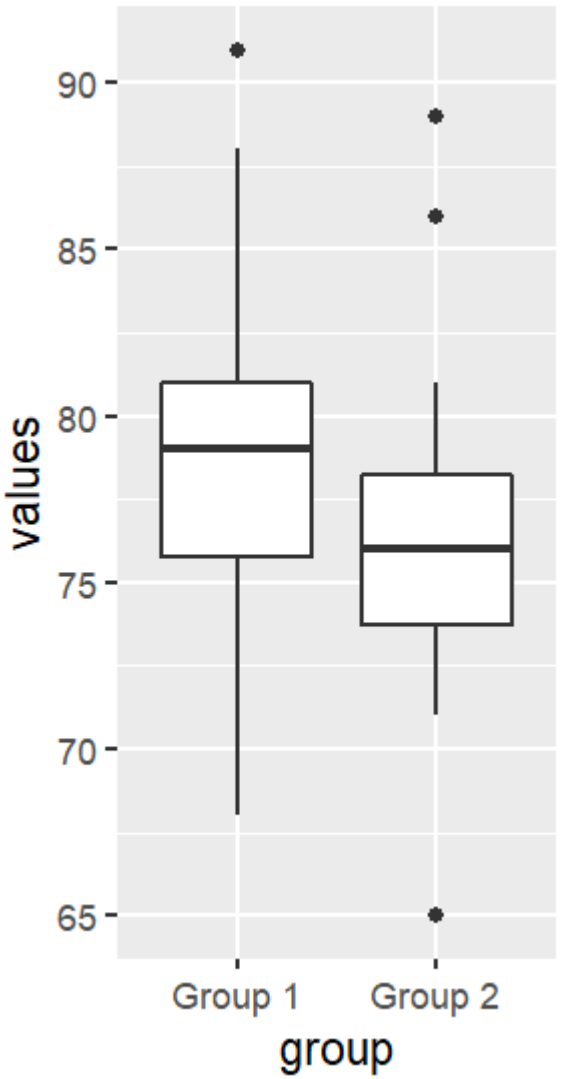
# Plots and Quantiles

Quantile Plots

QQ Plots

Boxplots

**Example**:



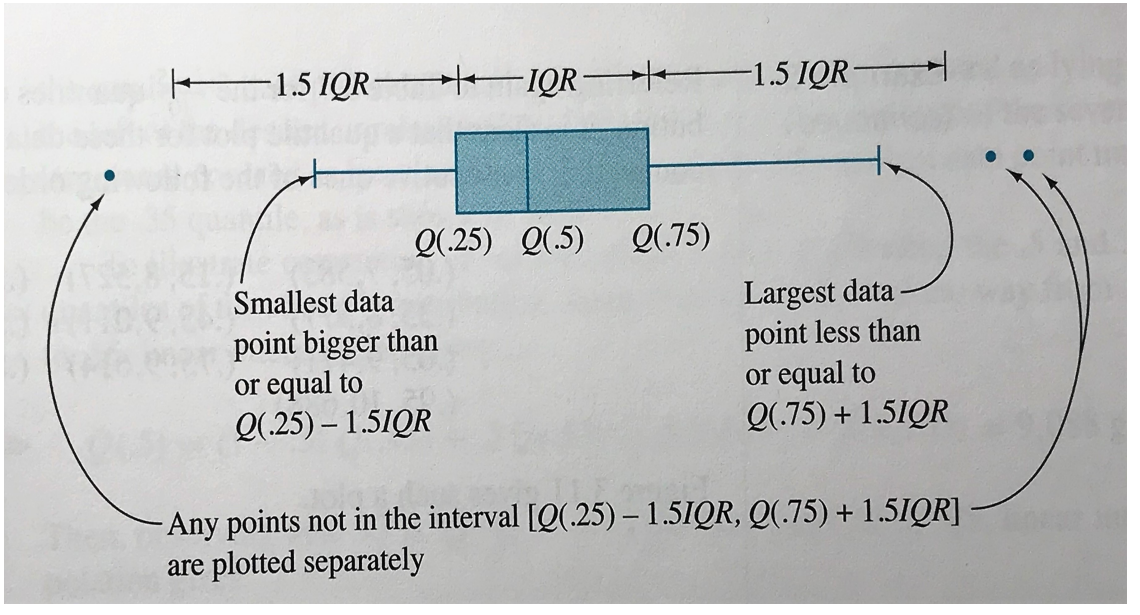
# Plots and Quantiles

Quantile Plots

QQ Plots

Boxplots

## Anatomy of a Boxplot



# Plots and Quantiles

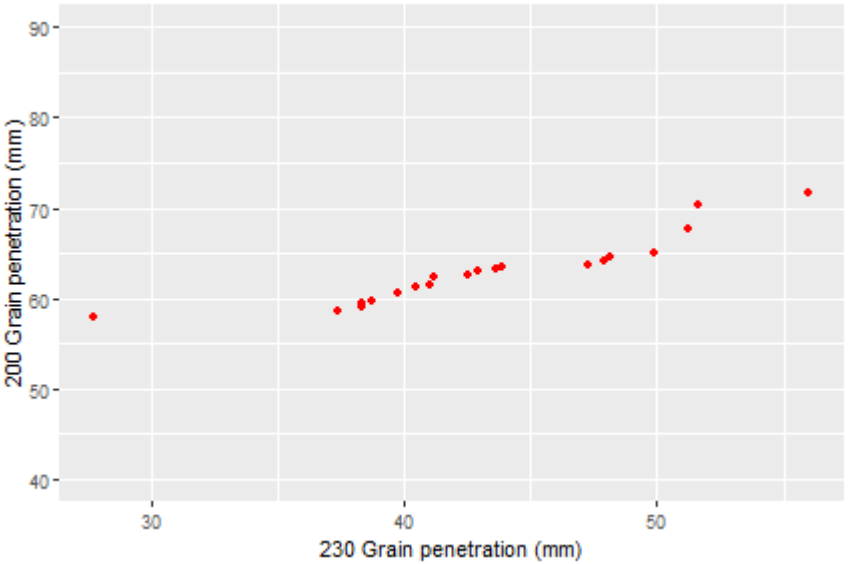
Quantile Plots

QQ Plots

Boxplots

## Recap: Example 6 of chapter 3: Bullet penetration depth

- 230 Grain penetration (mm)
- 200 Grain penetration (mm)




Except extreme lower values, it **seems** the two distributions have smiliar shapes; however, it still needs more attention to make a rough decision (consider boxplots).

# Plots and Quantiles

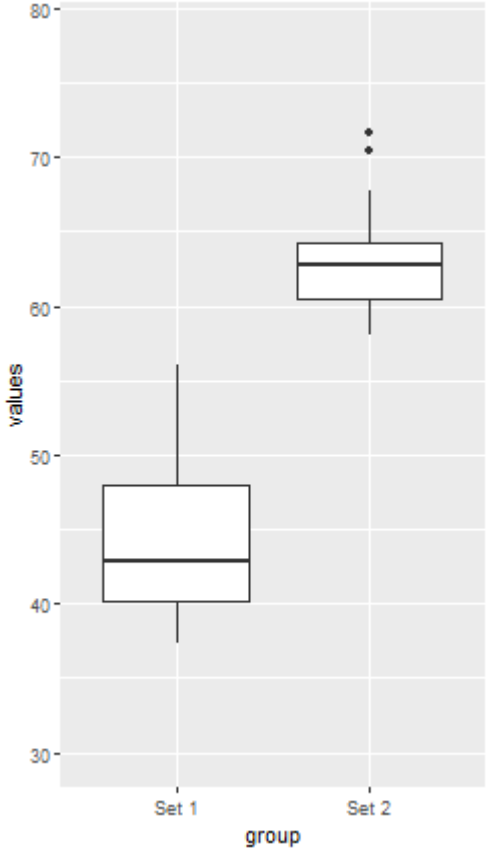
Quantile Plots

QQ Plots

Boxplots

## Recap: Example 6 of chapter 3: Bullet penetration depth

**Boxplot**



# Summarizing data Numerically

## Location and central tendency

## Measures of Spread

# Recap: Location and central tendency

Motivated by asking what is *normal/common/expected* for this data

**Mean**: A "fair" center value - $\frac{1}{n}\sum_{i=1}^{n} x_i$

**Mode**: The most commonly occurring value in set

**Median**: The value dividing the set in half (the middle of the values).

| 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 |

For group 1, the mean is 78.8, the median is 79, and the mode is 79.

For group 2, the mean is 76.45, the median is 76, and the mode is 76.

# Recap

## Location

## Spread

## Summaries of Variablity (Measures of Spread)

Motivated by asking what kind of *variability is seen in the data* or *how spread out* the data is.

**Range**: The difference between the highest and lowest values (Range = max - min)

**IQR**: The Interquartile Range, how spread out is the middle 50% (IQR = Q3 - Q1)

**Variance/Standard Deviation**: Uses squared distance from the mean.

| | Variance | Standard Deviation |
|---|---|---|
| **Population** | $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ | $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ |
| **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 Data Numerically

**Example**: Taking a sample of size 5 from a population we record the following values:

$$57, 60, 70, 59, 68$$

Find the variance and standard deviation of this sample.

# Example: Finding the Variance

Since we are told it is a sample, we need to use **sample variance**. The mean of 57, 60, 70, 59, 68 is 62.8

\[

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{5}(x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2\right) \\
&= \frac{1}{5-1}\left((57 - 62.8)^2 + (60 - 62.8)^2 + (70 - 62.8)^2 + (59 - 62.8)^2 + (68 - 62.8)^2\right) \\
&= \frac{1}{4}\left((-5.8)^2 + (-2.8)^2 + (7.2)^2 + (-3.8)^2 + (5.2)^2\right) \\
&= \frac{1}{4}(33.64 + 7.84 + 51.84 + 14.44 + 27.04) \\
&= 33.7
\end{aligned}
$$

`

]`

# Example: Finding the Standard Deviation

With $s^2$ known, finding $s$ is simple:

\[

$$
\begin{aligned}
s &= \sqrt{s^2} \\
&= \sqrt{33.7} \\
&= 5.8051701
\end{aligned}
$$

\]