Formulas

Stephanie Ranft S2459825

October 2, 2019

Statistics 1A
PSBE1-08

1 Descriptive statistics

1.1 The mean

(1) gives the formula for the mean of a sample of size n, where the true population has mean μ.
If you were to repeat your sampling (from the same population) a number of times, then you would have the sampling distribution of $\bar{y}$. If y is normally distributed with mean μ and standard deviation σ, i.e. $y \sim \mathcal{N}(\mu, \sigma^2)$, then the sampling distribution of the mean of y is given by $\bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$. This tells us that, if you were to take 500 samples (for example) of y, then calculate the mean of every sample $\bar{y}_j$, where $j = 1, 2, 3, \ldots, 500$, then the mean of the means $\bar{y}$ will be distributed normally with mean μ and standard deviation $\sigma/\sqrt{n}$. This seems a little confusing, however we can break it down here. If we draw one sample, then we have $\bar{y} = \sum y_i/n$. If we draw N samples (before, we had $N = 500$), then we have that the mean ($\bar{y}$) of the means ($\bar{y}_j$) is calculated by (2). So, you now have a sample of means $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N)$ and the distribution (the sampling distribution of the mean of y) is normal with mean μ and standard deviation $\sigma/\sqrt{n}$.
To clarify, $i = 1, 2, \ldots, n$ indexes the number of y's in any particular sample, and $j = 1, 2, \ldots, N$ indexes the sample. So, $y_{3,4}$ is the third value in sample 4.
You can play around with the sampling distribution by using this website.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1}$$

$$\bar{y} = \frac{\sum_{j=1}^{N} \bar{y}_j}{N} = \sum_{j=1}^{N} \frac{\sum_{i=1}^{n} y_{ij}}{n \cdot N} \tag{2}$$

Figure 2

1.2 Standard deviation

The intrinsic meanings of the words "standard" and "deviation" give you a great understanding of what this statistic is measuring: the average or standard distance (or deviation in distance) between any data point and the other data points in the sample. It measures dispersion of the data and gives the statistician an idea of how spread out their data set is, when they imagine it in a graph, for instance. It is calculated (literally) as the average distance between each data point in the sample and the sample mean, and the formula is given in (3).

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \tag{3}$$

The reason that we divide by $n-1$ instead of n is that we can be sure that the sample standard deviation s_y is an unbiased estimator of the population standard deviation σ. This is due to the shape of the square-root graph and not a part of your course. To think of this intuitively, we have the concept of "degrees of freedom": in a sample of n participants, once you have handed out $n-1$ name tags, the last name tag **has** to go to the nth person. For example, if I have 3 different exam papers and I hand one to each student; after I have handed out two of the three, then the last exam paper must go to the third person, hence 2 degrees of freedom. I have the freedom to assign a place to each of the $n-1$ data points in the sample, however once those $n-1$ spots have been filled, there is no freedom for the nth data point, therefore we have $n-1$ degrees of freedom.

Another way to think of degrees of freedom is imagine the n data points as n people standing along a ruler (comically huge ruler), and each person is standing at their height on the ruler, e.g. I would stand at the 167cm mark on the ruler. Once all n people are standing, I calculate their average distance from each other (the mean) and then the average distance each person has from the mean. With everyone standing in a straight line, there are $n-1$ gaps! Think about it, if you have a group of 5 people with spaces apart, and you count 4 spaces! So

when you are calculating the standard deviation for a sample, remember that this is always a start and finish to the "line" of data points, and between each is a gap; if there are n data points then there are $n-1$ gaps!

In a population, it is usually not possible to calculate the standard deviation (due to the large number of data points) however it is still a measure of dispersion; more on population standard deviation later. With respect to the previous example regarding gaps between people: if you have a line of people, and the line goes on forever, do you have a start or finish to the line? If you don't have a start or finish to a line of people, how do you measure the total number of gaps between each person? An easy way to do that would be to say that each person has to have a space behind them, so then the number of gaps equals the number of people.

2 Probability distributions

2.1 Symmetry about the mean

When we assume that the distribution of our random variable X is roughly normal with mean μ and standard deviation σ, we write:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \implies \mathbb{P}\left(X \leq x \,\middle|\, \mu, \sigma^2\right) = p. \tag{4}$$

In words, we describe a model for X (height of women, effectiveness of a therapy, etc.) which predicts the outcome of an event x (160 cm, effective, etc.) denoted as a probability p, based on the average of an infinite data set μ and how spread out the data points are σ.
It is important to remember that the bell-shaped graph is **symmetrical about the mean**:

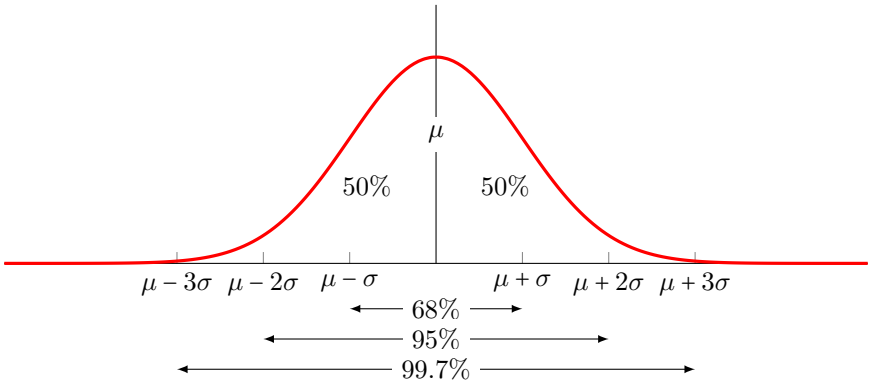


Figure 3: A normally distributed random variable X with mean μ and variance σ^2. 50% of the data lies either side of μ, and we can approximate the middle proportion using the 68-95-99.7 rule.

Using Figure 3, we can calculate probabilities with the use of a table, for example if you were asked about the lowest 16% of values for X, you would look for the x which satisfies:

$$\mathbb{P}\left(X \leq x\right) = 0.16. \tag{5}$$

$$\implies \mathbb{P}\left(X \leq x\right) + \mathbb{P}\left(X \geq \mu + (\mu - x)\right) = 0.16 + 0.16 \tag{6}$$

$$\mathbb{P}\left(X \leq x\right) + \mathbb{P}\left(X \geq 2\mu - x\right) = 0.32. \tag{7}$$

The previous conclusion comes from the symmetry about the mean: the area under the graph on the left of μ is equal to the area to the right. Thus, the probabilities are equal and, moreover, we know the entire area under the graph is equal to 1. Given that we have calculated the two outside extremes, what is the chunk in the middle?

$$\implies \mathbb{P}\left(x \leq X \leq 2\mu - x\right) = 1 - 0.32 = 0.68. \tag{8}$$

Now we know that we are looking for an x which is 1 standard deviations away from the mean, in this particular case we want the lowest so $x = \mu - \sigma$; if we wanted the top 16%, we would have $x = \mu + \sigma$. I've summarised information associated with this in the following table:

high/low	%	x
highest / lowest	50	μ
highest	16	$\mu+\sigma$
lowest		$\mu-\sigma$
highest	2.5	$\mu+2\sigma$
lowest		$\mu-2\sigma$
highest	0.15	$\mu+3\sigma$
lowest		$\mu-3\sigma$

Table 1: Associated x values for lowest/highest p-value ($p \times 100\%$).

Now that we have found the lowest 16%, what is left?

$$\mathbb{P}(X \leq \mu - \sigma) = 0.16. \tag{9}$$
$$\implies \mathbb{P}(X \geq \mu - \sigma) = 1 - 0.16 = 0.84. \tag{10}$$
$$= \mathbb{P}(X \leq \mu + \sigma) \tag{11}$$

This next picture (Figure 4) should help to visualise this.

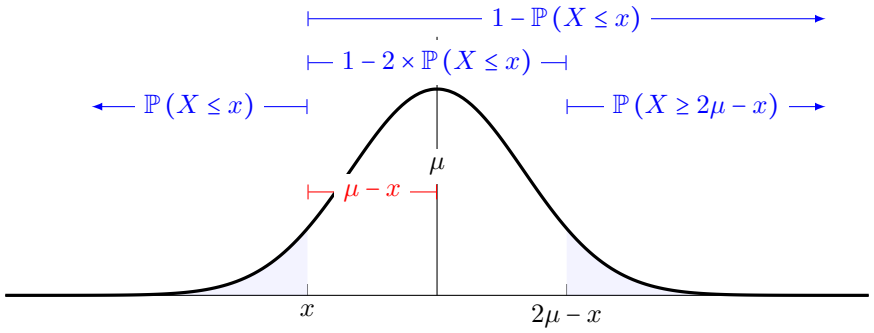


Figure 4

If we are not interested in any of the p-values given in Table 1, then how do we find x? The next section on z-scores will provide some illumination on the subject.

2.2 z-score

Assume, as we did before, that $X \sim \mathcal{N}(\mu, \sigma^2)$ and we want to find the x which satisfies $\mathbb{P}(X \leq x) = p$. We check Table 1, but it is not listed!! In order to find such an x, we need to refer to the z-table; say we wish to find the lowest 30% ($p = 0.300$). We find that this p-value lies somewhere between $z = -0.52$ and -0.53, and choose the most "conservative" z, i.e. the one closest to zero ($z = -0.52$).
If we only have a positive z-table, we need to find the z which corresponds to $p = 1 - 0.300 = 0.700$, which is $z = 0.52$. (Notice how z is symmetrical around the mean $\mu = 0$?) **Remember:** if you are looking for the an x corresponding to a p-value which is less than 0.5, then you are looking for a negative z-score. So, if you **only** have the positive z-table, look the z corresponding to $(1-p)$ and **remember** it is negative z that corresponds to p.
The next step is transform our z to x; in the past, you have transformed your x to a z-score, so this is exactly the converse.

$$\frac{x - \mu}{\sigma} = z \tag{12}$$

Suppose we have $\mu = 15$ and $\sigma = 2.2$. Substitute in $z = -0.52$, and μ and σ:

$$\frac{x - 15}{2.2} = -0.52 \tag{13}$$
$$\implies x - 15 = 2.2 \times (-0.52) = -1.144 \tag{14}$$
$$\implies x = 15 + (-1.144) = 13.856. \tag{15}$$

Therefore, if our random variable X is distributed normally with mean 15 and standard deviation 2.2, then the lowest 30% of values lies below 13.856. If we wanted to know the highest 30% of values ($p = 0.700$), we only need to change our negative z to positive:

$$\implies x = \overbrace{15 + 2.2 \times 0.52}^{\mu+\sigma\times z^*_{p=0.7}} = 15 + 1.144 = 16.144. \tag{16}$$

So, we have that the highest 30% of values lie above 16.144.

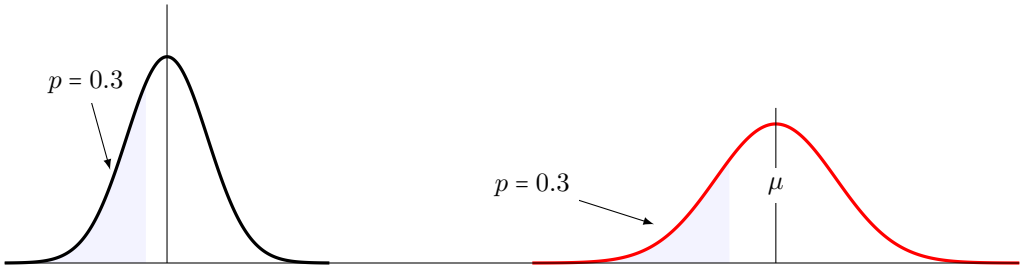


Figure 5: The graph on the right (red) is of $X \sim \mathcal{N}\left(\mu = 15, \sigma^2 = 2.2^2\right)$ and on the left (black) is $Z \sim \mathcal{N}(0, 1)$ (the standard normal distribution. The area under each graph corresponding to $p = 0.3$ is shaded.

Can you notice in Figure 5 that the shaded area on the left appears bigger than the shaded area on the right? This has to do with the larger variance for the one on the right; it is more spread out.
I will now summarise the steps:

1. Draw a quick sketch of your graph, clearly displaying the mean and the proportion you want to find. **Remember:** if you're looking for the highest proportion p, then you need to find an x satisfying

$$p = \mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x) \tag{17}$$
$$\implies \mathbb{P}(X \leq x) = 1 - p. \tag{18}$$

   Similarly, for the lowest proportion you are looking for an x satisfying $\mathbb{P}(X \leq x) = p$. If you are looking for a middle proportion p, consult Figure 4 and you will note that you need to find an x satisfying,

$$\mathbb{P}(X \leq x) = \frac{1 - p}{2}. \tag{19}$$

2. If, on your sketch, the x you seek is on the left of the mean, then you are looking for a **negative** z. Conversely, if x is on the right of the mean, then you are looking for a **positive** z.

3. Once you have ascertained which p-value corresponds to the z-table you have (positive/negative, left-/right-tailed), locate the z corresponding to your p. Substitute this z in the following equation, and solve for x:

$$x = \mu + \sigma \times z. \tag{20}$$

   N.B.: if you are looking for the middle proportion, so you will need to do this for both positive and negative z, resulting in two x's; a lower and an upper bound.

4. Now you can answer your original question:

$$\textit{The (highest/middle/lowest)} \underbrace{\quad}_{p} \textit{ proportion of values lie (above/between/below)} \underbrace{\quad}_{x}.$$

2.3 Normal distribution

You've already been working with the Normal distribution through your transformations to z-scores. Further, you have learnt about the Central Limit Theorem and assumptions of normality, which has aided in you statistical analysis. Aided by the CLT, if you have two random variables, X and Y, both independently and normally distributed, you are able to perform transformations to another random variable, say W, which is also normally distributed. For example, let

$$\left.\begin{array}{l} X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \\ Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \end{array}\right\} \implies W = a \times X + b \times Y \implies \begin{cases} \mu_W = a \times \mu_X + b \times \mu_Y \\ \sigma_W^2 = a^2 \times \sigma_X^2 + b^2 \times \sigma_Y^2 \end{cases} \implies W \sim \mathcal{N}\left(\mu_W, \sigma_W^2\right) \tag{21}$$

What (21) says is that if we have that X has mean μ_X and standard deviation σ_X, and Y has mean μ_Y and standard deviation σ_Y. Suppose we want to compare these distributions in some way, for example if we wanted to know the probability of the values of X being higher than the values of Y, we would look at the difference $W = X - Y$. So, $a = 1$ and $b = -1$, therefore W is distributed normally with mean $\mu_W = \mu_X - \mu_Y$ and standard deviation $\sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2}$. Then you could calculate your p-value as follows:

$$p = \mathbb{P}(X > Y) = \mathbb{P}(X - Y > 0) = \mathbb{P}(W > 0) = \mathbb{P}\left(\frac{W - \mu_W}{\sigma_W} > \frac{0 - \mu_W}{\sigma_W}\right) = \mathbb{P}\left(Z > -\frac{\mu_W}{\sigma_W}\right), \tag{22}$$

$$\text{where} \quad \begin{cases} \mu_W = \mu_X - \mu_Y \\ \sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2} \end{cases}$$

You can do this operation on any linear transformation of X and Y, **because they are normally distributed**. One of the homework questions was about the scores of female F and male M students, and how to compare them. We consider a different distribution of scores, such that $F \sim \mathcal{N}(112, 9)$ and $M \sim \mathcal{N}(106, 16)$, and want to know what's the probability that a woman got a score over 125, given that she scored worse than a man. If we translate these words into an equation, we're looking for the p-value satisfying:

$$p = \mathbb{P}(F > 125 \mid F < M). \tag{23}$$

What proportion of women scored worse than men? You construct you new variable W as follows:

$$W = F - M \implies \begin{cases} \mu_W = \mu_F - \mu_M \\ \qquad = 112 - 106 \\ \qquad = 6 \\ \\ \sigma_W = \sqrt{\sigma_F^2 + \sigma_M^2} \\ \qquad = \sqrt{9 + 16} \\ \qquad = \sqrt{25} \\ \qquad = 5 \end{cases} \tag{24}$$

$$\implies W \sim \mathcal{N}\left(6, 5^2\right) \tag{25}$$

The difference between a male's and female's score (females minus males) is distributed normally with a mean of 6 and a standard deviation of 5. This says that if you were to choose a random male and a random female from either group, and then compare their scores by subtracting the male's score from the female's, you expect the difference to 6 points, however it would be normal that the difference fluctuate between 1 and 11 points.

$$\implies \mathbb{P}(F < M) = \mathbb{P}(W < 0) \tag{26}$$

$$= \mathbb{P}\left(\frac{W - \mu_W}{\sigma_W} < \frac{0 - \mu_W}{\sigma_W}\right) \tag{27}$$

$$= \mathbb{P}\left(Z < -\frac{6}{5}\right) \tag{28}$$

$$= \mathbb{P}(Z < -1.20) \tag{29}$$

We find the p-value from the positive z-table, first by finding the p corresponding to positive $z = 1.20$, which is 0.8849. Then we have

$$\mathbb{P}(Z < -1.20) = \mathbb{P}(Z > 1.20) \tag{30}$$

$$= 1 - \mathbb{P}(Z < 1.20) \tag{31}$$

$$= 1 - 0.8449 = 0.1151. \tag{32}$$

Now, we know the proportion of females who scored worse than a man (11.51%), what proportion scored better than 125? We need to draw on Bayesian probabilities where the probability of event A occurring given that event B has **already occurred** is denoted $\mathbb{P}(A \mid B)$. This is equal to the probability of both events A and B occurring, restricted to the probability that event B occurs: $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$

$$\implies = \mathbb{P}(F > 125 \mid F < M) \tag{33}$$

$$= \frac{\mathbb{P}(F > 125 \cap F < M)}{\mathbb{P}(F < M)} \tag{34}$$

$$= \frac{\mathbb{P}(125 < F < M)}{\mathbb{P}(F < M)} \tag{35}$$

We want to know what portion of the normal distribution lies below M and above 125 (between the two), so we can just subtract the part of the graph below 125 from the part of the graph below M. See Figure 4 to visualise this better.

$$\implies = \frac{\mathbb{P}(F < M) - \mathbb{P}(F < 125)}{\mathbb{P}(F < M)} \tag{36}$$

$$= \frac{\mathbb{P}(W < 0) - \mathbb{P}\left(Z < \frac{125-122}{9}\right)}{\mathbb{P}(W < 0)} \tag{37}$$

$$= \frac{0.1151 - \mathbb{P}(Z < 0.33)}{0.1151} = \frac{0.1151 - 0.6293}{0.1151} = 0.72. \tag{38}$$

This tells us that 72% of the women who scored worse than a man, at least scored 125.

If the random variables are not independent, what do we do? If they are not independent, then $r \neq 0$ and also $\operatorname{Cov} X, Y \neq 0$. Suppose $W = aX + bY$ and $r \neq 0$, then

$$\sigma_W^2 = \operatorname{Var}(aX + bY) \tag{39}$$

$$= \operatorname{Var} aX + \operatorname{Var} bY + 2 \times \operatorname{Cov} aX, bY \tag{40}$$

$$= a^2 \operatorname{Var} X + b^2 \operatorname{Var} Y + 2ab \times \operatorname{Cov} X, Y \tag{41}$$

Recall the formula for $r(X,Y) = \operatorname{Cov} X, Y / (\sigma_X \times \sigma_Y)$, so $\operatorname{Cov} X, Y = r(X,Y) \times \sigma_X \times \sigma_Y$. We can rewrite this again as $\operatorname{Cov} X, Y = r(X,Y) \times \sqrt{\operatorname{Var} X \times \operatorname{Var} Y}$.

$$= a^2 \operatorname{Var} X + b^2 \operatorname{Var} Y + 2ab \times r(X,Y) \times \sqrt{\operatorname{Var} X \times \operatorname{Var} Y}. \tag{42}$$

In the case of our example, we have that X and Y are independent, however suppose they aren't and we have that $r = -0.7$.

$$\implies \sigma_W^2 = a^2 \operatorname{Var} F + b^2 \operatorname{Var} M + 2ab \times r(F,M) \times \sqrt{\operatorname{Var} F \times \operatorname{Var} M} \tag{43}$$

Recall that $W = F - M$ so $a = 1 = -b$.

$$= \operatorname{Var} F + \operatorname{Var} M + 2 \times (-0.7) \times \sqrt{\operatorname{Var} F \times \operatorname{Var} M} \tag{44}$$

$$= 9 + 16 - 1.4 \times \sqrt{9 \times 16} \tag{45}$$

$$= 25 - 16.8 = 8.2. \tag{46}$$

$$\implies \sigma_W = \sqrt{8.2} = 2.864. \tag{47}$$

2.4 Standard error

The standard error is short for "the standard error of the estimate for _____", where the underlined part can be "the mean", "the variance", etc. Quite bluntly, this is the error which is to be expected when you use a sample statistic to estimate the population value. In the case of the mean, we have the sampling distribution of the mean $\bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$ (see (2)), so the standard error of the estimate $\bar{y}$ for μ is

$$se = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}. \tag{48}$$

For the scope of this course, you do not need to know about any other standard errors, just that for estimating the mean using the sampling distribution.

3 Correlation

If two variables X and Y are correlated, we say that they are statistically associated, and in this course we refer to a linear relationship.

3.1 Kendall's τ

You start with a set of joint random variables (X, Y) and your collected data set $(x1_,y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Next, you classify any pair of observations, (x_i, y_i) and (x_j, y_j), as either

- **concordant** if the ranks for both elements agree, i.e. if $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$;
- **discordant** if the ranks do **not** agree, i.e. if $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$.

N.B.: if $x_i = x_j$ or $y_i = y_j$, then the pair is neither. We define Kendall's τ as:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{n(n-1)/2} \tag{49}$$

The denominator, $n(n-1)/2$, is the total number of pairs so τ is bounded! In fact, we have that $-1 \leq \tau \leq 1$; if $\tau = -1$, then we have **perfect disagreement** (if one ranking is the reverse of the other), and conversely if $\tau = 1$ then we have **perfect agreement** (if the rankings are the same). If we have that X and Y are **independent**, then we **expect** $\tau = 0$!! To see this, note the following equivalence

$$= \frac{2}{n(n-1)} \sum_{i<j} \underbrace{\operatorname{sgn}(x_i - x_j) \times \operatorname{sgn}(y_i - y_j)}_{\substack{= 1 \text{ if the pairs are concordant} \\ = -1 \text{ if the pairs are discordant}}} \tag{50}$$

So, $\tau = 0$ if and only if the number of concordant pairs equals the number of discordant pairs, i.e. X and Y are independent.
Assumptions:
τ measures the ordinal association between two measured quantities for the purposes of testing a (non-parametric) hypothesis of independence.

3.2 Spearman's ρ

Very similar to Kendall's τ (in it's use and assumptions), however it is suitable for both discrete (integers) and continuous (every decimal inbetween) **ordinal** variables; general correlation coefficient.
The first step is to order your data by ranks, for example:

$$(X, Y) = \begin{pmatrix} (7,4) \\ (5,7) \\ (8,9) \\ (9,8) \end{pmatrix} \longmapsto (\mathrm{rg}_X, \mathrm{rg}_Y) = \begin{pmatrix} (2,1) \\ (1,2) \\ (3,4) \\ (4,3) \end{pmatrix} \tag{51}$$

You may choose to order your pairs for X or for Y, but here I have chosen X; also, $n = 4$. The next step is to calculate Pearson's ρ (correlation coefficient):

$$\rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\operatorname{Cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{52}$$

An easier relation to remember:

$$= \frac{\sum_{i=1}^{n} z_{\mathrm{rg}_{X_i}} z_{\mathrm{rg}_{Y_i}}}{n-1}, \tag{53}$$

where $z_{\mathrm{rg}_{X_i}}$ is the z-score for the i-th ranked score of X; similarly, $z_{\mathrm{rg}_{Y_i}}$ of Y. If all rankings are **distinct integers**, then we can compute:

$$= 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{54}$$

where $d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$ is difference between the two ranks of each association. In our example,

$$\implies d = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \tag{55}$$

$$\implies d^2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{56}$$

$$\implies r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = 1 - \frac{6 \times 4}{4(4^2-1)} \tag{57}$$

$$= \frac{15-6}{15} \tag{58}$$

$$= \frac{9}{15} \tag{59}$$

$$= 0.6. \tag{60}$$

$$\implies t = \rho\sqrt{\frac{n-1}{1-r^2}} \sim t(n-2) \tag{61}$$

$$= 0.6\sqrt{\frac{3}{1-0.36}} \tag{62}$$

$$= 1.299 \sim t(3) \tag{63}$$

$$\tag{64}$$

3.3 Pearon's r

Pearson's r (or ρ, for a population) is a measure of the association (linear correlation) between two quantitative variables X and Y in a sample of size n; it is a ratio of the covariance and the standard deviations (of the sample):

$$r_{XY} = \frac{\text{Cov}\,(X,Y)}{s_X s_Y} \tag{65}$$

In order to simplify this, we substitute in the equations for covariance and standard deviation.

$$= \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} \tag{66}$$

We can omit the denominators of $n-1$ as they multiply and divide each other out to a constant of 1.

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{67}$$

Next, we multiple the brackets in the numerator, and square the brackets in the denominator.

$$= \frac{\sum_{i=1}^{n}(x_iy_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})}{\sqrt{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}\sqrt{\sum_{i=1}^{n}\left(y_i^2 - 2y_i\bar{y} + \bar{y}^2\right)}} \tag{68}$$

Take the summation and apply it to each term, and recall that $\sum \bar{x} = n\bar{x}$ as $\bar{x}$ is an assumed constant given the data.

$$= \frac{\sum_{i=1}^{n}x_iy_i - \bar{x}\sum_{i=1}^{n}y_i - \bar{y}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - 2\bar{y}\sum_{i=1}^{n}y_i + n\bar{y}^2}} \tag{69}$$

Now, we use that $n\bar{x} = \sum x_i$, and similarly for y:

$$= \frac{\sum_{i=1}^{n}x_iy_i - \bar{x}n\bar{y} - \bar{y}n\bar{x} + n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - 2n\bar{y}^2 + n\bar{y}^2}} \tag{70}$$

Again, we simplify this to our final expression.

$$= \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} \tag{71}$$

An equivalent, and easier to remember, expression for r is the **mean of the products of the standardised variables**:

$$r_{XY} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{s_X s_Y} \tag{72}$$

$$= \frac{\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{s_X s_Y}}{n-1} \tag{73}$$

$$= \frac{\sum_{i=1}^{n}z_{x_i}z_{y_i}}{n-1}. \tag{74}$$

If r is near to 1 (postively) or -1 (negatively), then we say that X and Y are correlated, and **not** independent; they are independent is r is near zero. If we have that $z_{x_i}z_{y_i}$ is positive, then x_i and y_i are both larger than their respective means or both smaller. Moreover, if $z_{x_i}z_{y_i}$ is negative, then $x_i < \bar{x}$ and $\bar{y} < y_i$, or $\bar{x} < x_i$ and $y_i < \bar{y}$. So, an r near to one tells us that (x_i, y_i) tend to be simultaneously greater than, or simultaneously less than, their respective means. The opposite is true for $r = -1$, as they are always on the opposite side of their respective means with regards to each other.

4 Short notes

5 number summary: min (Q_0 zero quartile, 0%), Q_1 (first quartile, 25%), Q_2 (median, second quartile, 50%), Q_3 (third quartile, 75%), max (Q_4 fourth quartile, 100%).
$IQR = Q_3 - Q_1$ interquartile range. For normally distributed R.V.'s, the distance from the mean to either quartile is about two-thirds of a standard deviation, so IQR equals approximately $(4/3)s$.
Rule for outliers: less than $Q_1 - 1.5 \times IQR$ or more than $Q_3 + 1.5 \times IQR$. For normally distributed R.V.'s, this makes a lot of sense as $1.5 \times IQR \approx (3/2) \times (4/3)s = 2s$, i.e. 2 standard deviations. So, outliers are those which lie more than 2 standard deviations away from the mean, i.e. upper and lower 2.5%.
Mean $\sum x/n$ is affected by outliers, whereas mode and median seldom are. Mean can also be called the expected value, i.e. the sum of the values an R.V. can take on multiplied by the probability:

$$\mathbb{E}(X) = \sum_x x \times p_X \tag{75}$$

$$= x_1 \times \mathbb{P}(X = x_1) + x_2 \times \mathbb{P}(X = x_2) + \cdots + x_n \times \mathbb{P}(X = x_n) \tag{76}$$

The above is for discrete variables; for continuous variables we use integrals but you don't need to know about that. Something else useful to know:

$$\mathrm{Var}(X) = \mathbb{E}\left(\left[X - \mathbb{E}(X)\right]^2\right) \tag{77}$$

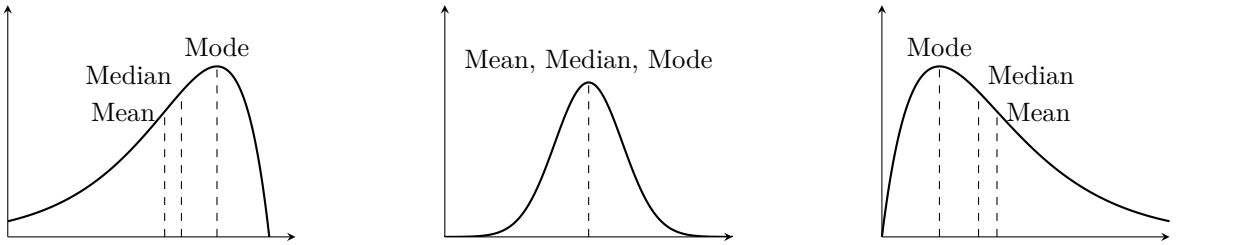
$$= \mathbb{E}(X^2) - \left[\mathbb{E}(X)\right]^2 \tag{78}$$

$$= \left[\sum_x x^2 \times p_X\right] - \left[\sum_x x \times p_X\right]^2. \tag{79}$$

This is the easiest formulation to remember. For question 22 of the exam, it was written in this way:

$$= \sum_x \left(x - \mathbb{E}(X)\right)^2 \times p_X. \tag{80}$$

Mode is highest frequency. Median is the middle value when ranked.
Right-skew: mode < Q_2 < mean (mean is on the right of the median).
Left-skew: mean < Q_2 < mode (mean is on the left of the median).



(a) Left-skewed distribution. (b) Normal distribution. (c) Right-skewed distribution.

Figure 6

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}{n}}. \tag{81}$$

Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}{n-1}}. \tag{82}$$

Back-to-back stem-and-leaf plots are used to compare frequency distributions of two data sets, e.g. exam grades (out of 30) for exams taken on different days.

Monday:	10	15	20	21	25	27	28	29	30
Friday:	3	8	14	17	18	26	27	27	28

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

Table 2: 5 number summary for Monday $(10, 17.5, 25, 28.5, 30)$ with $IQR = 11$ and for Friday $(3, 11, 18, 27, 28)$ with $IQR = 16$. Mean for Monday is 22.78 and for Friday is 18.67, meaning that Monday's distribution is left-skewed and Friday's is (slightly/hardly) right-skewed - Friday is distributed roughly normal.

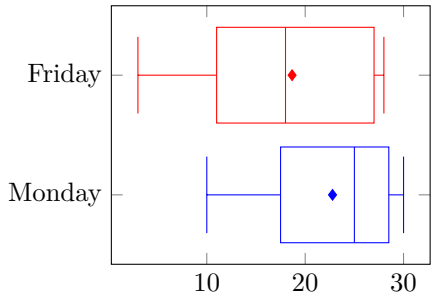


Figure 7: Box plot of exam data on different days, where the diamond displays the mean.

Every random variable which is normally distributed is symmetrical about it's mean which basically just refers to the mirror image effect of probabilities seen below. If the distribution of the variable has some type of skew, then it is not normal and does not exhibit the same symmetry.

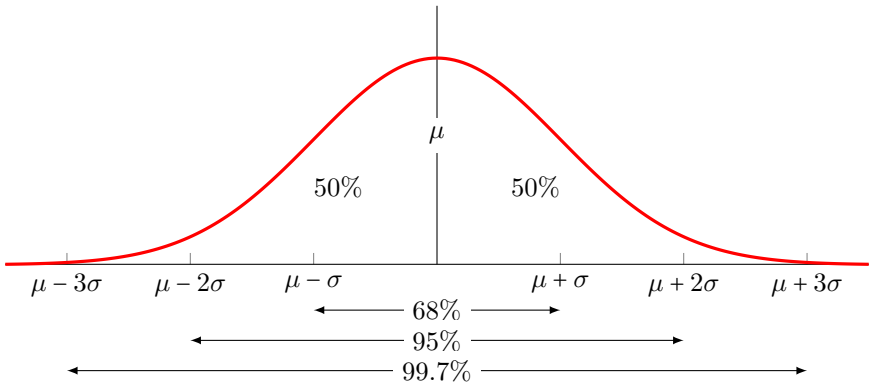


Figure 8: A normally distributed random variable X with mean μ and variance σ^2. 50% of the data lies either side of μ, and we can approximate the middle proportion using the 68-95-99.7 rule due to **symmetry about the mean**.

You can use this to check your answers! If you have a z-value which is approximately 1, or 2, or -3 (basically an integer $\pm$) you can use the 68-95-99.7 rule to estimate the p-value, e.g. $\mathbb{P}(Z < -2)$ can be estimated by remembering that in between -2 and 2 is about 95%, so what is left is about 5% - 2.5% each side, i.e. $\mathbb{P}(Z < -2) \approx 0.025$. Checking table A we have that $\mathbb{P}(Z < -2) = 0.0228$ which is near to our estimate. You can use this in all manner of questions that involve normally distributed random variables, e.g. comparing independent variables: the homework with $F \sim \mathcal{N}(\mu_F, \sigma_F^2)$ and $M \sim \mathcal{N}(\mu_M, \sigma_M^2)$ so $F - M \sim \mathcal{N}(\mu_F - \mu_M, \sigma_F^2 + \sigma_M^2)$. If you want to know if females are better than males then you're asking if a random female is better than a random male with high probability, i.e. find $\mathbb{P}(F > M) = \mathbb{P}(F - M > 0) = \mathbb{P}(Z > z)$ where z is the value found by solving:

$$z = \frac{0 - (\mu_F - \mu_M)}{\sqrt{\sigma_F^2 + \sigma_M^2}} = \frac{\mu_M - \mu_F}{\sqrt{\sigma_F^2 + \sigma_M^2}}. \tag{83}$$

Say this is a negative integer, e.g. $z = -1$, then $\mathbb{P}(Z > -1) = \mathbb{P}(Z < 1)$ (notice that flipping the inequality changes the negative to a positive - due to symmetry about the mean). You can estimate this by adding half of 68% to 50%, so it's roughly 84%; from table A, $\mathbb{P}(Z > -1) = 0.8413 = \mathbb{P}(Z < 1)$. You can also do this estimation using the critical z-values you know:

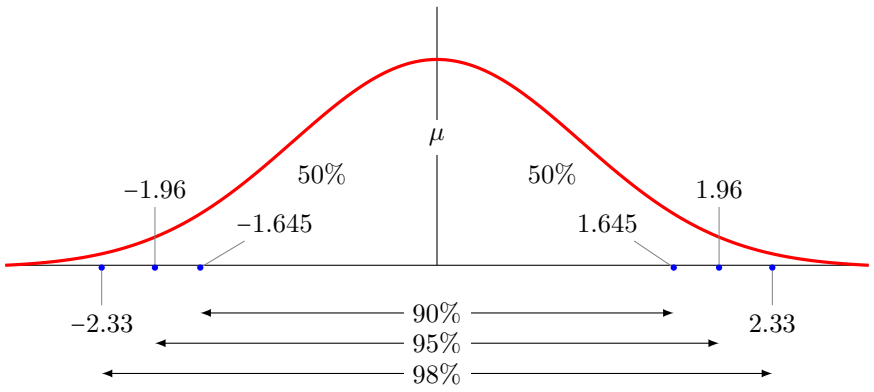


Figure 9: The standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. 50% of the data lies either side of μ, and we can calculate the middle proportions precisely using the critical values due to **symmetry about the mean**.

For **every** random variable X (with some distribution), the following always applies:

$$\mathbb{P}(X < x) = 1 - \mathbb{P}(X > x) \tag{84}$$

If the random variable has a continuous distribution, then:

$$\mathbb{P}(X \leq x) = \mathbb{P}(X < x) \implies \mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = 0. \tag{85}$$

The graph you are usually given for the standard normal distribution is $\mathbb{P}(Z < z)$ where z is positive. You can use re-arrangements of the following to determine the p-value regardless of whether you are looking for left- or right-tailed, or if you have a positive or negative z value:

$$\mathbb{P}(Z < z) = \begin{cases} 1 - \mathbb{P}(Z > z) \\ \mathbb{P}(Z > -z) \\ 1 - \mathbb{P}(Z < -z) \end{cases} \tag{86}$$

For discrete R.V.'s, you usually use $\leq$ and $\geq$ signs, but be aware of the following:

$$\mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x) = 1 - \mathbb{P}(X \geq x + 1). \tag{87}$$
$$\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X < x) = 1 - \mathbb{P}(X \leq x - 1). \tag{88}$$

For linear transformations, e.g. $W = aX + bY + c$ where $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ are normally distributed R.V.'s and a, b, c are some number, there are simple rules to remember:

$$\mu_W = \mathbb{E}(W) = \mathbb{E}(aX + bY + c) \tag{89}$$
$$= a\,\mathbb{E}(X) + b\,\mathbb{E}(Y) + c \tag{90}$$
$$= a\mu_X + b\mu_Y + c. \tag{91}$$
$$\sigma_W^2 = \operatorname{Var}(W) = \operatorname{Var}(aX + bY + c) \tag{92}$$

c is a constant, so it doesn't vary and we can take it out of this variance equation:

$$= \operatorname{Var}(aX) + \operatorname{Var}(bY) + 2\operatorname{Cov}(aX, bY) \tag{93}$$

Next, remember that $r(X,Y) = \operatorname{Cov}(X,Y)/\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$:

$$= a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab\left[r(X,Y)\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}\right] \tag{94}$$
$$= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\left(r(X,Y)\sigma_X\sigma_Y\right). \tag{95}$$

If $X \perp\!\!\!\perp Y$ (X and Y are independent) then $r(X,Y) = 0$, so

$$\sigma_W = \begin{cases} \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}, & \text{if } X \text{ and } Y \text{ are independent } (r = 0); \\ \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\left(r(X,Y)\sigma_X\sigma_Y\right)}, & \text{if } X \text{ and } Y \text{ are } \textit{not} \text{ independent } (r \neq 0). \end{cases} \tag{96}$$

Here is why you add twice the covariance (of the variable times the constant a or b) when computing the variance of a linear transformation:

$$\operatorname{Var}(aX + bY + c) = \mathbb{E}\left[\left((aX + bY + c) - \mathbb{E}(aX + bY + c)\right)^2\right] \tag{97}$$

The c disappears as it is a constant (doesn't vary) and we can simplify this to:

$$= \mathbb{E}\left((aX+bY)^2\right) - \left[\mathbb{E}\left(aX+bY\right)\right]^2 \tag{98}$$

We use now the identity: $(a+b)^2 = a^2+b^2+2ab$

$$= \mathbb{E}\left((aX)^2+(bY)^2+2abXY\right) - \left[a\,\mathbb{E}(X)+b\,\mathbb{E}(Y)\right]^2 \tag{99}$$

$$= \left[a^2\,\mathbb{E}\left(X^2\right)+b^2\,\mathbb{E}\left(Y^2\right)+2ab\,\mathbb{E}\left(XY\right)\right] - \left[\left[a\,\mathbb{E}(X)\right]^2+\left[b\,\mathbb{E}(Y)\right]^2+2ab\,\mathbb{E}(X)\,\mathbb{E}(Y)\right] \tag{100}$$

$$= \left[a^2\,\mathbb{E}\left(X^2\right)-a^2\left[\mathbb{E}(X)\right]^2\right]+\left[b^2\,\mathbb{E}\left(Y^2\right)-b^2\left[\mathbb{E}(Y)\right]^2\right]+\left[2ab\,\mathbb{E}\left(XY\right)-2ab\,\mathbb{E}(X)\,\mathbb{E}(Y)\right] \tag{101}$$

$$= a^2\underbrace{\left[\mathbb{E}\left(X^2\right)-\left[\mathbb{E}(X)\right]^2\right]}_{\mathrm{Var}(X)}+b^2\underbrace{\left[\mathbb{E}\left(Y^2\right)-\left[\mathbb{E}(Y)\right]^2\right]}_{\mathrm{Var}(Y)}+2ab\underbrace{\left[\mathbb{E}\left(XY\right)-\mathbb{E}(X)\,\mathbb{E}(Y)\right]}_{\mathrm{Cov}(X,Y)} \tag{102}$$

$$= a^2\,\mathrm{Var}\left(X\right)+b^2\,\mathrm{Var}\left(Y\right)+2ab\,\mathrm{Cov}\left(X,Y\right) \tag{103}$$

If the variables are independent, then their covariance is zero and subsequently so is r. Question 9 of the exam had the following linear transformation: $W = \mathrm{RUG}_1 - \mathrm{RUG}_2$ so $a = 1$ and $b = -1$ (the negative is important), and

$$\mathrm{Var}\left(\mathrm{RUG}_1-\mathrm{RUG}_2\right) = (1)^2\,\mathrm{Var}\left(\mathrm{RUG}_1\right)+(-1)^2\,\mathrm{Var}\left(\mathrm{RUG}_2\right)+2\times(1)\times(-1)\,\mathrm{Cov}\left(\mathrm{RUG}_1,\mathrm{RUG}_2\right) \tag{104}$$

$$= \mathrm{Var}\left(\mathrm{RUG}_1\right)+\mathrm{Var}\left(\mathrm{RUG}_2\right)-2\times\underbrace{r\left(\mathrm{RUG}_1,\mathrm{RUG}_2\right)\sigma_{\mathrm{RUG}_1}\sigma_{\mathrm{RUG}_2}}_{\mathrm{Cov}\left(\mathrm{RUG}_1,\mathrm{RUG}_2\right)}. \tag{105}$$

N.B.: the formal notation for normal R.V.'s is $\mathcal{N}\left(\text{mean, variance}\right) = \mathcal{N}\left(\mu,\sigma^2\right)$, however it is really clear that your teacher swaps between putting standard deviation and variance. *BE CAREFUL IN THE EXAM* - the correct answer for question 10 in the exam should have been $\mathcal{N}\left(220,14.4\right)$ and not $\mathcal{N}\left(220,\sqrt{14.4}\right) = \mathcal{N}\left(220,3.8\right)$. So please check your answers to ensure that both are not answer possibilities!!

$$r(x,y) = \frac{\mathrm{Cov}\left(x,y\right)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}} = \frac{\mathrm{Cov}\left(x,y\right)}{s_x s_y} = \frac{\sum\left(x_i-\bar{x}\right)\times\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2}\times\sqrt{\sum\left(y_i-\bar{y}\right)^2}} = \frac{\sum\left(z_{x_i}\times z_{y_i}\right)}{n-1}. \tag{106}$$

The easiest way to remember Pearson's correlation coefficient r is as the sum of multiplied z-scores divided by the degrees of freedom $n-1$, which can be calculated from the sample as follows (or using your calculator):

$$\begin{array}{ccccc} \left(x_i-\bar{x}\right) & \mapsto & \left(x_i-\bar{x}\right)^2 & \mapsto & s_x = \dfrac{\sum\left(x_i-\bar{x}\right)^2}{n-1} \\ & \searrow & & \swarrow & \\ & & z_{x_i} = \dfrac{x_i-\bar{x}}{s_x} & & \end{array}$$

Do this for both x and y, multiply the z-scores pairwise, then add all of those multiplied z-scores together and divide by $n-1$.

[**Instructions for CASIO fx-82MS**] Your calculator is an important tool, and was not meant to only perform simple operations, and the following is a guide to conducting linear regression.

1. Turn it on, and press SHIFT and then the MODE to select CLR, then 3 and then = twice to clear all your previous inputs and modes.

2. Press MODE and then

   Press 2 for univariate data: For example, Question 3 of Homework 1 asks you find the find the mean, median and standard deviation for the following data set (18, 19, 20, 21, 24, 45) of student ages. To input this, type 18 and press M+ afterwards. You should now see $n = 1$ on the screen, which tells you that you have saved the first data point. Now press 19 and then M+, then 20 and then M+, and so on. You are finished when you have entered all 6 data points and the screen says $n = 6$.
   If you have entered a data point incorrectly you can use the up and down keys to cycle through the data points until you see the incorrect one, and then just type the correct number and press = to save the correction. If you wish delete a data point, use the up and down keys to cycle through the data points until you see the incorrect one, and then press SHIFT and then M+ to

select CL.
If you wish to know the sum of squared values $\sum x^2$, the sum of values $\sum x$ or the sample size n, press SHIFT and then 1 to select S-SUM, then press the appropriate number for your selection and then press = to display the value.
If you wish to know the sample mean $\bar{x}$, the population standard deviation σ_x or the sample standard deviation s_x, press SHIFT and then 2 to select S-VAR, then press the appropriate number for your selection and then press = to display the value.
For the aforementioned data set, you would compute the sum of squared values $\sum x^2 = 4127$, the sum of values $\sum x = 147$, the sample size $n = 6$, the sample mean $\bar{x} = 24.5$, the population standard deviation $\sigma_x = 9.359$ and the sample standard deviation $s_x = 10.252$. The median is half of the sum of 3rd and 4th ranked data values, i.e. $(20+21)/2 = 20.5$ is the median.

Press 3 and then 1 for bivariate data: For example, Question 1 of Homework 3 asks you find the find the correlation coefficient for the following data set $\{(166,182),(164,178),(166,180),(165,178),(170,181),(165,181)\}$ of the heights of women and their husbands. To input this, type 166,182 and press M+ afterwards. You should now see $n = 1$ on the screen, which tells you that you have saved the first data point. Now press 164,178 and then M+, then 166,165 and then M+, and so on. You are finished when you have entered all 6 data points and the screen says $n = 6$.
If you have entered a data point incorrectly you can use the up and down keys to cycle through the data points until you see the incorrect one, and then just type the correct data value and press = to save the correction. If you wish delete a data point, use the up and down keys to cycle through the data points until you see the incorrect one, and then press SHIFT and then M+ to select CL.
If you wish to know the sum of squared x values $\sum x^2$, the sum of x values $\sum x$ or the sample size n, press SHIFT and then 1 to select S-SUM and then press the appropriate number for your selection. You can press the right and left arrow keys to cycle through the different menu options: sum of squared y values $\sum y^2$, the sum of y values $\sum y$, or the sum of point-wise multiplied x and y values $\sum x_i y_i$.
If you wish to know the sample mean of x, $\bar{x}$, the population standard deviation of x, σ_x, or the sample standard deviation of x, s_x, press SHIFT and then 2 to select S-VAR and then press the appropriate number for your selection. You can press the right and left arrow keys to cycle through the different menu options: sample mean of y, $\bar{y}$, the population standard deviation of y, σ_y, the sample standard deviation of y, s_y, the linear regression intercept coefficient A, the linear regression slope coefficient B, or the linear correlation coefficient r. You may notice that on the last cycle of the menu there are the options $\hat{x}$ and $\hat{y}$: press AC, then type the y_i value for which you wish to predict $\hat{x}_i$, then find the $\hat{x}$ option in the menu and press =. Do the same if you wish to find response $\hat{y}_i$ based on input x_i.
For the aforementioned data set, you would compute the linear regression intercept coefficient A=112.091, the linear regression slope coefficient B=0.4091, and the linear correlation coefficient $r = 0.513$. An example of $\hat{x}$ and $\hat{y}$: $\hat{x}_i(180) = 166$ and $\hat{y}_i(168) = 180.8$.

If $r = 1$ then positive correlation; $r = 0$ then independent; $r = -1$ then negative correlation. r is invariant under linear transformations but may be affected by the sign, i.e. if $W = a + bY$ then

$$r(X,W) = r(X, a+bY) = \frac{\mathrm{Cov}(X, a+bY)}{s_X s_{a+bY}} = \frac{b\,\mathrm{Cov}(X,Y)}{\sqrt{b^2} s_X s_Y} = \mathrm{sign}(b)\frac{\mathrm{Cov}(X,Y)}{s_X s_Y} \tag{107}$$

$$= \mathrm{sign}(b) \times r(X,Y). \tag{108}$$

If $r > 0$ (positive) then $(x_i - \bar{x}, y_i - \bar{y})$ have more concordant than discordant pairs, and conversely if $r < 0$ (negative) then $(x_i - \bar{x}, y_i - \bar{y})$ have more discordant than concordant pairs. See Kendall's τ below for the definition of concordant and discordant. If $r \neq 0$ then you can impose a **regression of dependent y on independent** x (order is important here: regression of y on x implies we model $\hat{y}$): $\hat{y}_i = b_0 + b_1 x_i$, where the coefficients are

$$b_0 = \text{intercept} = \bar{y} - b_1 \bar{x}; \tag{109}$$

$$b_1 = \text{slope} = r(x,y)\frac{s_y}{s_x} = \frac{\mathrm{Cov}(x,y)}{s_x^2}. \tag{110}$$

If x and y are *already* standardised, then $b_0 = 0$ and $b_1 = r(x,y) = [\sum(xy)]/(n-1)$. If $r < 0$ then $b_1 < 0$ (negative correlation = negative slope), and conversely if $r > 0$ then $b_1 > 0$ (positive correlation = positive slope); the slope

of the regression line depends *only* on the sign of r. If you are given R^2 (percentage of variance in y explained by x) and b_1, then $r = \text{sign}(b_1)\sqrt{R^2}$. Furthermore, $r(\hat{y}, y) = |r(x, y)|$ - this was an exam question once upon a time. It is important to remember that the regression line $\hat{y}_i = b_0 + b_1 x_i$ **must pass through the means**, i.e. pass through $(\bar{x}, \bar{y})$:

$$\hat{y}_i(\bar{x}) = b_0 + b_1(\bar{x}) = [\bar{y} - b_1\bar{x}] + b_1\bar{x} = \bar{y}. \tag{111}$$

Simpson's paradox: correlation ≠ causation, e.g. children's learning ability and taking vitamins. In a simple example, if a teacher were to ask the children in their class who was given vitamins by their parents in the morning and then contrast with the child's grades, they may find that taking vitamins and good grades are positively correlated. However, this does not take into account that parents who give vitamins might have more money than those who do not, and can therefore better provide for their children resulting in more focus at school (less turbulent home life).
Lurking variable: variables which are unknown or difficult to control and thus are not accounted for. With reference to the previous example, the parents' having a high income and administering vitamins both influence the grade of a child in a positive way.
Confounding variables: variables which attribute to the thing you are trying to measure in an unexpected way resulting in incorrect analysis of data - what you thought were two independent variables might be related through the response variable, e.g. birth weight and age of the mother. It is well documented the increased risk of neonatal health issues when the mother is over or around 50, so the mother's age must be taken into account when conducting studies on pregnant women in this manner.
Re-read chapter 10.1 on causality.
General theory - re-read chapter 2 about variables (discrete, quantative, nominal, etc.), random sampling, parameter v.s. statistic, etc. *LEARN THESE DEFINITIIONS BY HEART.*
Probability (Bayesian): $\mathbb{P}(A \text{ or } B)$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \underbrace{\mathbb{P}(A \cap B)}_{\mathbb{P}(A \text{ and } B)}. \tag{112}$$

Conditional probability: $\mathbb{P}(A \text{ given } B)$ - B happened, so what's the probability of A occurring?

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{113}$$

$\mathbb{P}(A \cap B)$ is the joint probability distribution, which we can use to find the marginal probability distributions $\mathbb{P}(A)$ and $\mathbb{P}(B)$. Marginal probability:

$$\mathbb{P}(A) = \sum_{x \in B} \mathbb{P}(A \cap B) \tag{114}$$

$$= \mathbb{P}(A \cap \{B = x_1\}) + \mathbb{P}(A \cap \{B = x_2\}) + \cdots + \mathbb{P}(A \cap \{B = x_n\}). \tag{115}$$

Independence: $A \perp\!\!\!\perp B$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \times \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \tag{116}$$

Mutual exclusivity: $\mathbb{P}(A \text{ and } B) = 0$ - A cannot happen at the same time as B

$$\mathbb{P}(A \cap B) = 0 \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \tag{117}$$

"At least one" = 1 − "none":

$$\mathbb{P}(\text{event } A \text{ occurs at least once}) = 1 - \mathbb{P}(\text{event } A \text{ never occurs}) \tag{118}$$

	Event A	Event B	Marginal
Event C	$\mathbb{P}(A \cap C)$	$\mathbb{P}(B \cap C)$	$\mathbb{P}(C)$
Event D	$\mathbb{P}(A \cap D)$	$\mathbb{P}(B \cap D)$	$\mathbb{P}(D)$
Marginal	$\mathbb{P}(A)$	$\mathbb{P}(B)$	$\mathbb{P}(\Omega) = 1$

Table 3

X	0	1	2	3	...	n
p_X	$\mathbb{P}(X = 0)$	$\mathbb{P}(X = 1)$	$\mathbb{P}(X = 2)$	$\mathbb{P}(X = 3)$	...	$\mathbb{P}(X = n)$

Table 4: Mean/expected value: $\mathbb{E}(X) = \mu_X = \sum x \times p_X$. Standard deviation/spread: $\sqrt{\text{Var}(X)} = \sigma_X = \sum x^2 \times p_X - \mu_X^2$.

A Bernoulli trial is an experiment with two outcomes, "success" and "failure". The probability of success is $0 \leq p \leq 1$ and the probability of failure is $q = 1 - p$. If you perform more than one Bernoulli trial, say n trials, then you can model this with Binomial distribution, which gets its name from the use of the Binomial coefficient $\binom{n}{k}$. Formally, $X \sim \mathcal{B}in(n, p)$ so then,

$$\mathbb{P}(X = k \mid n;\ p) = \binom{n}{k} p^k q^{n-k} \tag{119}$$

$$= \frac{n!}{k!(n-k)!} p^k q^{n-k} \tag{120}$$

$$= \binom{n}{n-k} p^k q^{n-k} \tag{121}$$

$$= \mathbb{P}(X = n - k \mid n;\ q) \tag{122}$$

For example, suppose the probability of success is $p = 0.6$ and you want to find the probability that you have exactly 6 successes in 10 trials; $X \sim \mathcal{B}in(10, 0.6)$ and find $\mathbb{P}(X = 6)$.

$$\mathbb{P}(X = 6 \mid 10;\ 0.6) = \binom{10}{6} 0.6^6 0.4^{10-6} \tag{123}$$

$$= \frac{10!}{6!(10-6)!} 0.6^6 0.4^4 \tag{124}$$

$$= \frac{10 \times 9 \times 8 \times 7}{2 \times 3 \times 4} 0.046656 \times 0.0256 \tag{125}$$

$$= 5 \times 3 \times 2 \times 7 \times 0.0011943936 \tag{126}$$

$$= 0.250822656 \tag{127}$$

Remember: this is the same probability of exactly 4 failures in 10 trial, when the probability of any 1 failure is 0.4.

$$= \mathbb{P}(X = 4 \mid 10;\ 0.4) \tag{128}$$

In the same way, we can work out the probabilities for all possible values of X:

k	0	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}(X = k)$	0.000105	0.00157	0.0106	0.0425	0.1115	0.201	0.251	0.215	0.121	0.0403	0.00605

Table 5

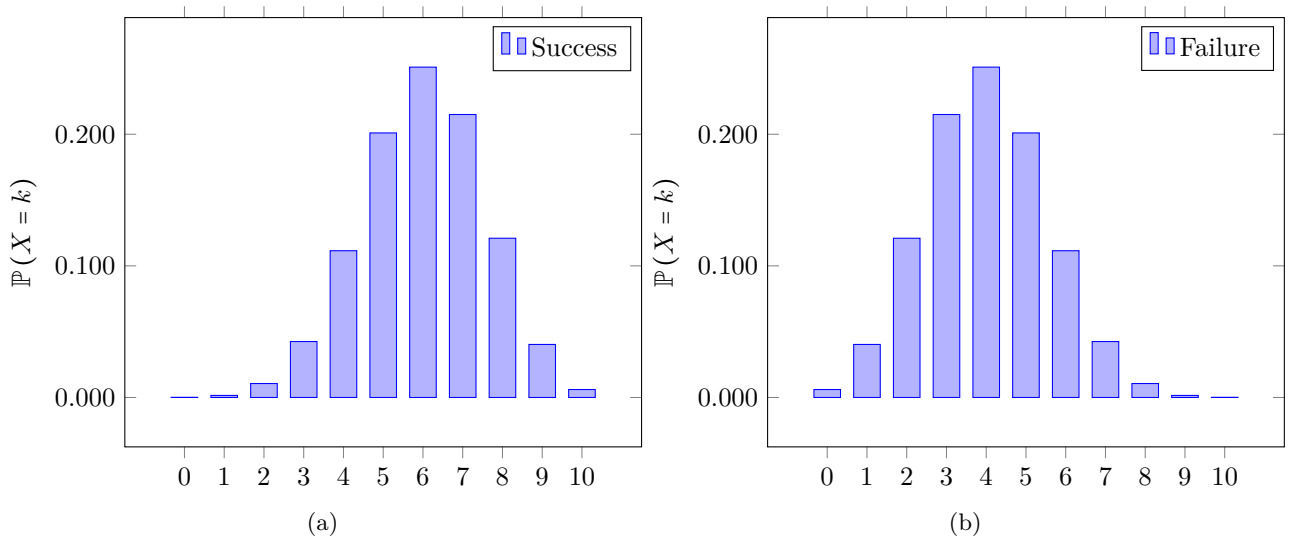


Figure 10: Probability of success fig. 10a and failure fig. 10b for $X \sim \mathcal{B}in(10, 0.6)$. Notice that the histograms are mirror images of each other? This is because the probability of k successes is equal to the probability of $10 - k$ failures, for $k = 0, 1, \ldots, 10$. This mirror imagery holds for all Binomial distributed random variables. For $\mathbb{P}(X \leq k)$, you need to imagine stacking the histogram bars on top each other from 0 up to k, and conversely for $\mathbb{P}(X \geq k)$ you imagine stacking the histogram bars on top each other from 10 down to k.

If $X \sim \mathcal{B}in(n, p)$ (binomial) and both np and $n(1-p)$ are greater than 5, we can approximate using the Normal distribution with mean np and variance $np(1-p)$. Ensure you remember the **continuity correction**!!

Binomial is discrete but normal is continuous, and a the probability that a continuous variable equals a number is always zero, e.g. $\mathbb{P}(X \leq 5) \approx \mathbb{P}(Y < 5.5)$ and $\mathbb{P}(X = 7) \approx \mathbb{P}(6.5 < Y < 7.5)$ where $Y \sim \mathcal{N}(np, np(1-p))$.
Sampling distribution of the mean of X: if X has mean μ and variance σ^2 (no assumptions about normality of X - can be skewed) then by conducting random sampling of size $n \geq 30$ (N samples, each of size n) ensures that the distribution of the mean $\bar{X}$ is approximately normal with the same mean μ and with variance σ^2/n, i.e. the variance decreases as n increases and the distribution centres around the population mean μ. This is due to Law of Large Numbers (LLN), which implies the Central Limit Theorem holds (CLT) - discusses central tendency, i.e. data tends towards the centre (the mean) as you increase the sample size.

$$\text{Cohen's } d\text{: } d = \frac{x - \mu}{\sigma}, \quad \text{where } X \sim \left(\mu, \sigma^2\right). \tag{129}$$

Cohen's d tells you how many standard deviations away from the mean your sample statistic x lies (the z-score if X is normally distributed).

$$\text{Kendall's } \tau\text{: } \tau = \frac{\#C - \#D}{n(n-1)/2}, \quad \text{where there is no distribution.} \tag{130}$$

Kendall's τ tells you how rank correlated the sample data are, where $\tau = 0$ indicates that there is no relationship and $\tau = 1$, that there is a perfect relationship. Used for measuring the correlation between variables that are without distribution, i.e. categorical data. **Concordant pairs:** as the rankings for X increase they also increase for Y, and similarly when the rankings for X decrease they also decrease for Y. **Discordant pairs:** as the rankings for X increase they decrease for Y, and similarly when the rankings for X decrease they increase for Y. **Steps to calculate Kendall's τ:**

1. Take the following data set:

Category 1	**Category 2**
6	5
10	8
2	6
1	3
9	10
3	2
5	10
7	2
8	7
5	2

Table 6

2. Create two new columns with their rankings within the categories: in category 1, we have the score of 5 twice which occupy the 4th and 5th ranks so they each get a rank score of $(4+5)/2 = 2.5$ - you divide by 2 as there are two of them. In category 2, we have a score of 2 three times which occupy the 1st, 2nd and 3rd ranks so they each get a rank score of $(1+2+3)/3 = 2$ - you divide by 3 as there are three of them. In category 2, you also have have the score of 10 twice which occupy the 9th and 10th ranks so they each get a rank score of $(9+10)/2 = 9.5$ - you divide by 2 as there are two of them.

3. Order either category 1 or 2 by increasing rank, and the associated pair from the other category (not sorted) will move with it. Here is the table sorted by category 1: Doing this allows you to see con- and discordance more readily. Looking at the first pair in Table 8, $(1, 4)$, and compare it to the other pairs in the following way:

   (a) The total number of comparisons you need to conduct is $n(n-1)/2$. Here $n = 10$, so the total is $10 \times 9/2 = 45$.

   (b) 1 is less than 4, and comparing it with the next pair, $(2, 6)$, 2 is less than 6. This is a concordant pair.

   (c) The next pair, $(3, 2)$, is discordant with $(1, 4)$ as 3 is more than 2.

   (d) Continue this for the total 45 pair matchings until you can count the total number of con- and discordant pairs. Deduct the number of discordant pairs from the number of concordant pairs, and divide by the total to calculate τ. This gives you the ratio of agreement, i.e. as one category goes up so does the other, and vice versa, so they must influence each other.

Category 1	Category 2
6	5
10	8
2	6
1	4
9	9.5
3	2
4.5	9.5
7	2
8	7
4.5	2

Table 7

Category 1	Category 2
1	4
2	6
3	2
4.5	2
4.5	9.5
6	5
7	2
8	7
9	9.5
10	8

Table 8

4. As you have already ordered category 1, it helps to add an additional column displaying the inequality between the paired data, such as in Table 9. All the pairs that have the < are concordant with each other, and discordant with the pairs that have the >. There are 4 <'s and 6 >'s, and $\binom{4}{2}+\binom{6}{2} = 4\times3/2+6\times5/2 = 21$.

   (a) The first pair is <, and below that are 6 >'s. So our running total for discordant pairs is 6.

   (b) The second pair is <, and below that are 6 >'s. So our running total for discordant pairs is $6+6 = 12$.

   (c) The third pair is >, and below that are 2 <'s. So our running total for discordant pairs is $12+2 = 14$.

   (d) The fourth pair is >, and below that are 2 <'s. So our running total for discordant pairs is $14+2 = 16$.

   (e) The fifth pair is <, and below that are 4 >'s. So our running total for discordant pairs is $16+4 = 20$.

   (f) The sixth, seventh and eighth pairs are >, and below them all is 1 <. So our running total for discordant pairs is $20+1+1+1 = 23$.

   (g) The ninth pair is <, and below that is 1 >. So our running total for discordant pairs is $23+1 = 24$.

   N.B.: count down the list and do not count anything twice.

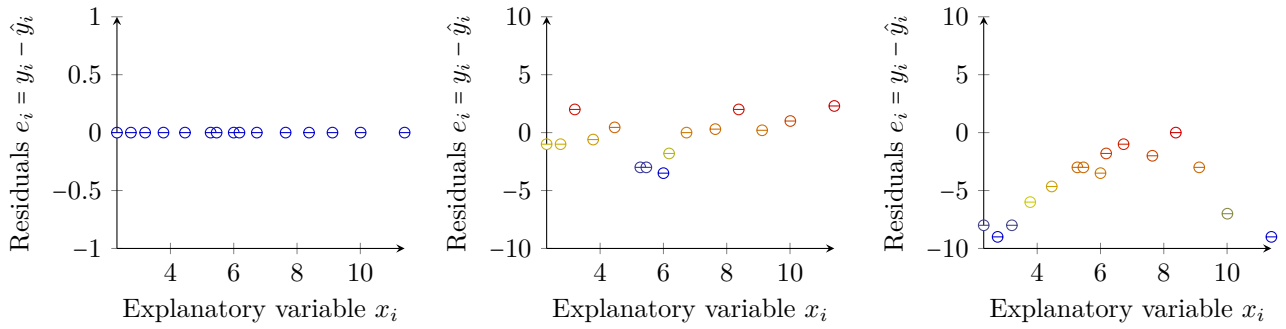
5. The total is 45, and 24 is accounted for by the number of discordant pairs, meaning that the remaining 21 is accounted for by the number of concordant pairs. Recall: there are 4 <'s and 6 >'s, and $\binom{4}{2}+\binom{6}{2} = 4\times3/2+6\times5/2 = 21$.

$$\tau = \frac{21-24}{45} = \frac{-3}{45} = -0.0667 \tag{131}$$

   It doesn't matter that τ is negative - if you change the arrangement of the categories, then it will be the same value only positive. It is clear that these two categories do not share any rank correlation, and thus do not show agreement.

Category 1		Category 2
1	$<$	4
2	$<$	6
3	$>$	2
4.5	$>$	2
4.5	$<$	9.5
6	$>$	5
7	$>$	2
8	$>$	7
9	$<$	9.5
10	$>$	8

Table 9



(a) (b) (c)

Figure 11

How to view explanatory-residual plots: if the residuals $e_i = y_i - \hat{y}_i$ are plotted against the explanatory variable x_i you can determine a lot about the appropriateness of a simple linear regression model to the data points (x_i, y_i). Figure 11a tells us that the predicted variables were always *exactly equal* to the observed response variables, i.e. $\hat{y}_i = y_i$ for every data point x_i and $i = 1, 2, \ldots, n$. This means for the linear correlation coefficient $r = 1$, or $r = -1$ if the relationship is negative, and we call this a perfect linear relationship.
Figure 11b displays an appropriate use of the simple linear regression model as the pattern is randomised - if there were an observable pattern, you would deduce that the residuals are not independent of the explanatory variable, which is one of your main assumptions for simple linear regression! If you have a U-shaped graph like in Figure 11c, then you can conclude that the linear model was not appropriate to model the data, and perhaps a quadratic or logarithmic model is more appropriate.
How to compare residual and scatter plots: if it is evident from the scatter plot of the response and explanatory variables (x, y) that there is no linear relationship, you can already assume that Figures 11a and b will **not** be the corresponding residual-explanatory variable plot. You can see from the scatter plot that they are not linearly correlated so this will be evident in the scatter plot by displaying a pattern, e.g. U-shaped.
For example, question 15 of the exam: it is evident that the pattern is not linear and **remember** that the regression line *must* pass through the point $(\bar{x}, \bar{y})$. This means that the regression line would go through the biggest cluster of data points whilst trying to reduce the distance between itself and the data points. So, kind of skateboard your data point through the biggest cluster of data points and then pivot on $(\bar{x}, \bar{y})$ as the central point to put it through as many data points as possible. Your approximate regression line that you have superimposed on the scatter plot is the $e_i = 0$ line on the regression plot, meaning that if you spin your scatter plot so that this is parallel with your eyes you can have a rough look at what the residual plot might look like.