3 Descriptive statistics

Engineering data are always variable. Given precise enough measurement, even constant process conditions produce different responses. Thus, it is not the individual data values that are important, but their **distribution**. We will discuss simple methods that describe important distributional characteristics of data.

Definition 3.1. *Descriptive statistics* is the use of plots and numerical summaries to describe data without drawing any formal conclusions.

Through the use of *descriptive statistics*, we seek to find the following features of data sets:

1. Center

   the point that the data are closest to on average.

2. Spread

   how wide the data look, how varied the points are

3. Shape

   common patterns / trends that are present in the data

4. Outliers

   points that lie way beyond the rest of the data (weird points)

3.1 Graphical and tabular displays of quantitative data

numerical values

Almost always, the place to start a data analysis is with appropriate graphical and tabular displays. When only a few samples are involved, a good plot can tell most of the story about data and drive an analysis.

in an analysis, start by plotting!

3.1.1 Dot diagrams and stem-and-leaf plots

⇒ at most hundreds (?)

When a study produces a small or moderate amount of **univariate quantitative data**, a *dot diagram* can be useful.

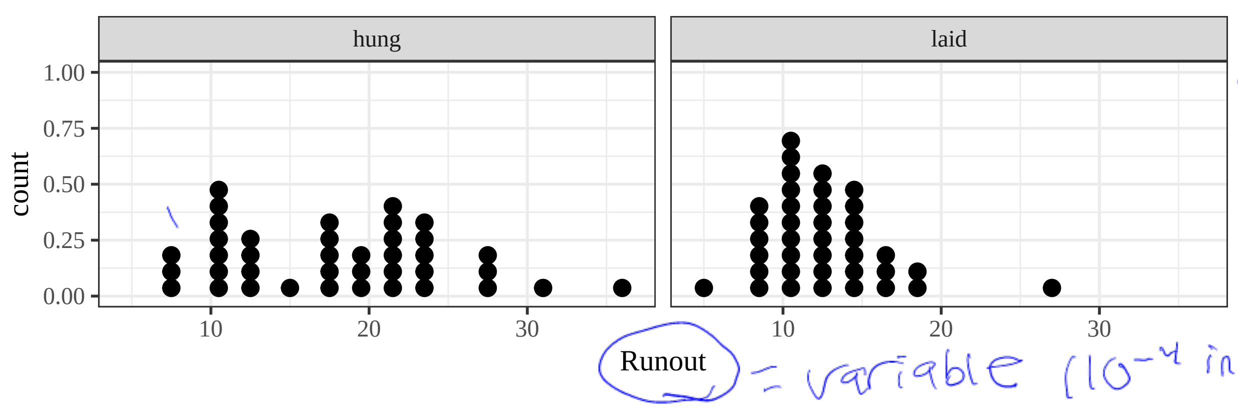
only a single characteristic is observed (or important)

numeric (ch. 1).

Definition 3.2. A *dot diagram* shows each observation as a dot placed at the position corresponding to its numerical value along a number line.

Example 3.1 (Heat treating gears, cont'd)**.** Recall the example from Chapter 1. A process engineer is faced with the question, "How should gears be loaded into a continuous carburizing furnace in order to minimize distortion during heat treating?" The engineer conducts a well-thought-out study and obtains the runout values for 38 gears laid and 39 gears hung.

39 hung = 39 dots.



38 laid = 38 dots

= variable (10^{-4} inches)

Figure 1: Dot diagrams of runouts.

each dot represents a gear in the study.

laid values are generally smaller and more consistent than hung gears.

Example 3.2 (Bullet penetration depth, pg. 67)**.** Sale and Thom compared penetration depths for several types of .45 caliber bullets fired into oak wood from a distance of 15 feet. They recorded the penetration depths (in mm from the target surface to the back of the bullets) for two bullet types.

1 quant. var   1 quant var

200 grain jacketed bullets	230 grain jacketed bullets
63.8, 64.65, 59.5, 60.7, 61.3, 61.5, 59.8, 59.1, 62.95, 63.55, 58.65, 71.7, 63.3, 62.65, 67.75, 62.3, 70.4, 64.05, 65, 58	40.5, 38.35, 56, 42.55, 38.35, 27.75, 49.85, 43.6, 38.75, 51.25, 47.9, 48.15, 42.9, 43.85, 37.35, 47.3, 41.15, 51.6, 39.75, 41

Table 1: Bullet penetration depths (mm)

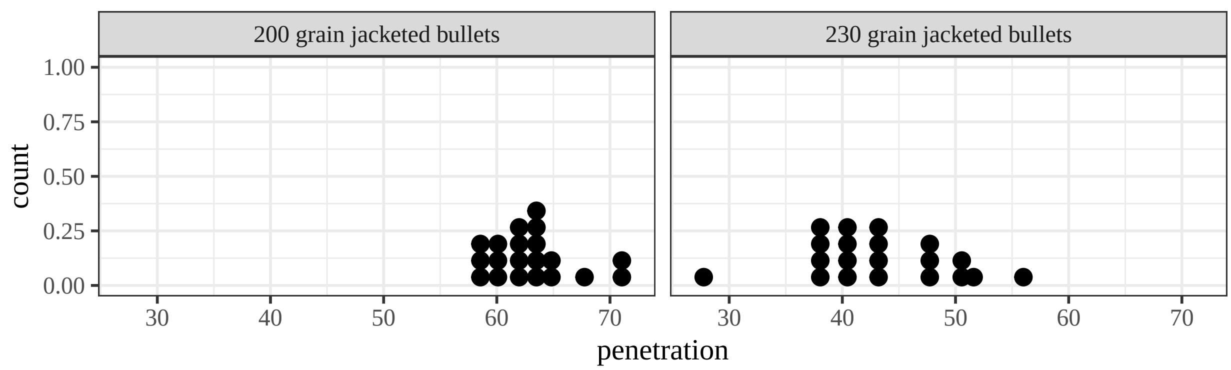


Figure 2: Dot diagrams of penetration depths.

- 200 grain bullets center around a larger value
- 230 grain bullets have a larger spread

Dot diagrams are good for getting a general feel for the data (and can be done with pencil and paper), but do not allow the recovery of the exact values used to make them.

Definition 3.3. A *stem-and-leaf plot* is made by using the last few digits of each data point to indicated where it falls.

↳ does allow the recovery of exact values

Example 3.3 (Heat treating gears, cont'd)**.**

"leaves" second digit of each obsvn

05, 08, 08, 09, 09, 09, 09

"stem" first digit of each obsvn

hung	laid
7, 8, 8, 10, 10, 10, 10, 11, 11,	5, 8, 8, 9, 9, 9, 9, 10, 10, 10,
11, 12, 13, 13, 13, 15, 17, 17,	11, 11, 11, 11, 11, 11, 11, 12,
17, 17, 18, 19, 19, 20, 21, 21,	12, 12, 12, 13, 13, 13, 13, 14,
21, 22, 22, 22, 23, 23, 23, 23,	14, 14, 15, 15, 15, 15, 16, 17,
24, 27, 27, 28, 31, 36	17, 18, 19, 27

Table 2: Thrust face runouts (.0001 in.)

0	5 8 8 9 9 9 9
1	0 0 0 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 6 7 7 8 9
2	7

Laid Gears

→ leaf digits 0–4 (5 values)
→ leaf digits 5–9 (5 values)

8 8 7	0	5 8 8 9 9 9 9
3 3 3 2 1 1 1 0 0 0 0	1	0 0 0 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4
9 9 8 7 7 7 7 5	1	5 5 5 5 6 7 7 8 9
4 3 3 3 3 2 2 2 1 1 1 0	2	
8 7 7	2	7
1	3	
6	3	

Hung Gears | Laid Gears

"side to side"

equal STEM intervals b/c they allow an equal # of leaf digits (0–9, all have 10 possible values)

3.1.2 Frequency tables and histograms

Dot diagrams and stem-and-leaf plots are useful for getting to know a data set, but they are not commonly used in papers and presentations.

Definition 3.4. A *frequency table* is made by first breaking an interval containing all the data into an appropriate 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, frequencies, relative frequencies, and cumulative relative frequencies can be added.

Example 3.4 (Heat treating gears, cont'd).

3/38

choice of interval is arbitrary:
① intervals are equal sizes
② rounded numbers are preferred

Runout (.0001 in)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency	
5-8	|||	3	.079	.079	
9-12	~~||||~~ ~~||||~~ ~~||||~~ |||	18	.474	.553	← 0.79+0.474
13-16	~~||||~~ ~~||||~~ ||	12	.316	.868	← 0.553+0.316
17-20	||||	4	.105	.974	← 0.868+0.105
21-24		0	0	.974	
25-28	|	1	.026	1.000	← 0.974+0.026
		38	1.000		

↳ # of data points

Table 3: Frequency table for laid gear thrust face runouts.

Example 3.5 (Bullet penetration depth, cont'd).

5/20

Runout (.0001 in)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency	
58-59.99	~~||||~~	5	.25	.25	
60.00-61.99	|||	3	.15	.40	← 0.25+0.15
62.00-63.99	~~||||~~ |	6	.30	.70	
64.00-65.99	|||	3	.15	.85	
66.00-67.99	|	1	.05	.90	
68.00-69.99		0	0	.90	
70.00-71.99	||	2	.10	1.000	
		20	1.000		

↳ # of data points

Table 4: Frequency table for 200 grain penetration depths.

easy!

frequency table $\Longrightarrow$ histogram

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

Definition 3.5. A *(frequency or relative frequency) histogram* is a kind of bar chart used to portray the shape of a distribution of data points.

Guidelines for making histograms:

1. use intervals of equal length (2mm)
2. show the entire vertical axis starting at ~~zero~~
3. Avoid breaking either axis
4. keep a uniform scale for axes (tick marks)
5. center bars of appropriate heights at the midpoint of the intervals

Example 3.6 (Bullet penetration depth, cont'd)**.**

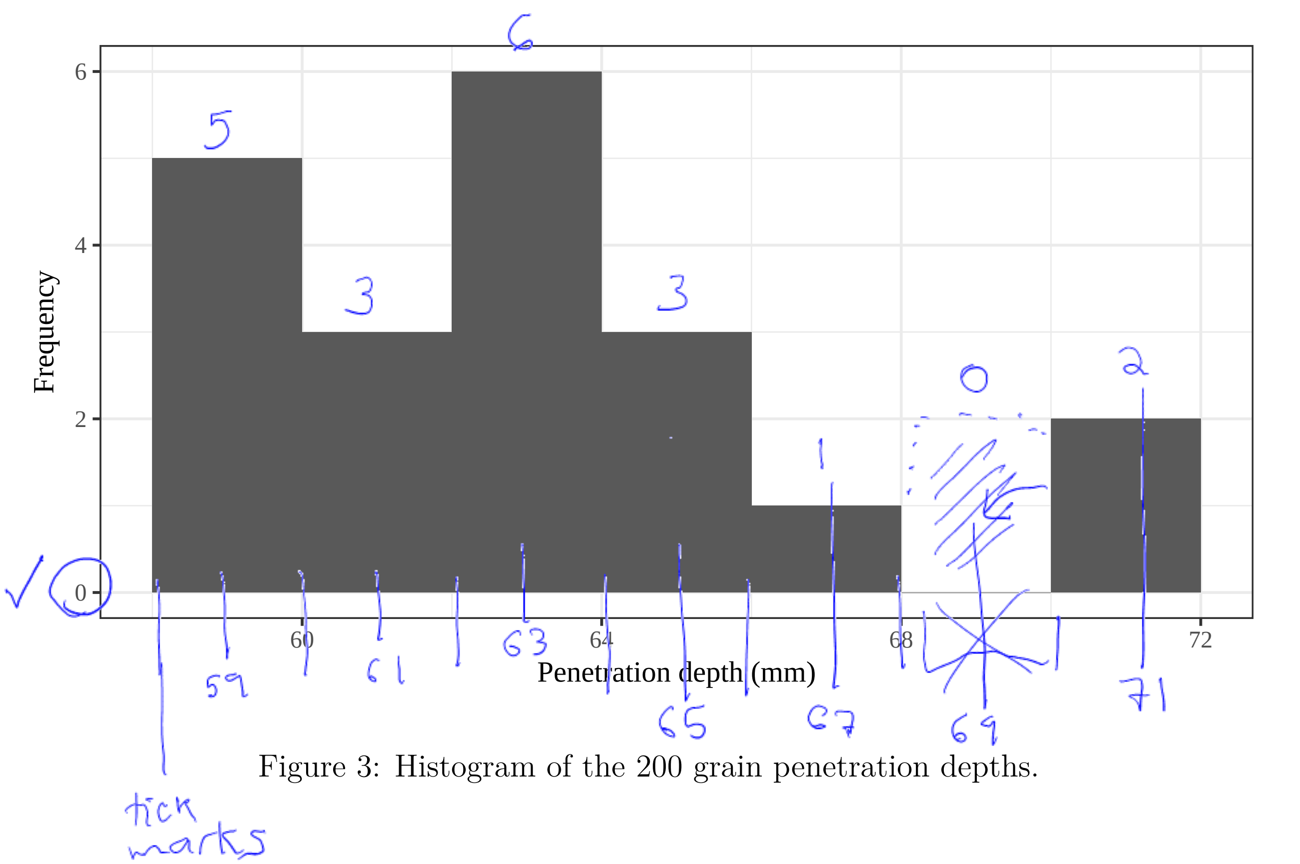


Figure 3: Histogram of the 200 grain penetration depths.

use intervals of 4!

Example 3.7 (Histogram)**.** Suppose you have the following data:

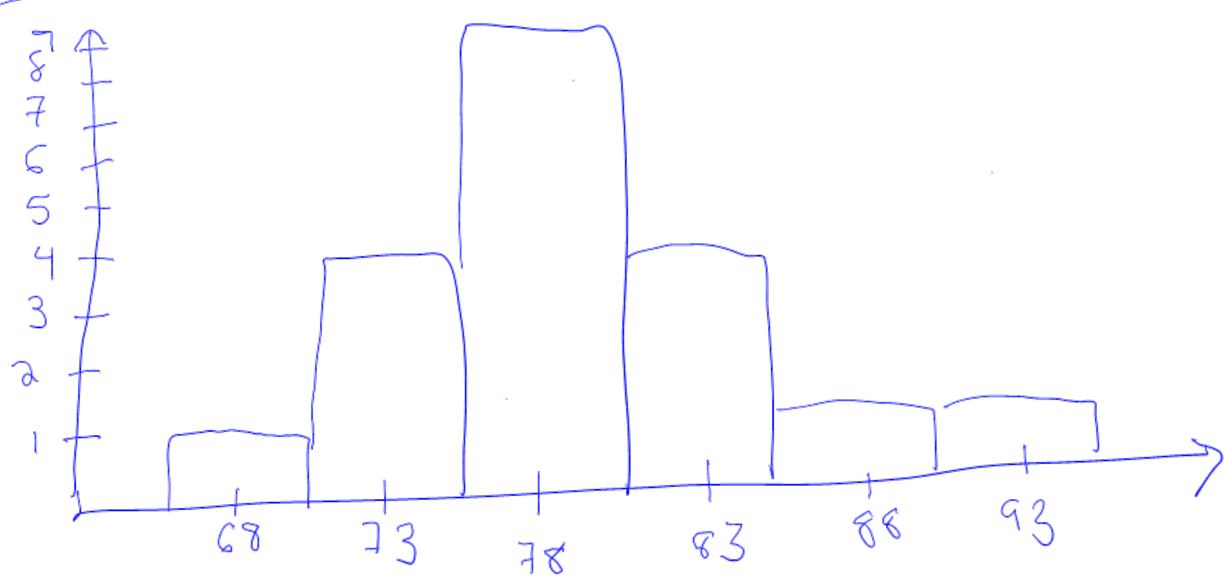
$$74, 79, 77, 81, 68, 79, 81, 76, 81, 80, 80, 78, 88, 83, 79, 91, 79, 75, 74, 73$$

. Create the corresponding *frequency table* and *frequency histogram*.

Frequency table

class	Tally	Freq	Rel. Freq.	c.r. Freq
66-70	|	1	1/20 = 0.05	0.05
71-75	||||	4	0.20	0.25
76-80	~~||||~~ ||||	9	0.45	0.70
81-85	||||	4	0.20	0.90
86-90	|	1	0.05	0.95
91-95	|	1	0.05	1
		20	1	

Histogram



"center"
1 2 3

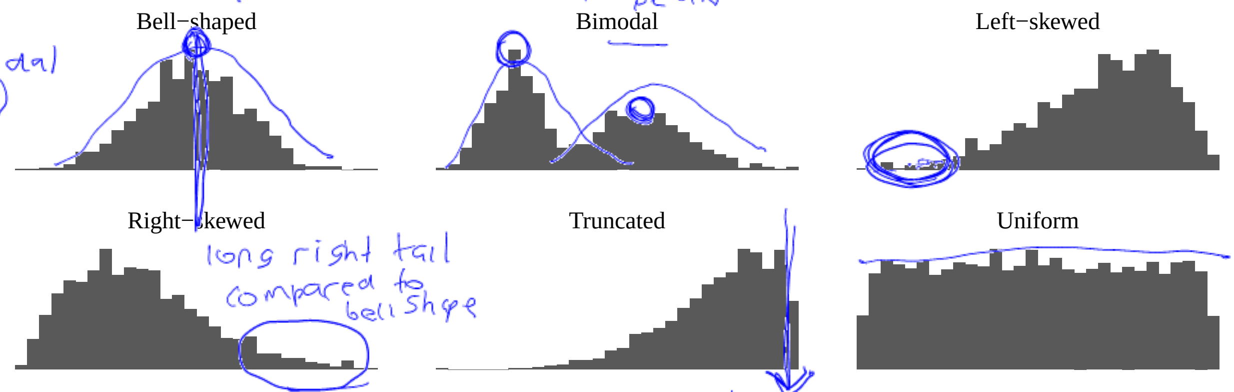
Why do we plot data? Information on location, spread, and shape is portrayed clearly in a histogram and can give hints as to the functioning of the physical process that is generating the data.

1 peak in center
symmetrical
"normal"

2 "modes"/ "peaks"

long left tail compared to bell shape

unimodal (1 peak)



long right tail compared to bell shape

flat

cut-off

Figure 4: Common distributional shapes.

If data on the diameters of machined metal cylinders purchased from a vendor produce a histogram that is decidedly **bimodal**, this suggests

The machining was done on 2 machines or by 2 operators or at 2 times, etc...

If the histogram is **truncated**, this might suggest

The cylinders have been 100% inspected, removing all cylinders w/ excess diameters

3.1.3 Scatter plots

Dot-diagrams, stem-and-leaf plots, frequency tables, and histograms are univariate tools. But engineering questions often concern multivariate data and *relationships between the variables.*
quantitative

Definition 3.6. A *scatterplot* is a simple and effective way of displaying potential relationships between two quantitative variable by assigning each variable to either the x or y axis and plotting the resulting coordinate points.

Example 3.8 (Orange trees)**.** Jim and Jane want to know the relationship between an orange tree's age (in days since 1968-12-31) and its circumference (in mm). They recorded the data for 35 orange trees.

35 points

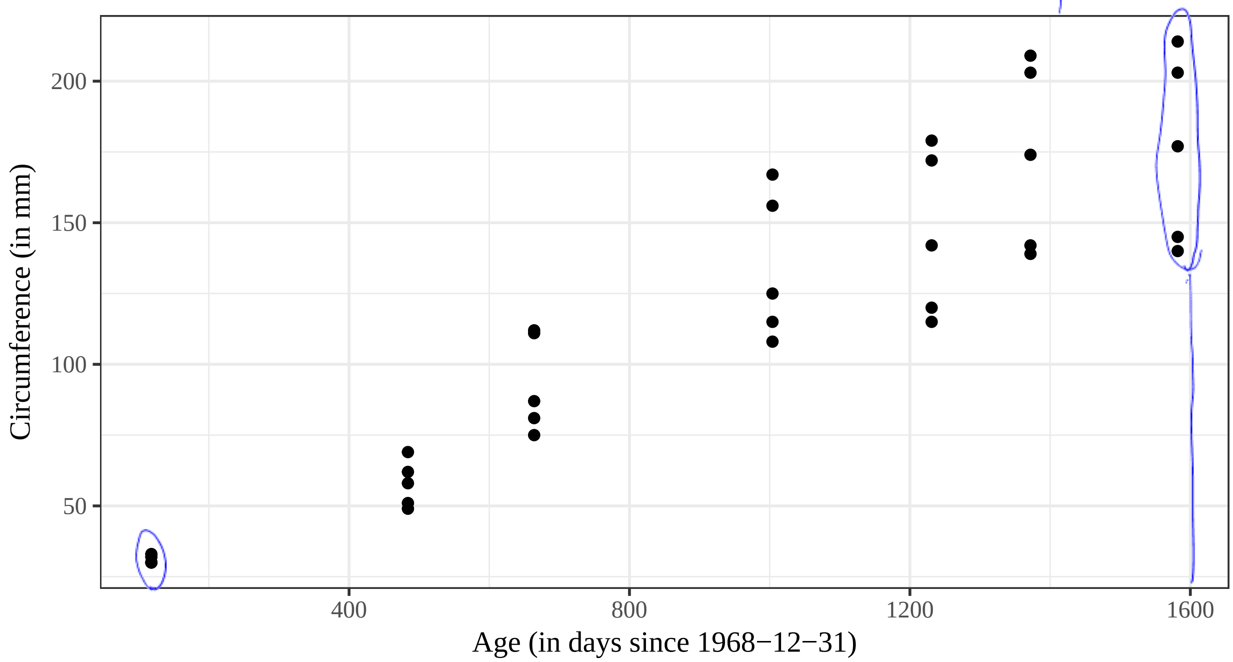
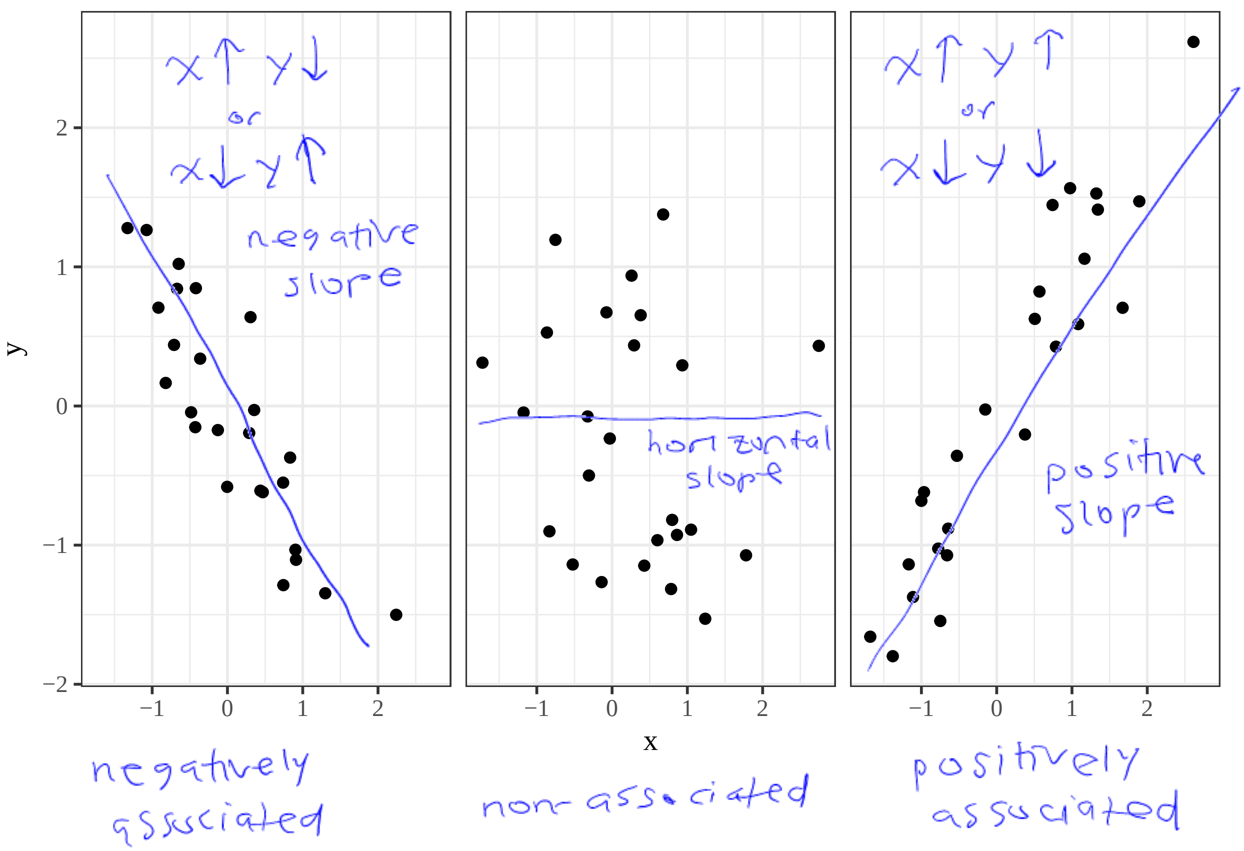


Figure 5: Scatterplot of 35 trees' age and circumference.

older trees associated w/ larger circumferences
(positive association/relationship)
older trees have more variability

There are three typical association/relationship between two variables:



Definition 3.7. A *run chart* is a basic graph that displays data values in a time sequence in the order in which the data were generated.

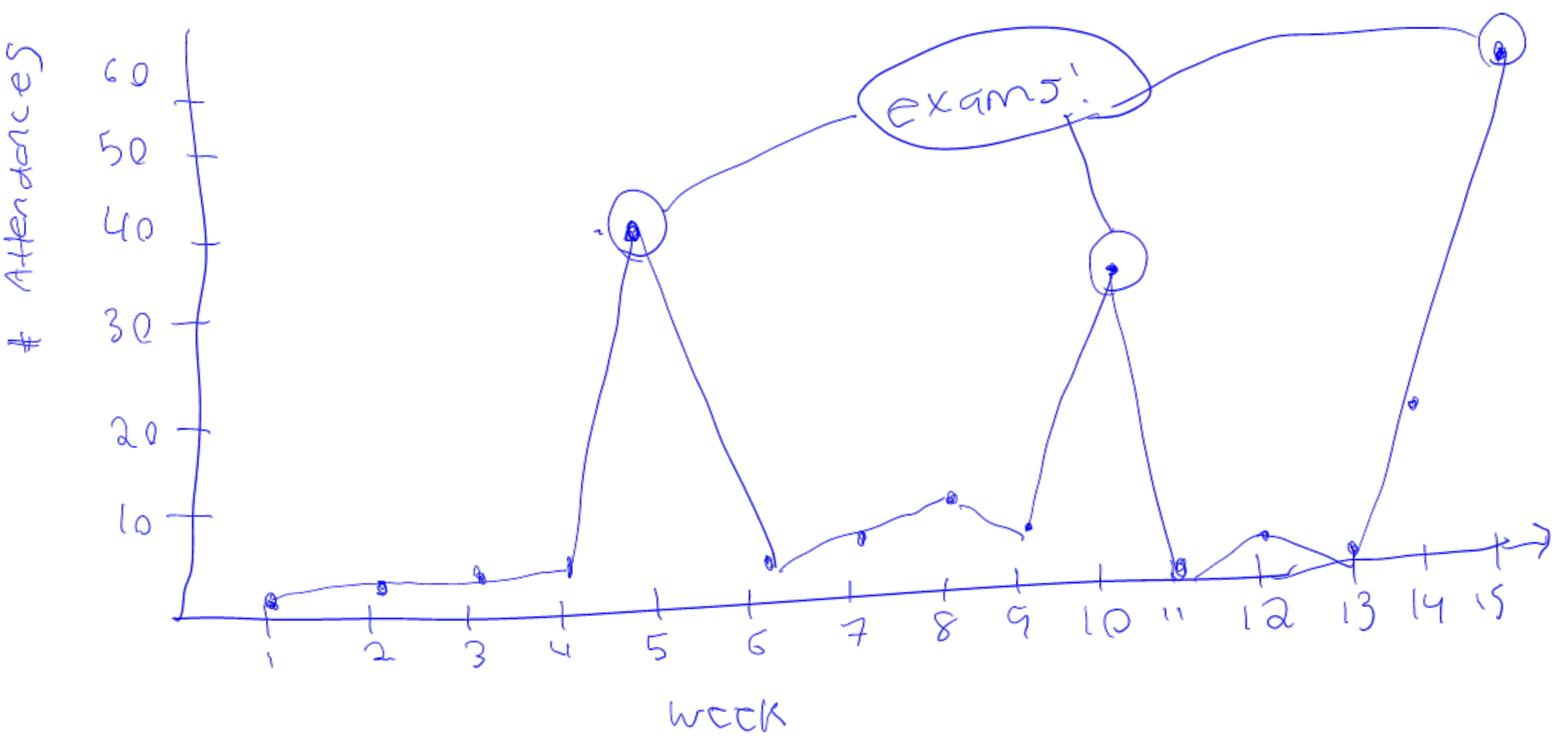
Example 3.9 (Office hours)**.** A professor collects data on the number of students that come to her office hours per week during the course of the semester.

ordered time → quantitative variable

Week	Attendance
1	0.00
2	1.00
3	4.00
4	5.00
5	40.00
6	2.00
7	5.00
8	10.00
9	7.00
10	30.00
11	0.00
12	4.00
13	3.00
14	19.00
15	60.00

see trend over time

Table 5: Weekly attendance in office hours for a semester.



3.2 Quantiles

0% – 100%

Most people are probably familiar with the idea of *percentiles*.

Definition 3.8. The p^{th} *percentile* of a data set is a number greater than $p\%$ of the data and less than the rest.

> "You scored at the 90^{th} percentile on the SAT" means that your score was higher than 90% of the students who took the test and lower than the other 10%

> "Zorbit was positioned at the 80^{th} percentile of the list of fastest growing companies compiled by INC magazine." means Zorbit was growing faster than 80% of the companies in the list and slower than the other 20%.

"quantiles"

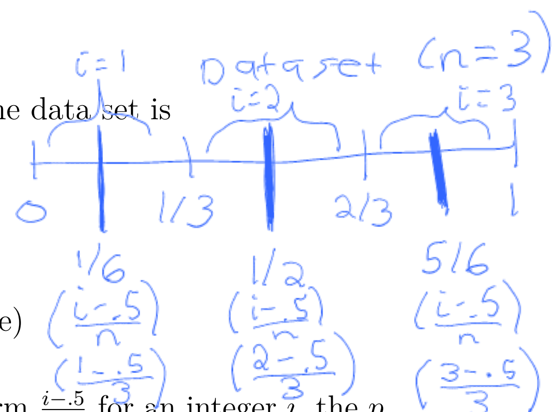
It is often more convenient to work in terms of fractions between 0 and 1 than percentages.

Definition 3.9. For a data set consisting of n values that when ordered are $x_1 \leq x_2 \leq \cdots \leq x_n$,

1. if $p = \frac{i-.5}{n}$ for a positive integer $i \leq n$, the p *quantile* of the data set is

$$Q(p) = Q\left(\frac{i-.5}{n}\right) = x_i$$

(the ith smallest data point will be called the $\frac{i-.5}{n}$ quantile)

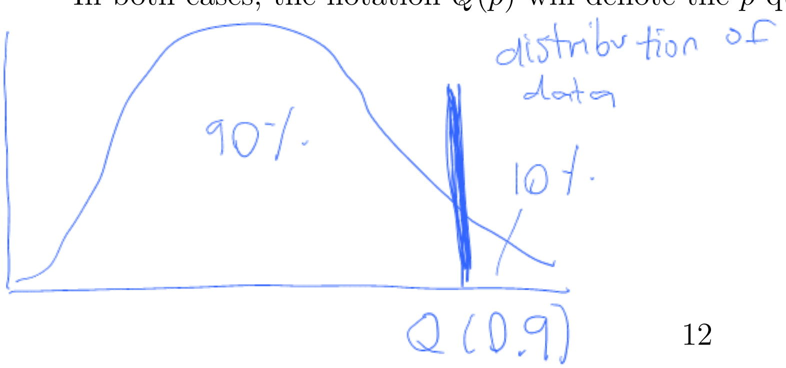


2. for any number p between $\frac{.5}{n}$ and $\frac{n-.5}{n}$ that is not of the form $\frac{i-.5}{n}$ for an integer i, the p *quantile* of the data set will be obtained by linear interpolation between the two values of $Q\left(\frac{i-.5}{n}\right)$ with corresponding $\frac{i-.5}{n}$ that bracket p.

In both cases, the notation $Q(p)$ will denote the p quantile.

p quantile of a distn is a number such that a fraction p of the distn lies to the left and a fraction 1-p of the distn lies to the right.

e.g.



Example 3.10 (Breaking strengths of paper towels, pg. 79). Here is a study of the dry breaking strength (in grams) of generic paper towels. $\boxed{n=10}$

i	$\underbrace{\frac{i-0.5}{n}}_{=p}$	ith smallest data point ($x_i = Q(\frac{i-0.5}{n}) = Q(p)$)
1	0.05	7583
2	0.15	8527
3	0.25	8572
4	0.35	8577
5	0.45	9011
6	0.55	9165
7	0.65	9471
8	0.75	9614
9	0.85	9614
10	0.95	10688

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

Table 6: Ten paper towls breaking strengths (in grams).

$Q(0.85) = 9614$

$Q(0.25) = 8572$

① Let $p_i = \frac{i-0.5}{n}$, $i = 1, 2, \ldots, n$

② Define $Q(p_i) = x_i$, $i = 1, 2, \ldots, n$

Ⓐ If $p = p_j$ from some index j, then $Q(p) = Q(p_j)$

Ⓑ otherwise, linearly interpolate $Q(p)$:

① $i' = np + 0.5$

$\lceil i' \rceil$ = "ceiling" ⇒ next integer above our value

$\lfloor i' \rfloor$ = "floor" ⇒ next integer below our value

ⓘⓘ $Q(p) = (\lceil i' \rceil - i')\, x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)\, x_{\lceil i' \rceil}$

$Q(0.5) = Q(p)$

① $i' = np + 0.5 = (10)(0.5) + 0.5 = 5.5$

$\lceil i' \rceil = 6$

$\lfloor i' \rfloor = 5$

ⓘⓘ $Q(p) = (6 - 5.5)\, x_5 + (5.5 - 5)\, x_6$

$= (0.5)(9011) + (0.5)(9165)$

$= 9088$ grams

$Q(0.5) = \frac{0.5 - 4.5}{5.5 - 4.5} = 0.5$ btwn 0.45 to 0.55

0.5 0.5

0.45 0.5 0.55

$\Rightarrow Q(0.5) = 0.5\, Q(0.45) + 0.5\, Q(0.55)$

$= 0.5(9011) + 0.5(9165)$

$= 9088$ grams

$Q(0.18) = Q(p)$

① $i' = np + 0.5 = (10)(0.18) + 0.5 = 2.3$

$\lceil i' \rceil = 3$

$\lfloor i' \rfloor = 2$

ⓘⓘ $Q(0.18) = (3 - 2.3)\, x_2 + (2.3 - 2)\, x_3$

$= (0.7)(8527) + (0.3)(8572)$

$= 8540.5$ grams

$Q(0.18) = \frac{0.18 - 0.15}{0.25 - 0.15} = 0.3$ btwn 0.15 to 0.25

0.3 0.7

0.15 0.18 0.25

$\Rightarrow Q(0.18) = 0.3\, Q(0.15) + 0.7\, Q(0.25)$

$= 0.3(8572) + 0.7(8527)$

$= 8540.5$ grams

$Q(.94)$

① $i' = np + 0.5 = (10)(.94) + 0.5 = 9.9$

$\lceil i' \rceil = 10$ $\lfloor i' \rfloor = 9$

② $Q(p) = (10 - 9.9)\, x_9 + (9.9 - 9)\, x_{10}$

$= (0.1)(9614) + 0.9(10688)$

$= 10580.6$ g.

$\frac{.94 - .85}{.95 - .85} = .9$ between .85 and .95

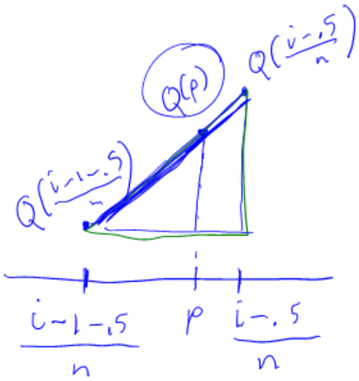
.9 .1

.85 .94 .95

$\Rightarrow Q(.94) = .1\, Q(.85) + .9\, Q(.95)$

$= .1(9614) + .9(10688)$

$= 10580.6$ g.



$$\frac{Q(p) - Q\left(\frac{i-1-.5}{n}\right)}{p - \frac{i-1-.5}{n}} = \frac{Q\left(\frac{i-.5}{n}\right) - Q\left(\frac{i-1-.5}{n}\right)}{\frac{i-.5}{n} - \frac{i-1-.5}{n}}$$

Definition 3.10. $Q\overset{= x_1}{\left(\frac{1-.5}{n}\right)}$ is called the *minimum* and $Q\overset{= x_n}{\left(\frac{n-.5}{n}\right)}$ is called the *maximum* of a distribution.

Definition 3.11. $Q(.5)$ is called the *median* of a distribution. $Q(.25)$ and $Q(.75)$ are called the *first (or lower) quartile* and *third (or upper) quartile* of a distribution, respectively.

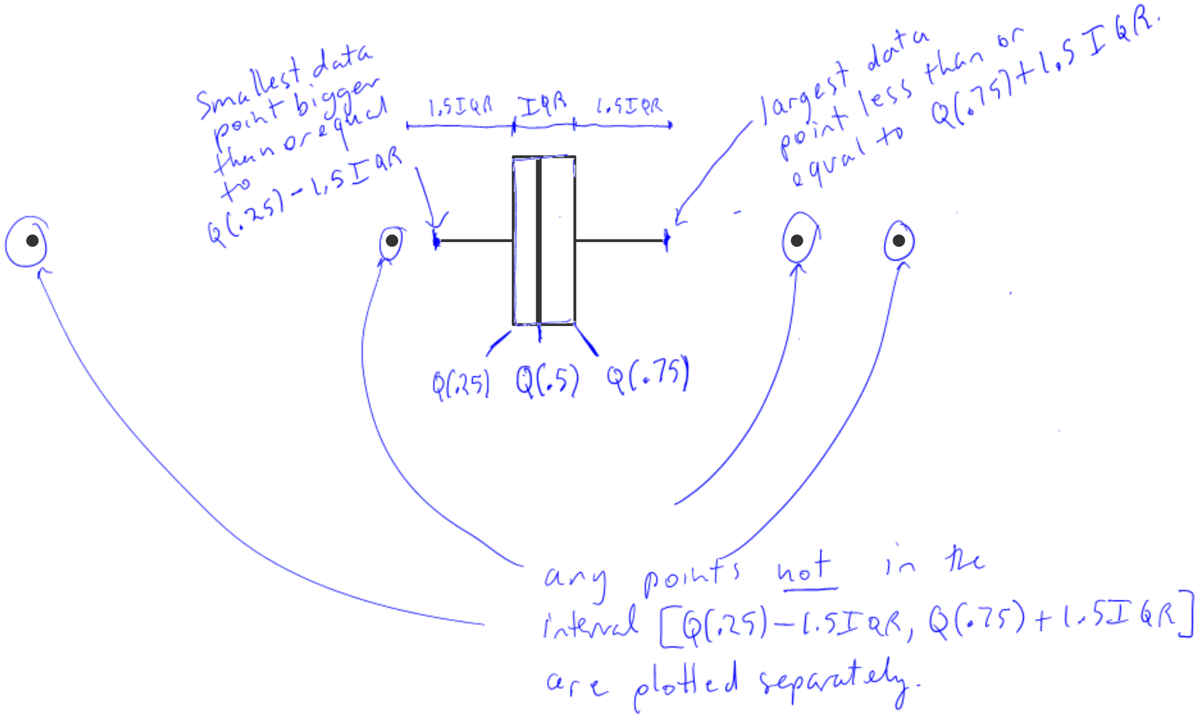
Definition 3.12. The *interquartile range (IQR)* is defined as $IQR = Q(.75) - Q(.25)$.

Definition 3.13. An *outlier* is a data point that is larger than $Q(.75) + 1.5 * IQR$ or smaller than $Q(.25) - 1.5 * IQR$.

Most points should be below $Q(.75) + 1.5\, IQR$ or above $Q(.25) - 1.5\, IQR$.

3.2.1 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.



Example 3.11 (Bullet penetration depths, cont'd).

i	$\frac{i-.5}{20}$	200 grain bullets	230 grain bullets
1	0.025	58.000	27.750
2	0.075	58.650	37.350
3	0.125	59.100	38.350
4	0.175	59.500	38.350
5	0.225	59.800	38.750
6	0.275	60.700	39.750
7	0.325	61.300	40.500
8	0.375	61.500	41.000
9	0.425	62.300	41.150
10	0.475	62.650	42.550
11	0.525	62.950	42.900
12	0.575	63.300	43.600
13	0.625	63.550	43.850
14	0.675	63.800	47.300
15	0.725	64.050	47.900
16	0.775	64.650	48.150
17	0.825	65.000	49.850
18	0.875	67.750	51.250
19	0.925	70.400 ←	51.600
20	0.975	71.700	56.000

Table 7: Quantiles of the bullet penetration depth distributions.

create a boxplot for 230 grain bullets

$Q(.25) = .5Q(.225) + .5Q(.275) = .5(38.75) + .5(39.75) = 39.25$ mm

$Q(.5) = .5Q(.475) + .5Q(.525) = .5(42.55) + .5(42.9) = 42.725$ mm

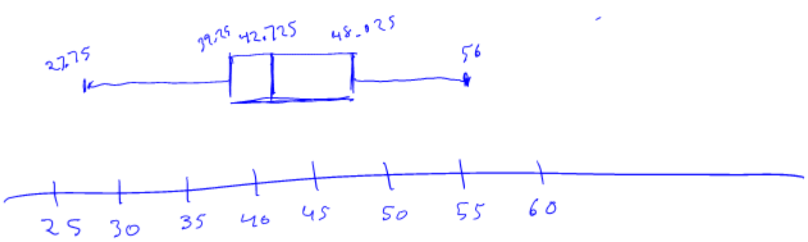
$Q(.75) = .5Q(.725) + .5Q(.775) \doteq .5(47.9) + .5(48.15) = 48.625$ mm

$IQR = Q(.75) - Q(.25) = 48.025 - 39.25 = 8.775$

$1.5\,IQR = 13.163$

$$\left.\begin{array}{l} Q(.75) + 1.5\,IQR = 61.88\text{mm} \\ Q(.25) - 1.5\,IQR = 26.087\text{mm} \end{array}\right\} \Rightarrow \text{no outliers}$$

230 grain bullets



do the same for 200 grain bullets and plot the two distributions (boxplots) side by side

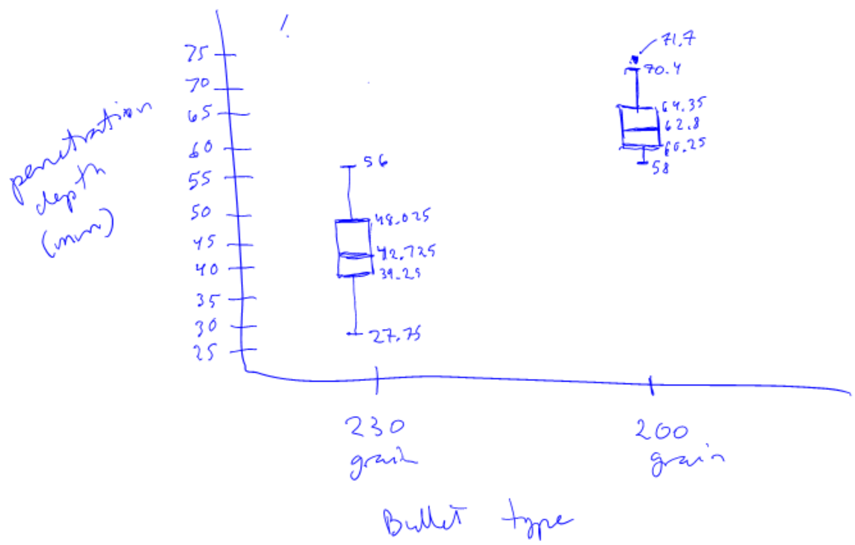
200 grain bullets

$$Q(.25) = .5(59.8) + .5(60.7) = 66.25$$
$$Q(.5) = .5(62.65) + .5(62.95) = 62.8$$
$$Q(.75) = .5(64.05) + .5(64.65) = 64.35$$

$$IQR = 64.35 - 60.25 = 4.1$$

$$1.5\,IQR = 6.15$$

$$Q(.75) + 1.5\,IQR = 70.5$$
$$Q(.25) - 1.5\,IQR = 54.1$$



3.2.2 Quantile-quantile (Q-Q) plots

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

We could compare two histograms, dotplots, boxplots.

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

"equal shape" = linearly related quantile functions.

Definition 3.14. A *Q-Q plot* for two data sets with respective quantile functions Q_1 and Q_2 is a plot of ordered pairs (scatterplot) $(Q_1(p), Q_2(p))$ 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 $i = 1, \ldots, n$.

two univariate, numeric, same size.

Example 3.12 (Bullet penetration depth, cont'd)**.**

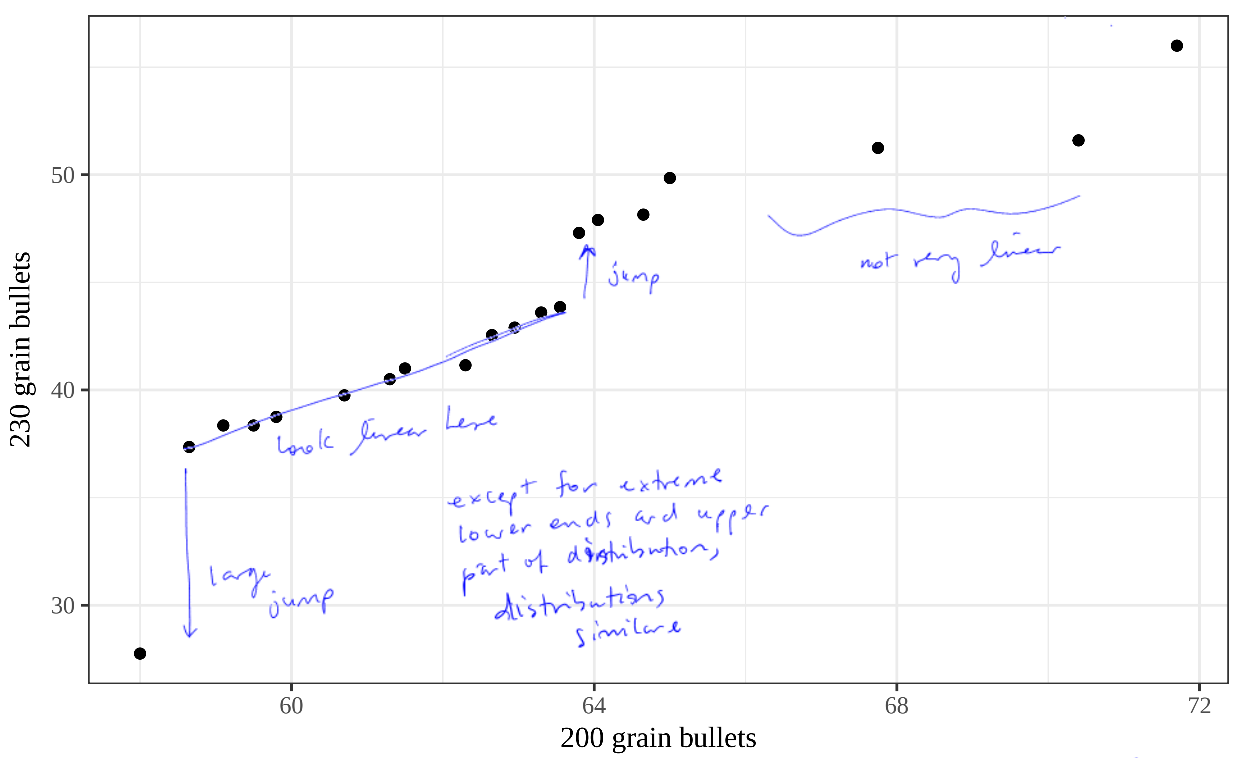


Figure 6: Q-Q plot for the bullet penetration depths.

overall, not super linear.

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 coordinates coming from the first data set and the vertical ones from the second.

Example 3.13 (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

Table 8: Two artificial data sets

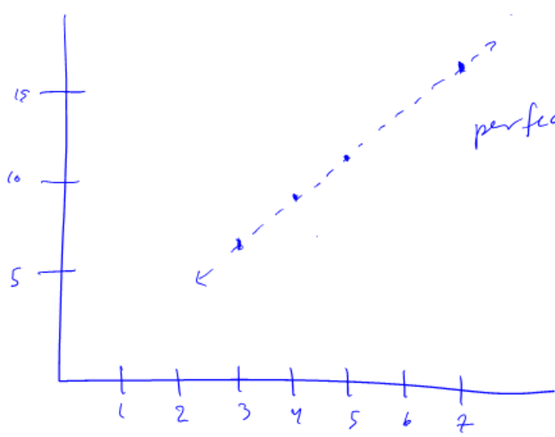
1, Order

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

2. pair off.

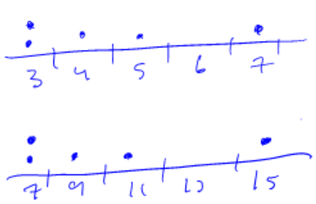
(3,7)
(3,7)
⋮
(7,15)

3, scatter plot



perfectly linear => exact same distributional shape.

Dot diagrams:



scale/location differ, but same shape.

3.2.3 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*.

Definition 3.15. 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 $i = 1, \ldots, n$.

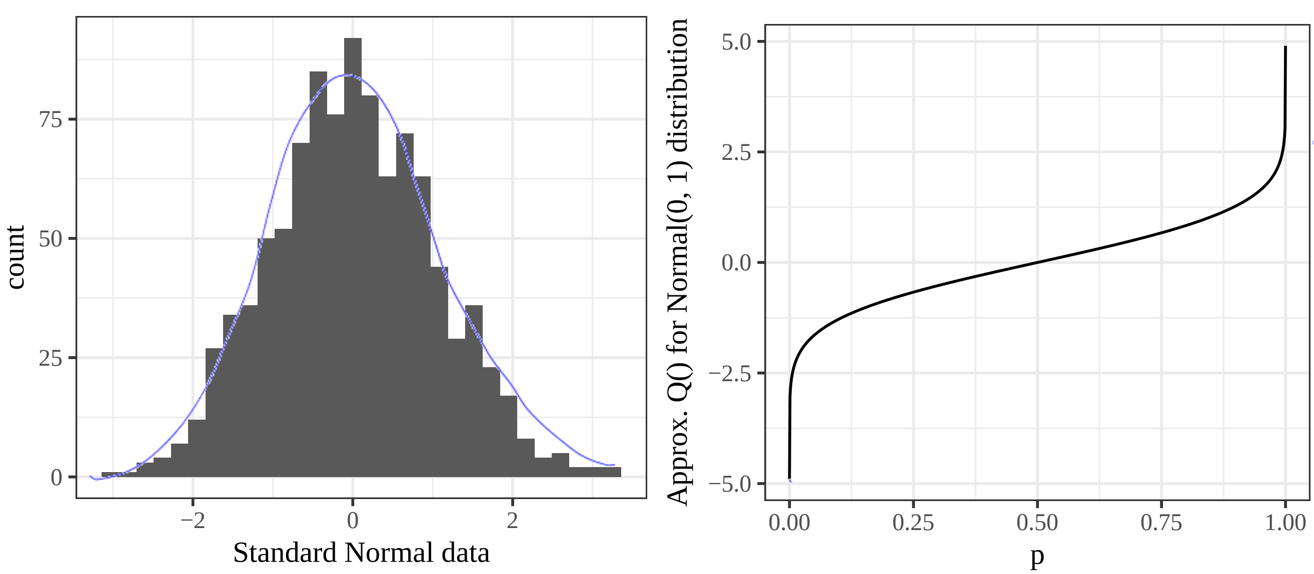
bell shaped

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

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

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*.

"normal plot"



Example 3.14 (Breaking strengths of paper towels, cont'd).

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

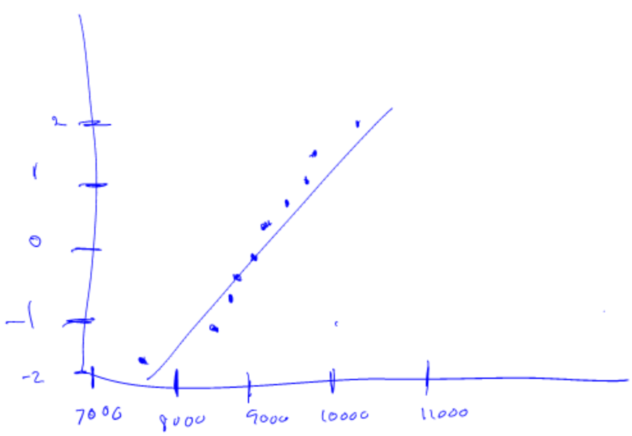
$p = \frac{i-.5}{10}$, $i = 1, \ldots, 10$.

ordered data

i	$\frac{i-.5}{10}$	Breaking strength Q()	Standard Normal Q()
1	0.05	7583	-1.64
2	0.15	8527	-1.04
3	0.25	8572	-0.67
4	0.35	8577	-0.39
5	0.45	9011	-0.13
6	0.55	9165	0.13
7	0.65	9471	0.39
8	0.75	9614	0.67
9	0.85	9614	1.04
10	0.95	10688	1.64

(7583, −1.64)
(8527, −1.04)
(10,688, 1.64)

Table 9: Breaking strength and standard Normal quantiles.



roughly straight
$\Rightarrow$ breaking strength data is roughly bell shaped.

3.3 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

For example keep daily average and range measurements.

Definition 3.16. A *numerical summary (or statistic)* is a number or list of numbers calculated using the data (and only the data).

3.3.1 Measures of location

or 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*.

Definition 3.17. The *(arithmetic) mean* of a sample of quantitative data $(x_1, \ldots, x_n)$ is

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

Definition 3.18. 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.

$\lceil \cdot \rceil$ ceiling function, smallest integer greater than or equal to the argument

i.e. $\lceil 0.2 \rceil = 1$

$\lceil 1 \rceil = 1$

Example 3.15 (Measures of location)**.** Calculate the three measures of location for the following data.

$$0, 1, 1, 2, 3, 5$$

3.3.2 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.

Definition 3.19. The *range* of a data set consisting of ordered values $x_1 \leq \cdots \leq x_n$ is

$$R = x_n - x_1.$$

Definition 3.20. The *sample variance* of a data set consisting of values $x_1, \ldots, x_n$ is

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

The *sample standard deviation*, s, is the nonnegative square root of the sample variance.

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

Example 3.16 (Measures of spread)**.** Calculate the four measures of spread for the following data.

$$0, 1, 1, 2, 3, 5$$

Example 3.17 (Sensitivity to outliers)**.** Which measures of center and spread differ drastically between the x_is and the y_is? Which ones are the same?

$$x_i : 0, 1, 1, 2, 3, 5$$
$$y_i : 0, 1, 1, 2, 3, 817263489$$

3.3.3 Statistics and parameters

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

Definition 3.21. Numerical summarizations of sample data are called (sample) *statistics*. Numerical summarizations of population and theoretical distributions are called (population or model) *parameters*.

Definition 3.22. If a data set, $x_1, \ldots, x_N$, represents an entire population, then the *population (or true) mean* is defined as

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

Definition 3.23. If a data set, $x_1, \ldots, x_N$, 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*, σ is the nonnegative square root of σ^2.

3.4 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 3.18 (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.

	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
...	...	...	...	...	...	...	...	...	...	...	...

Table 10: Car data for 1973-1974 models.

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

cyl	Frequency	Relative Frequency	Cumulative Frequency
4.00	11	0.34	0.34
6.00	7	0.22	0.56
8.00	14	0.44	1.00

Table 11: Frequency table for car cylinders.

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

Definition 3.24. A *bar plot* presents categorical data with rectangular bars with lengths proportional to the values that they represent (usually frequency of occurrence).

Example 3.19 (Cars, cont'd)**.**