Descriptive statistics

Table of contents

Graphical Summaries Table

Categorical

Categorical variables are variables that are qualitative. When given categorical variables, we generally use a frequency table.

Given the research question of if more men died in the sinking of *Titanic* then women, we will get the variables *gender* and *survived*. We can represent these conditional events in a two-way table:

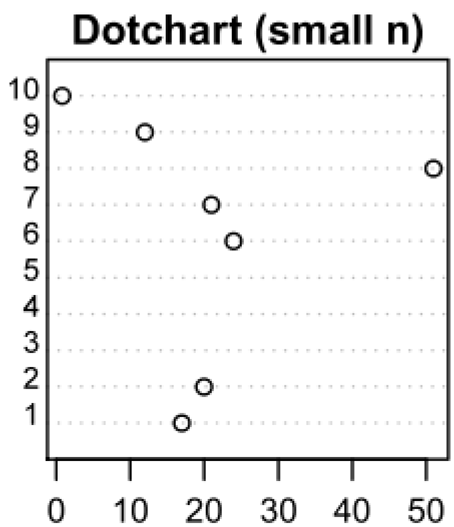
	Survived	Died
Male	142	709
Female	308	154

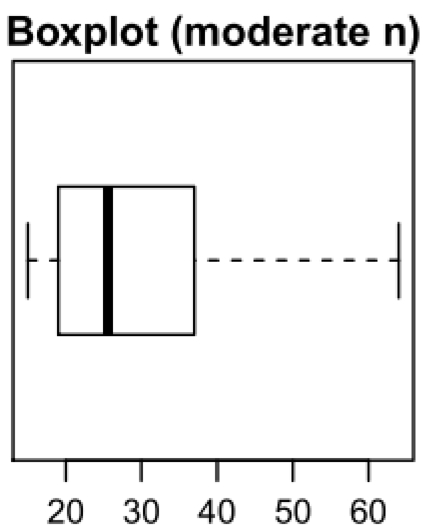
Quantitative

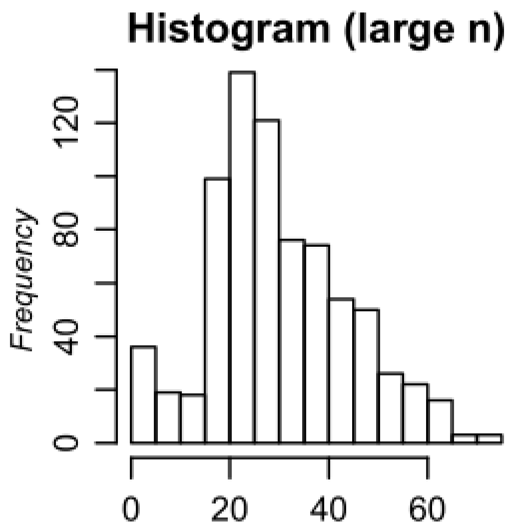
For quantitative variables, we want to consider three main things:

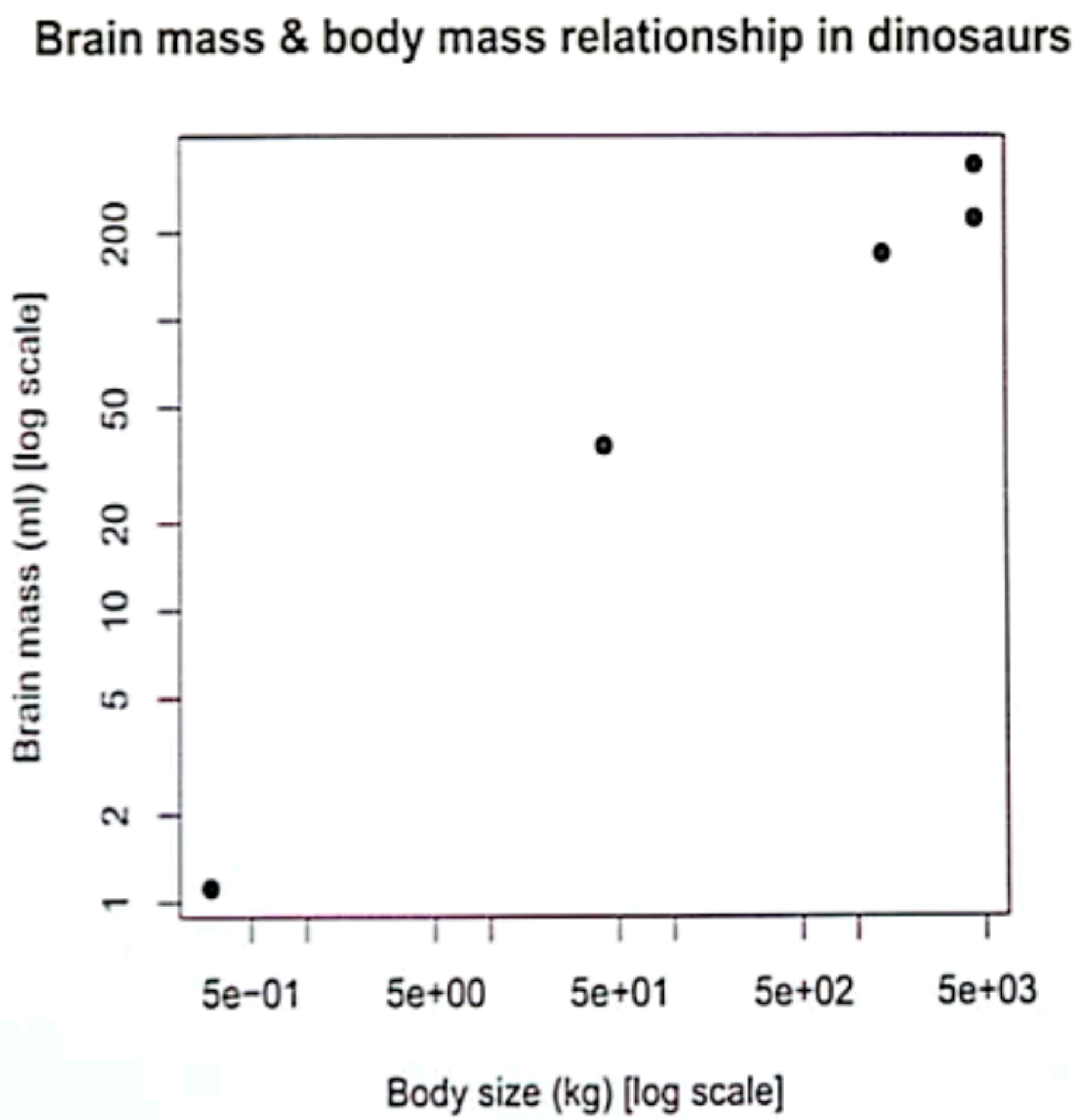
- Location: which is a measure of where *most* of our data lies around
- Spread: how our variables are distributed in relation to our location
- Shape: the general shape of the distribution

Graphical and numerical summaries of quantitative data






Histogram (large n)
Frequency
0
40
80
120
0
20
40
60


Brain mass & body mass relationship in dinosaurs
Brain mass (ml) [log scale]
1
2
5
10
20
50
200
5e−01
5e+00
5e+01
5e+02
5e+03
Body size (kg) [log scale]

For small n, we use the dotchart, which has the range of values on the x-axis and then the frequency of that value on the y-axis.

For medium n, we use the boxplot, which gives us the the IQR (the box), with the thick line representing the median. The whiskers represent 1.5x IQR in both directions - values outside of this range are considered outliers.

For large n, we use a histogram. Histograms use ranges of values and group them for frequency. Approximately, there will be
$\sqrt{n}$ bars/buckets, which we group these values into. The values fit into a range given by the amount of buckets dividing the total range equally.

The *kernel density estimator (KDE)*:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - x_i)$$

for some weighting $w(x)$, which is usually normal density with mean 0 and standard deviation h. How to choose bandwith h has a lot of research behind it which is beyond the scope of this course. It basically provides a **smooth histogram**.

For associations between **two quantitative variables**, we use a scatter plot, which just put dots on the (x, y) pairings of data. This lets us consider if there is a relationship between the two variables, which is commonly called regression.

Location formulas/descriptions

The *sample mean:*

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

is a natural measure of location of a quantitative variable.

The *sample median*:

$$\tilde{x}_{0.5} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}) & \text{if } n \text{ is even} \end{cases}$$

Spread formulas/descriptions

The *sample variance:*

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

is a common measure a spread. Let us justify some parts of variance

- We divide by $\frac{1}{n-1}$ due to the need of an unbiased estimator.
- We square the distance from the mean - as the sum without the square sums to zero, and we also ensure all negative spreads become positive

The *sample standard deviation*:

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

which measures the average "distance" from the mean.

The *interquartile range (IQR)*:

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

Shape formulas/descriptions

Skewness explains how the distribution is *tailed* - **right skewness** explains a distribution with a long right tail, and vice versa. Skewness can be estimated by:

$$\hat{\kappa}_1 = \frac{1}{(n-1)s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Thus, is κ is large, this indicates our graph is skewed. Interestingly, skewness gives us insights into *outliers* - outliers must be investigated; are they special cases, extreme events or just data errors?

Summarising associations between two quantitative variables

Consider a pair of samples from two quantitative variables:

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}$$

We would like to understand how the x and y variables are related. Analysis of two quantitative variables is commonly called *regression* (likely **linear** regression). We largely use *scatterplots* (which is explained above in the graphical summary part).

An effective **numerical** summary of the *linear* relationship between two quantitative variables is the **correlation coefficient (r)**:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $\bar{a}$ and s_a is the sample mean and sample s.d for the variable a. To explain the formula:

- We find the deviation for each point from the mean, and "scale" it with the standard deviation.
- We then take the scaled product difference between x and y, then scale it again by $\frac{1}{n-1}$, as we have lost a degree of freedom.
- As $|r| \to 1$, we get a stronger correlation between the two variables.

Transforming data

Transforming data is usually done to change the scale data is measured on, as well as improve data properties.

Linear Transformation

A *linear transformation* of a sample from a quantitative variable, from $\{x_1, x_2, \ldots, x_n\}$ to $\{y_1, y_2, \ldots, y_n\}$ satisfies:

$$y_i = a + bx_i$$

for each i and $b \neq 0$. Given x and y, we are finding the *linear* equation that transforms it to y.

Linear transformations can have effects on our data and statistics. For some *measure of location* m_y, we will get:

$$m_y = a + bm_x$$

then we say that m is a measure of location.

If m_x is a *measure of spread* in the same units as x:

$$m_y = |b| m_x$$

If m_x is a *measure of shape* then:

$$m_y = \begin{cases} m_x & \text{if } b > 0 \\ -m_x & \text{if } b < 0 \end{cases}$$

Explanation:
Thus, as the data of x changes, the measure of location *moves with the data*. For measures of spread - the spread itself has no notion of location; but the *scale* is changing. A measure of

shape should be invariant under any change of scale (if a circle get 2x bigger, it's still a circle).
The above can be proven by applying $y_i = a + bx_i$ into the formulas.

Example: Dinosaur body mass (x) was measured in kilograms.
If we transforms the body mass data into grams instead (y), how will the mean body mass, standard deviation and correlation be calculated from y relative to how they were measured in x ?
First, declare that:

$$y_i = 1000x_i$$

- $\bar{y} = 1000\bar{x}$
- $s_y = 1000s_x$
- $r_y = r_x$

What if it was a y was log transformed, ergo $y_i = log x_i$?

z-score

The z-score, or standardised score of a quantitative variable is definsed as:

$$z = \frac{x - \bar{x}}{s_x}$$

The z-score is a measure of unusualness; $z = 1$ shows one standard deviation away, and etc.