

# Boxplots

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

## Boxplots

Quantiles are useful in making *boxplots*, an alternative to dot diagrams or histograms.

The boxplot shows less information, but many can be placed side by side on a single page for comparisons.

smallest data  
Point bigger  
than is equal  
to  
 $Q(0.25) - 1.5 \times IQR$

$1.5 \times IQR$   
IQR

$1.5 \times IQR$

largest data  
Point less  
than or equal  
to  
 $(Q(0.75) + 1.5 \times IQR)$

$Q(0.25)$

$Q(0.5)$

$Q(0.75)$

any points NOT in the interval

$[Q(0.25) - 1.5 \times IQR, Q(0.75) + 1.5 \times IQR]$

Summarizing

Intro

Purpose

Descriptive statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

are plotted separately.

## 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.

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

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

	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

**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

$$\bullet \text{ } 1.5 * \text{IQR} = 8.25$$

while Group 2 has IQR = 5 and

$$\bullet \text{ } 1.5 * \text{IQR} = 7.5$$

# Summarizing

## Example:

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

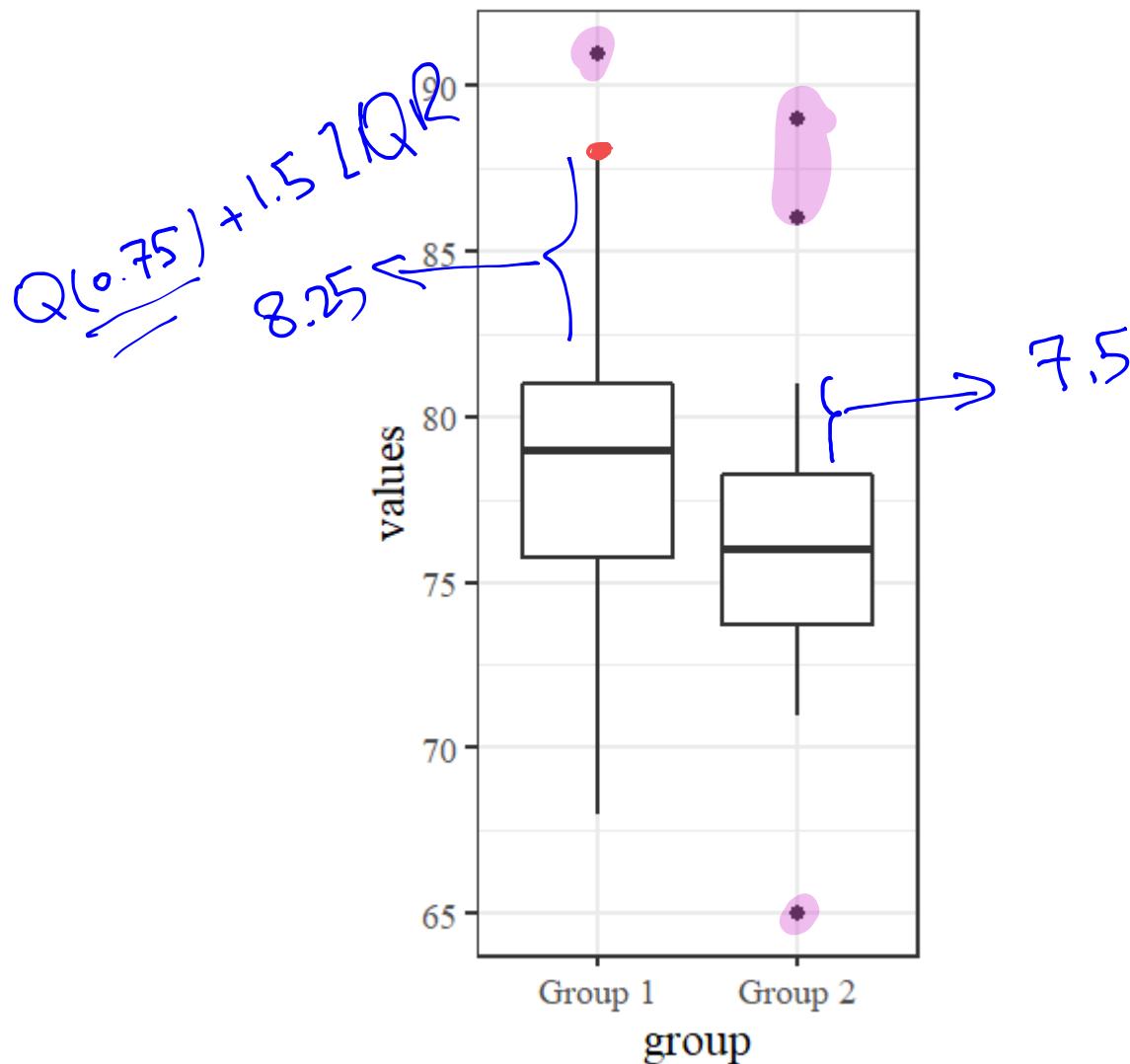
Histogram

Scatter plots

Center Stats

Quantiles

Boxplot



Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

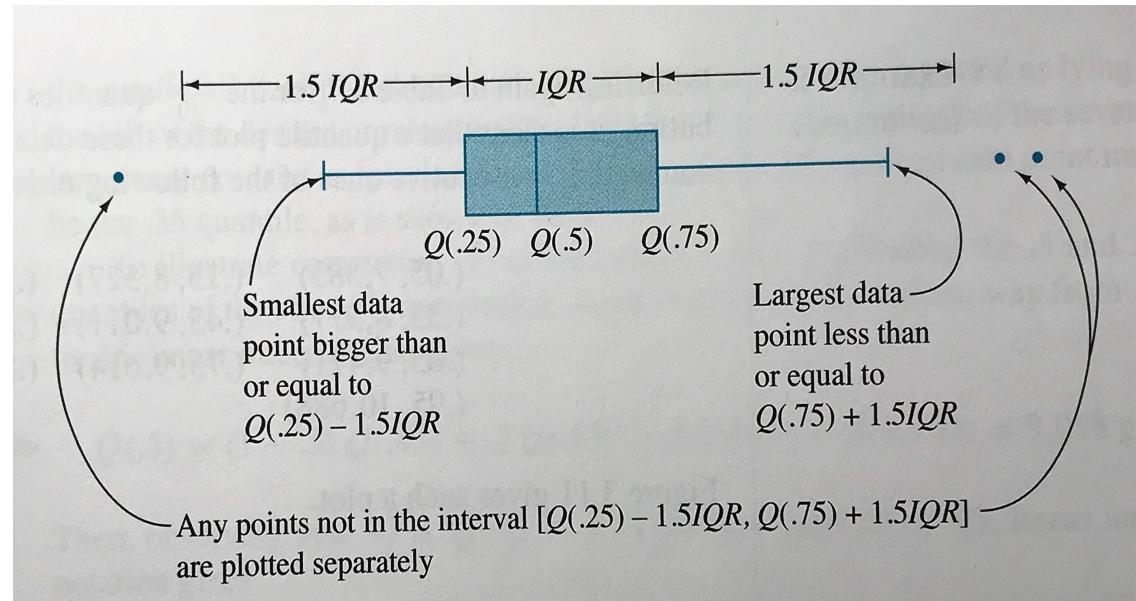
Scatter plots

Center Stats

Quantiles

Boxplot

## Anatomy of a Boxplot



# Summarizing

## Intro

## Purpose

## Descriptive statistics

## Plots

## Freq Tables

## Histogram

## Scatter plots

## Center Stats

## Quantiles

## Boxplot

## Example:[Bullet penetration depths, cont'd]

i	$\frac{i - .5}{20}$	200 grain bullets	230 grain bullets
1	0.025	58.00	27.75
2	0.075	58.65	37.35
3	0.125	59.10	38.35
4	0.175	59.50	38.35
5	0.225	59.80	38.75
6	0.275	60.70	39.75
7	0.325	61.30	40.50
8	0.375	61.50	41.00
9	0.425	62.30	41.15
10	0.475	62.65	42.55
11	0.525	62.95	42.90
12	0.575	63.30	43.60
13	0.625	63.55	43.85
14	0.675	63.80	47.30
15	0.725	64.05	47.90
16	0.775	64.65	48.15
17	0.825	65.00	49.85
18	0.875	67.75	51.25
19	0.925	70.40	51.60
20	0.975	71.70	56.00

63

63

70

for 230%

$$Q(0.25) = x_5 + (np - 5 + 0.5)(x_6 - x_5) = 39.25 \text{ mm}$$

$$Q(0.5) = x_{10} + (np - 10 + 0.5)(x_{11} - x_{10}) = 42.725 \text{ mm}$$

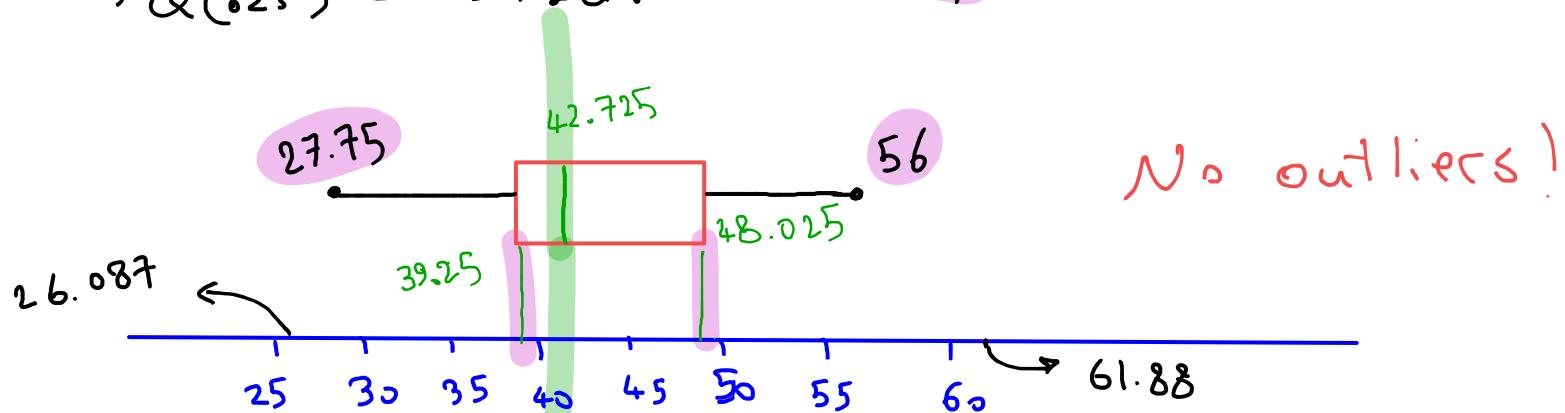
$$Q(0.75) = x_{15} + (np - 15 + 0.5)(x_{16} - x_{15}) = 48.025 \text{ mm}$$

$$IQR = Q(0.75) - Q(0.25) = 48.025 - 39.25 = 8.775$$

$$1.5 * IQR = 13.163$$

$$\left\{ \begin{array}{l} Q(0.75) + 1.5 * IQR = 61.88 \text{ mm} \end{array} \right.$$

$$\left. \begin{array}{l} Q(0.25) - 1.5 * IQR = 26.087 \text{ mm} \end{array} \right.$$



Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

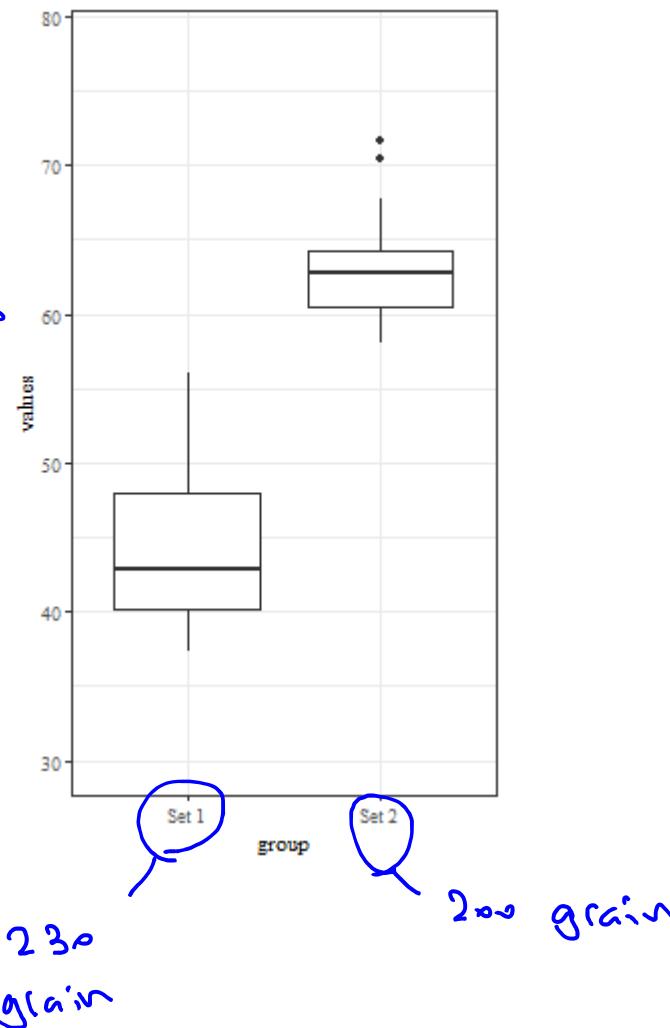
Center Stats

Quantiles

Boxplot

Example:[Bullet penetration depths, cont'd]

side-by-side  
Boxplots for  
the two types  
of bullets.



Quantile-quantile (Q-Q) plots

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

## Quantile-quantile (Q-Q) plots

Often times, we want to compare the **shapes** of two distributions.

→ compare histograms, boxplots, scatter plots

→ A more sensitive way is to make a single plot based on the quantile functions for two distributions.

A **Q-Q plot** for two data sets with respective quantile functions  $Q_1$  and  $Q_2$  is a plot of ordered pairs  $(Q_1(p), Q_2(p))$  for appropriate values of  $p$ . When two data sets of size  $n$  are involved, the values of  $p$  used to make the plot will be  $\frac{i-.5}{n}$  for  $i = 1, \dots, n$ .

"equal shape" = linearly related quantile functions.

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

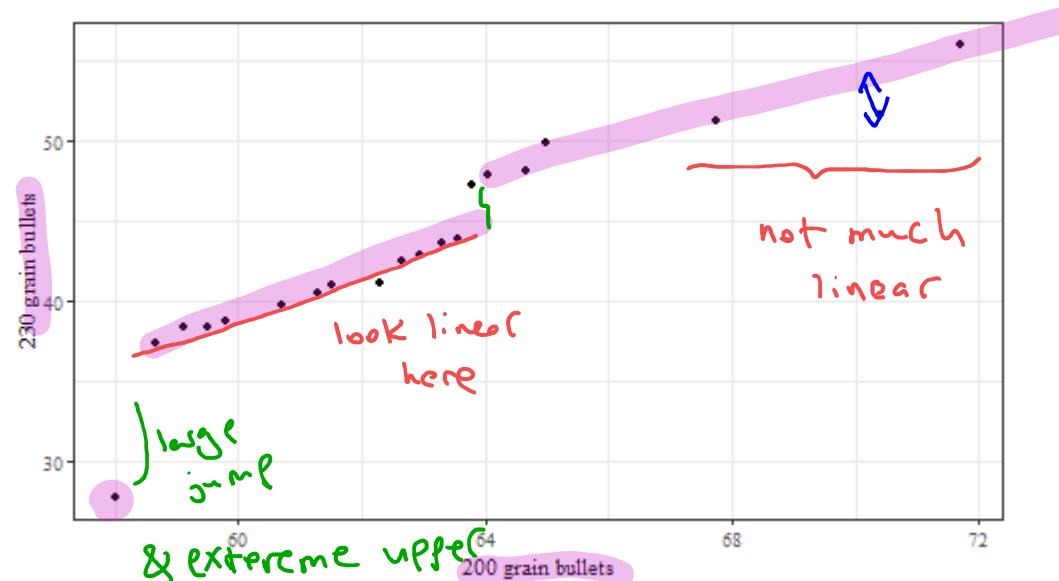
QQ-Plot

## Quantile-quantile (Q-Q) plots

**Example:** [Bullet penetration depth, cont'd]

**Example:** [Bullet penetration depth cont'd]

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



Except extreme lower values, it **seems** the two distributions have similar shapes; however, it still needs more attention to make a rough decision (consider boxplots).  $\Rightarrow$  overall : not super linear.

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

To make a Q-Q plot for two data sets of the same size,

1. order each from the smallest observation to the largest,
2. pair off corresponding values in the two data sets
3. plot ordered pairs, with the horizontal coordinate coming from the first data set and the vertical ones from the second.

**Example:** [Q-Q plot by hand]

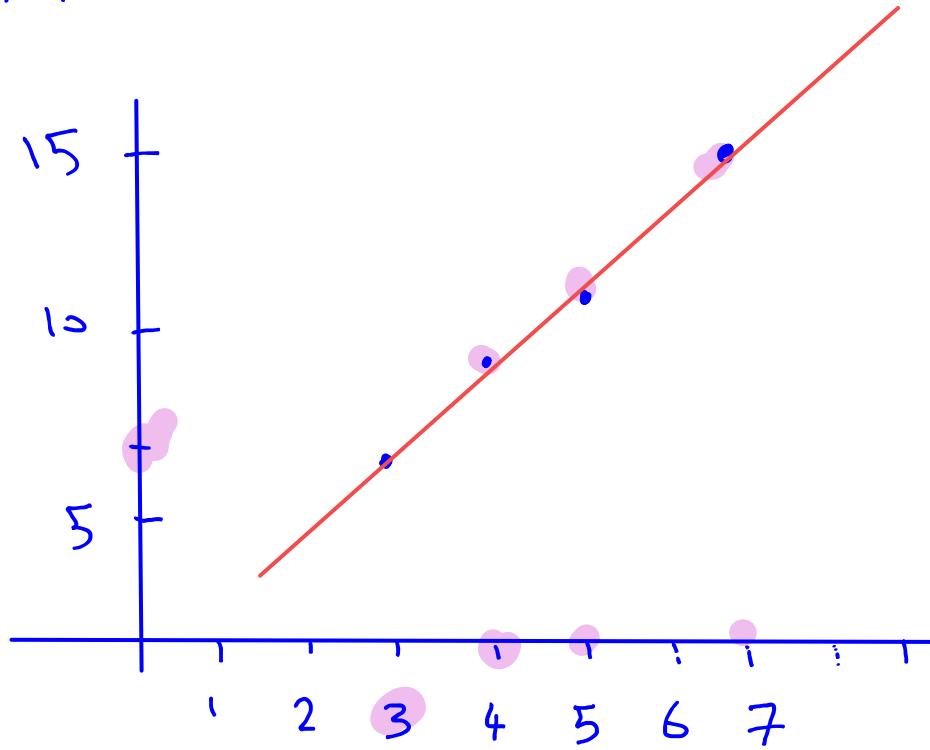
Make a Q-Q plot for the following small artificial data sets.

Data set 1	Data set 2
3, 5, 4, 7, 3	15, 7, 9, 7, 11

① order :  
Set 1 :  $\boxed{3}$   $\boxed{3}$   $\boxed{4}$   $\boxed{5}$   $\boxed{7}$   $\rightarrow \boxed{5}$   
Set 2 :  $\boxed{7}$   $\boxed{7}$   $\boxed{9}$   $\boxed{11}$   $\boxed{15}$   $\boxed{16}$   $17$   $20$

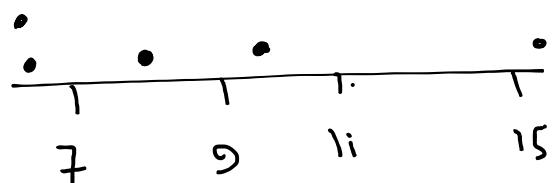
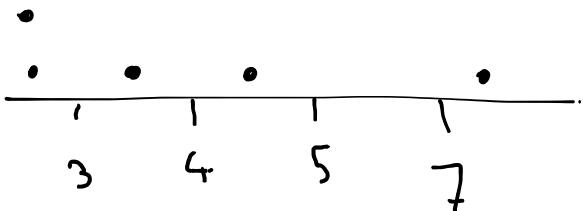
② pair off  
 $(3, 7)$      $(5, 11)$   
 $(3, 7)$      $(7, 15)$   
 $(4, 7)$

### ③ scatter plot



perfectly linear!  $\Rightarrow$  The two data sets have  
exactly the same shape.

using dot diagrams



scale/location differ  
but the same  
shape!

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

## 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.

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

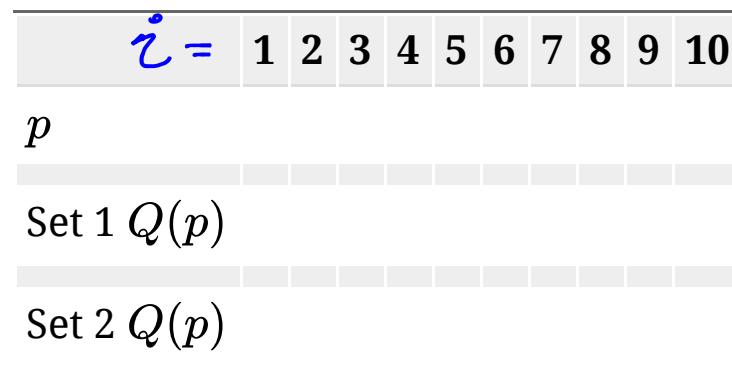
Boxplot

QQ-Plot

## 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



Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

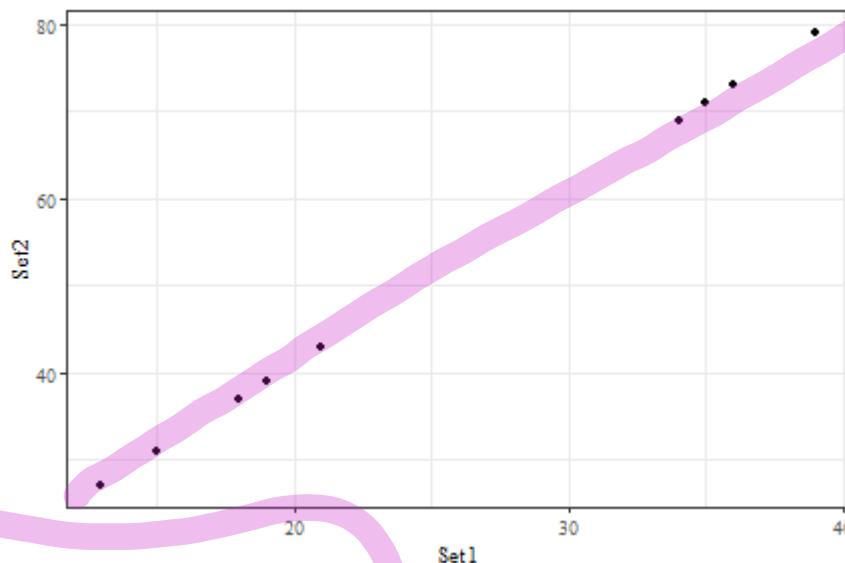
Quantiles

Boxplot

QQ-Plot

## 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
$Q_1(p)$	13	15	18	19	21	34	35	35	36	39
$Q_2(p)$	27	31	37	39	43	69	71	71	73	79



$$F(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

## 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.

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

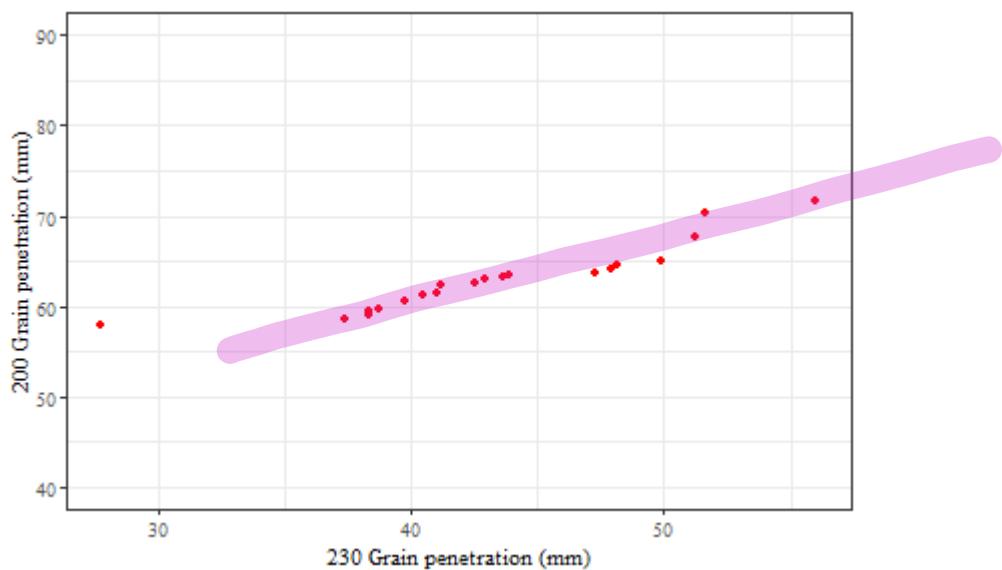
Boxplot

QQ-Plot

## 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 similar shapes; however, it still needs more attention to make a rough decision (consider boxplots). Might want to figure out what has caused the extreme value

## Theoretical quantile-quantile plots

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

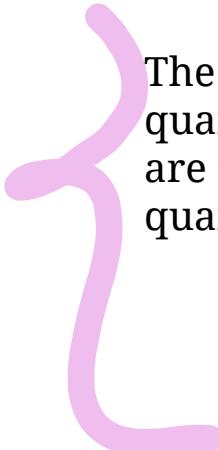
Boxplot

QQ-Plot

## Theoretical quantile-quantile plots

Q-Q plots are useful when comparing two finite data sets, but a Q-Q plot can also be used to compare a data set and an expected shape, or *theoretical distribution*.

A **theoretical Q-Q plot** for a data set of size  $n$  and a theoretical distribution, with respective quantile functions  $Q_1$  and  $Q_2$  is a plot of ordered pairs  $(Q_1(p), Q_2(p))$  for  $p = \frac{i-.5}{n}$  where  $i = 1, \dots, n$ .



The most famous theoretical Q-Q plot occurs when quantiles for the *standard Normal* or *Gaussian* distribution are used. A simple numerical approximation to the quantile function for the Normal distribution is

$$Q(p) \approx 4.9(p^{.14} - (1-p)^{.14}).$$

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

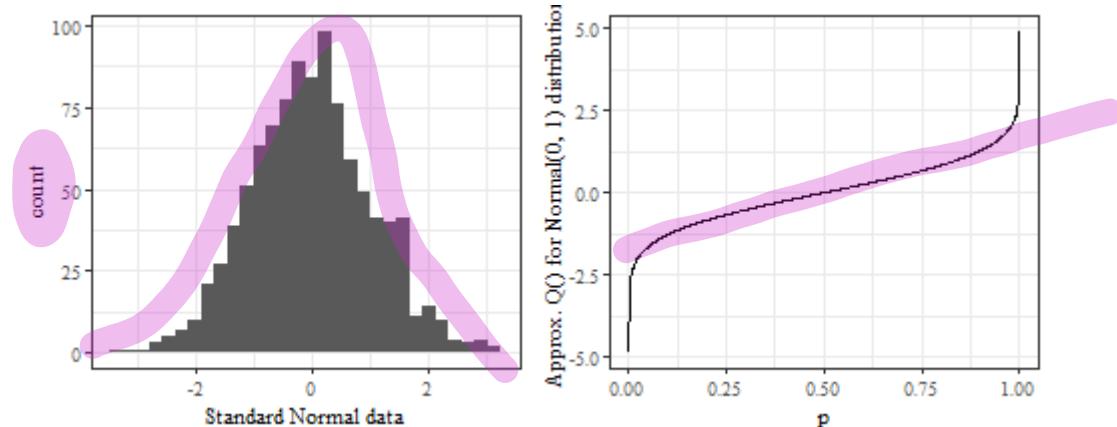
Boxplot

QQ-Plot

## Theoretical quantile-quantile plots

The standard Normal quantiles can be used to make a theoretical Q-Q plot as a way of assessing how bell-shaped a data set is.

The resulting plot is called a **normal Q-Q plot**.



Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

## Quantile-Quantile Plots:

### Summary

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?"

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

**Example:**[Breaking strengths of paper towels, pg. 79]

Here is a study of the dry breaking strength (in grams) of generic paper towels.

test	strength
1	8577
2	9471
3	9011
4	7583
5	8572
6	10688
7	9614
8	9614
9	8527
10	9165



## Summarizing

### Intro

### Purpose

### Descriptive statistics

### Plots

### Freq Tables

### Histogram

### Scatter plots

### Center Stats

### Quantiles

### Boxplot

### QQ-Plot

Example:[Breaking strengths of paper towels, pg. 79]

i	$\frac{i - .5}{20}$	Breaking strength $Q(i)$	Q0	Standard Normal $Q\theta$
1	0.05	7583		-1.6448536
2	0.15	8527		-1.0364334
3	0.25	8572		-0.6744898
4	0.35	8577		-0.3853205
5	0.45	9011		-0.1256613
6	0.55	9165		0.1256613
7	0.65	9471		0.3853205
8	0.75	9614		0.6744898
9	0.85	9614		1.0364334
10	0.95	10688		1.6448536

Summarizing

Intro

Purpose

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

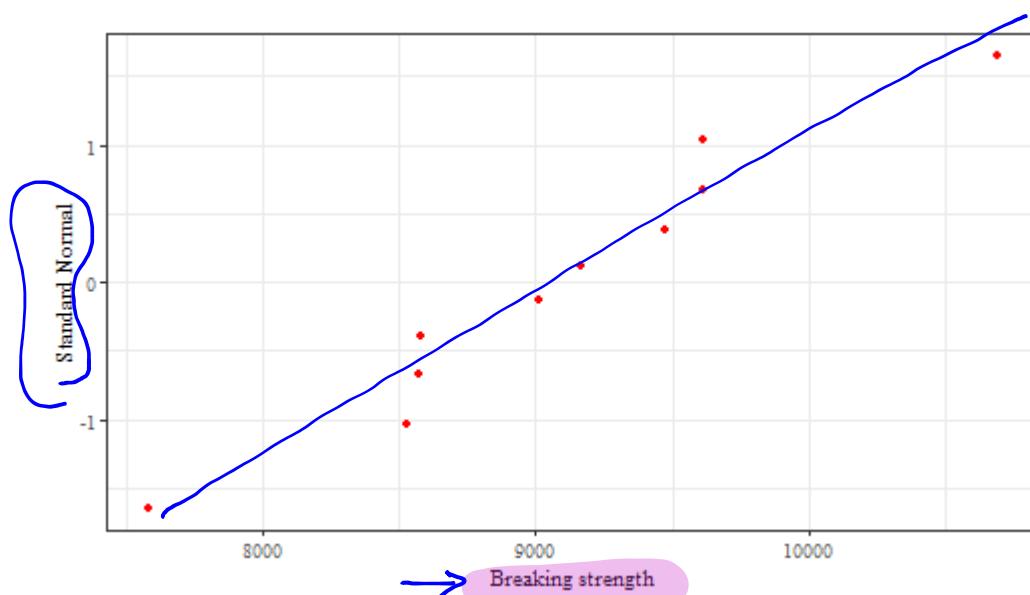
Quantiles

Boxplot

QQ-Plot

## Theoretical quantile-quantile plots

**Example:** [Breaking strengths of paper towels, pg. 79]



roughly straight => breaking strength  
data is roughly  
bell-shaped.

# Summarizing data Numerically

Location and central tendency

Measures of Spread

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Numerical summaries

When we have a large amount of data, it can become important to reduce the amount of data to a few informative numerical summary values. Numerical summaries highlight important features of the data

A **numerical summary (or statistic)** is a number or list of numbers calculated using the data (and only the data).

### Measures of location

An "average" represents the center of a quantitative data set. There are several potential technical meanings for the word "average", and they are all *measures of location*.

The **(arithmetic) mean** of a sample of quantitative data  $(x_1, \dots, x_n)$  is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Numerical summaries

### Measures of location

The mode of a discrete or categorical data set is the most frequently-occurring value.

We have also seen the median,  $Q(.5)$ , which is another measure of location. A shortcut to calculating  $Q(0.5)$  is

- $Q(0.5) = x_{\lceil n/2 \rceil}$  if  $n$  is odd
- $Q(0.5) = (x_{n/2} + x_{n/2+1})/2$  if  $n$  is even.

### Example:[Measures of location]

Calculate the three measures of location for the following data.

$$0, 1, \boxed{1}, 2, 3, 5 \quad \text{med.} = \frac{2+1}{2} = 1.5$$

$$\bar{x} = \frac{1}{6} (0 + 1 + 1 + 2 + 3 + 5) = \underline{\underline{2}}$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Numerical summaries

### Measures of spread

Quantifying variation in a data set can be as important as measuring its location. Again, there are many way to measure the spread of a data set.

- The **range** of a data set consisting of ordered values  $x_1 \leq \dots \leq x_n$  is

$$R = x_n - x_1.$$

- The **sample variance** of a data set consisting of values  $x_1, \dots, x_n$  is

$$\text{Pop. Variance} \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

$$\text{Sample Variance} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The **sample standard deviation**,  $s$ , is the nonnegative square root of the sample variance.  $\Rightarrow S = \sqrt{s^2}$

We have also seen the  $IQR$ ,  $Q(.75) - Q(.25)$ , which is another measure of spread.

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Numerical summaries

### Measures of spread

#### Example:[Measures of spread]

Calculate the four measures of spread for the following data.

0, 1, 1, 2, 3, 5

$$\bar{x} = 2$$

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

$$\begin{aligned} n &= 6 \\ s^2 &= \frac{1}{6-1} \left[ (0-2)^2 + (1-2)^2 + (1-2)^2 \right. \\ &\quad \left. + (3-2)^2 + (5-2)^2 \right] \end{aligned}$$

$$= \frac{1}{5} (7+9) = 16/5$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

$$s = \sqrt{\frac{16}{5}}$$

## Numerical summaries

### Measures of spread

**Example:** [Sensitivity to outliers] Which measures of center and spread differ drastically between the  $x_i$ s and the  $y_i$ s? Which ones are the same?

$$x_i : 0, 1, 1, 2, 3, 5$$

$$y_i : 0, 1, 1, 2, 3, 81, 72, 63, 48, 9$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Statistics and parameters

It's important now to stop and talk about terminology and notation.

- Numerical summarizations of sample data are called (sample) **statistics**. Numerical summarizations of population and theoretical distributions are called (population or model) **parameters**.
- If a data set,  $x_1, \dots, x_N$ , represents an entire population, then the **population (or true) mean** is defined as

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i.$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Statistics and parameters

- If a data set,  $x_1, \dots, x_N$ , represents an entire population, then the **population (or true) variance** is defined as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2.$$

The **population (or true) standard deviation**,  $\sigma$  is the nonnegative square root of  $\sigma^2$ .

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Categorical and count data

So far we have talked mainly about summarizing quantitative, or measurement, data. Sometimes, we have categorical or count data to summarize. In this case, we can revisit the *frequency table* and introduce a new type of plot.

### Example:[Cars]

Fuel consumption and 10 aspects of automobile design and performance are available for 32 automobiles (1973–74 models) from 1974 Motor Trend US Magazine.

Summarizing

Descriptive statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical summaries

### Example:[Cars]

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
...	...	...	...	...	...	...	...	...	...	...	...

We can construct a frequency table for the *cylinder* variable.

cyl	Frequency	Relative Frequency	Cumulative Frequency
4	11	0.34375	0.34375
6	7	0.21875	0.56250
8	14	0.43750	1.00000

From this frequency data, we can summarize the categorical data graphically.

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

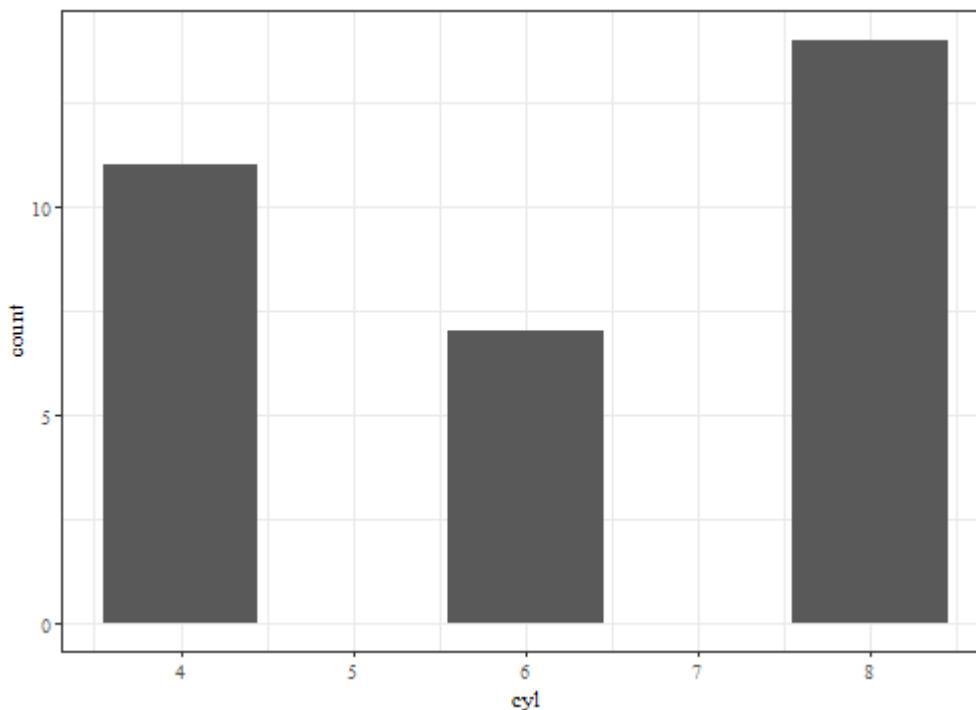
Boxplot

QQ-Plot

Numerical  
summaries

- A **bar plot** presents categorical data with rectangular bars with lengths proportional to the values that they represent (usually frequency of occurrence).

### Example:[Cars, cont'd]



Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

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

68, 54, 60, 66, 58

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 68, 54, 60, 66, 58 is 61.2

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^5 (x_i - \bar{x})^2 \\&= \frac{1}{n-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2) \\&= \frac{1}{5-1} ((68 - 61.2)^2 + (54 - 61.2)^2 + (60 - 61.2)^2 + (66 - 61.2)^2 + (58 - 61.2)^2) \\&= \frac{1}{4} ((6.8)^2 + (-7.2)^2 + (-1.2)^2 + (4.8)^2 + (-3.2)^2) \\&= \frac{1}{4} (46.24 + 51.84 + 1.44 + 23.04 + 10.24) \\&= 33.2\end{aligned}$$

Summarizing

Descriptive  
statistics

Plots

Freq Tables

Histogram

Scatter plots

Center Stats

Quantiles

Boxplot

QQ-Plot

Numerical  
summaries

## Example: Finding the Standard Deviation

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

$$s = \sqrt{s^2}$$

$$= \sqrt{33.2}$$

$$= 5.7619441$$

