Formulas

Stephanie Ranft S2459825

October 17, 2019

Statistics 2
PSBE2-07

Variance

Pooled variance

Consider a test between I independent samples. Whilst we cannot assume that they are all from the same population (and hence have the same variance), the Central Limit Theorem allows us to conclude that the pooled variance converges to the true variance. To see how this works, and in order to better understand when and where to use pooled variance:

$$s_p^2 = \frac{\sum_{i=1}^{I}(n_i-1)\,s_i^2}{\sum_{i=1}^{I}(n_i-1)}, \qquad \text{where } I \text{ is the total number of groups.} \tag{1}$$

$$= \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2 + \cdots + (n_I-1)\,s_I^2}{\underbrace{(n_1-1)+(n_2-1)+\cdots+(n_I-1)}_{n-I}} \tag{2}$$

If we were to assume that one of the samples was a lot bigger than the others, suppose sample k ($n_k >> n_i$ for every sample), then we can assume that the pooled variance will converge to s_k^2:

$$= \frac{\frac{n_1-1}{n_k-1}s_1^2 + \cdots + s_k^2 + \cdots + \frac{n_I-1}{n_k-1}s_I^2}{\frac{n_1-1}{n_k-1} + \cdots + \frac{n_k-1}{n_k-1} + \cdots + \frac{n_I-1}{n_k-1}} \tag{3}$$

$$\sim s_k^2. \tag{4}$$

The reason for this has something to do with the "power" of having a large n; that the sample is reliably similar to the population (lower standard error). This is an important point in sample collection, because if all samples except one have a really small size (<30) but one sample is substantially larger (>100), then the statistician can more readily believe the results of the largest sample (due to CLT - recall formula for SE). If we have that all of the samples are nearly the same size (choose $n_k \approx n_1 \approx \cdots \approx n_I$), then there is no dominating variance and we can assume the following:

$$s_p^2 \approx \frac{(n_k-1)\,s_1^2 + (n_k-1)\,s_2^2 + \cdots + (n_k-1)\,s_I^2}{(n_k-1)+(n_k-1)+\cdots+(n_k-1)} \tag{5}$$

$$= \frac{\cancel{(n_k-1)} \times \left(s_1^2 + s_2^2 + \cdots + s_I^2\right)}{I \times \cancel{(n_k-1)}} \tag{6}$$

$$= \frac{\sum_{i=1}^{I} s_i^2}{I} \tag{7}$$

$$= \bar{s}^2, \qquad \text{which is the mean of the variances.} \tag{8}$$

This indicates that the pooled variances is the weighted average of variances, where the highest weight is given to the largest sample size. This is to ensure that our pooled variance is the best estimator of the population variance. Another way to look at this is to look at the fact that $s_i^2 = \sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2/(n_i-1)$, where j indexes the persons in a group.

$$\implies s_p^2 = \frac{\sum_{j=1}^{n_1}\left(y_{1j}-\bar{y}_1\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j}-\bar{y}_2\right)^2 + \cdots + \sum_{j=1}^{n_I}\left(y_{Ij}-\bar{y}_I\right)^2}{n-I} \tag{9}$$

$$= \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}{n-I} \tag{10}$$

$$= \frac{\text{SSE}}{\text{df}_E} = MSE. \tag{11}$$

It is important to know when exactly to use pooled v.s. unpooled variance, but if we can **assume that the variances of the populations are equal**, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$, then we use pooled variance.
For $I = 2$ (comparing two means), we can use the t statistic to determine the results of our test:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\left(\underbrace{n_1 + n_2}_{n} - 2\right) \tag{12}$$

$$\implies t^2 = \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_p^2 \times \underbrace{\frac{n_1+n_2}{n_1 n_2}}_{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{13}$$

$$= \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2 / \left(n_1 + n_2\right)}{s_p^2 / n_1 n_2} \tag{14}$$

If $n_1 \approx n_2$

$$\implies t^2 \approx \frac{\frac{n}{2}\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_p^2} \sim F\left(1, n_1 + n_2 - 2\right). \tag{15}$$

For $I > 2$ (comparing multiple means), we use the F statistic:

$$F = \frac{\text{MSG}}{\text{MSE}} \tag{16}$$

$$= \frac{\text{SSG}/\text{df}_G}{\text{SSE}/\text{df}_E} \tag{17}$$

$$= \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(\bar{y}_i - \bar{y}\right)^2 / (I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 / \underbrace{(n-I)}_{\substack{(n-1)-(I-1) \\ =\text{df}_T - \text{df}_G}}} \tag{18}$$

$$= \frac{\sum_{i=1}^{I} n_i \times \left(\bar{y}_i - \bar{y}\right)^2 / (I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 / (n-I)} \sim F\left(\underbrace{I-1}_{\text{df}_G}, \underbrace{n-I}_{\text{df}_T - \text{df}_G}\right) \tag{19}$$

More on this when we discuss ANOVA...

Unpooled variance

If we cannot see from the data that the independent samples are drawn from the same population, or that the I variances are similar, then we use unpooled variance. In general, we do not consider using this for $I > 2$ (for comparing more than two means). It is mathematically possible, but in practice it is seldom used and certainly not a part of the scope of this course. The main reason being that the calculations required to determine the degrees of freedom is quite complicated (see (21)), and most psychologists use a computer. So, the **only time you use unpooled variance is when you have observable differences between the two groups**. So, you could note that $s_1^2 >> s_2^2$ or perhaps the n_i of each group is low (<30) and unequal; it is something you need to determine for yourself. For example, if your test was about two machines from different manufacturers and whether they can complete the same task in the same amount of time, then you would use unpooled variance. If you were exploring behavioural data two culturally different countries, you would use unpooled variances.
<u>Rule of thumb:</u> if it's only two groups and you can assume/predict that the variances are unequal, use unpooled.

$$s_{up} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{20}$$

$$\implies t = \frac{\bar{y}_1 - \bar{y}_2}{s_{up}} \sim t(k), \qquad \text{where } k = \frac{(n_1 - 1) \times (n_2 - 1)}{(n_2 - 1) \times C^2 + (1 - C)^2 \times (n_1 - 1)} \tag{21}$$

$$\text{and } C = \frac{s_1^2 / n_1}{s_{up}^2}. \tag{22}$$

Paired data

I don't think I need to refer much to this, however if you are testing whether there is any effect before and then after, consider this to be paired data. In that case, you don't use either pooled or unpooled variance, but look instead at the variance of differences. That is, transform your data from x and y to $d = x - y$, for instance.

ANOVA

One-way

In Table 1 is an incomplete tabulated output for a test comparing the means between groups. Can you fill in the missing values?

	SS	df	MS	F	sig.
G	91.467		45.733		0.021
E	276.400	27			
T	367.867				

Table 1: ANOVA one-way table

Some background on the test: a teacher wants to know if the starting level of her pupils affects the mean length of time to complete the exam. Formulate the null and alternative hypotheses.

$$H_0: \tag{23}$$
$$H_a: \tag{24}$$

Summarise your findings of this test: In Table 1, can have that the sum of squares between the groups (SSG) is 91.467 and that the mean square between groups (MSG) is 45.733. You want to find out how these two are related, and notice that $45 \times 2 = 90$, which is indeed the degrees of freedom for groups ($\text{df}_G = 2$). Now, you can conclude that you have 3 groups in total ($I = 3$).
Moving on the row marked 'E': there is an evident relationship between the mean squared error within each group (MSE) of 276.400 and the degrees of freedom for the error (df_E) of 27. That is, $\text{df}_E \times 10 \approx \text{MSE}$; using the calculator, you find that MSE is equal to 10.237. Given that df_E ($= n - I$; $I = 3$) is 27, we know that each group has 30 participants ($n = 30$).
Now that we know n, and hence df_T is 29, we can calculate the variance of our data (MST) as 12.685.
In order to calculate our F statistic, we need to under exactly what it is: F is the ratio of variation between and within groups. There are three distinct cases which we will look at now.

$0 \leq F < 1$

In this case (refer to Figure 1), we know that the variance in the data which can be explained by the differences between the groups, is less than the variance within the groups themselves. Either, there is not much difference between the groups, or there is a lot of variation within the groups. In both cases, you would need to perform post-hoc tests, such as contrasts, to confirm or deny your findings. Evidently if $F = 0$, then there is no variation between the groups, i.e. $\bar{y}_1 = \bar{y}_2 = \cdots = \bar{y}_I = \bar{y}$.

$$F = \frac{\text{MSG}}{\text{MSE}} < 1 \tag{25}$$
$$\implies \frac{\text{SSG}}{\text{df}_G} < \frac{\text{SSE}}{\text{df}_E} \tag{26}$$
$$\implies \text{variance between} < \text{variance within} \tag{27}$$

Figure 1

$F \approx 1$

We can note the implications of an F near to one as 'good'; we conclude that the variance in the data is not solely due to variance between groups but equally within. We call this as 'good' because of the complications that arise when we find significant results (more on this in a moment). This F tells us that the variations between the groups is proportional to the variance within the groups themselves, so the effects of being in any particular group are not evident in the data. A very simple example of this is test scores between schools: you might want to test whether attending a more prestigious has any outcome on the results of the students themselves. So, you would have I equal to the number of schools (ranked in order of prestige) and J equal to the number of students at each school (in a particular graduating year). If you received an F value close to 1, then you would conclude that there is no significant advantage benefited to

students who attend a prestigious school, in terms of grades. Additionally, note the following relationship:

$$\left.\begin{array}{l} F = \frac{\operatorname{Var} y - s_p^2}{s_p^2} = \frac{\operatorname{Var} y}{s_p^2} - 1 \\ F \approx 1 \end{array}\right\} \implies \frac{\operatorname{Var} y}{s_p^2} \approx 2. \tag{28}$$

This says that, if the total variation in y is twice the size of the collective variation within each group, then any observed variation between the groups in our sample is acceptable (accept null hypothesis).

F > 1

In this case, we can conclude that we have a significant result: there is a great difference between the groups. With respect to the previous example, we would conclude that the prestige of a school has an effect on the grades of attendees; if F is **much** larger than 1, we could say that the effect is **profound** or immense. Within each school, there is not much variation in the data compared with the variation between the schools (e.g. $I = 3$): Figure 2 is a pictorial example of $F > 1$, which would lead us to assume that further tests (e.g. contrasts) need to be performed in order to determine which school benefits the greatest advantages to it's attendees.

So, back to our table! We can discern that we will have an $F > 1$, as $10 \times 4.5 = 45$, and indeed we have $F = 4.467$ (see the completed Table 2). Relating this to the significance of 0.021 found by the software, the probability of achieving an F more extreme than $f^*_{(2,27\,|\,\alpha=0.05)} = 3.354$ in any repetition is 0.021, which is less than our $\alpha = 0.05$.

The usual method of formulating the null and alternate hypotheses (as you may have figured, by now), is

$$H_0 : \mu_1 = \mu_2 = \mu_3, \text{ i.e. there is no difference between the three groups;} \tag{29}$$

$$H_a : \text{there exists a difference somewhere between the groups.} \tag{30}$$

Using the F statistic and p-value, we reject the null hypothesis at a 5% significance level in favour of the alternate hypothesis, and conclude that at least one of the groups is different from the others. In order to determine which group differs the most, we might formulate some contrasts based on our plot of the data, similar to the one in Figure 2. For example if one of the groups, say group three, was distinctly further away from groups one and two, we might decide to find an a, such that $0 \leq a \leq 1$ and test:

$$H_0 : \frac{a\,(\mu_1 + \mu_2)}{2} = (1 - a)\mu_3 \tag{31}$$

If we find that an a near to zero provides us with a insignificant results (i.e. not reject H_0), then we know that group three dominates. However, if an a near to one has this provision, then we know that a groups one and two equally dominate. Here, dominate means that they greatly contribute to the difference between groups (i.e. large comparable mean/s).

The background on the data set tells us that this is indeed about schooling, but only one particular test (limited factors $\implies$ ANOVA I, and not ANOVA II). We may conclude that the findings of the test imply that the starting level of the pupils does indeed affect the mean length of time to complete the exam. Without knowing any more information, such as individual group means $\bar{y}_i$, standard deviations sd_i and number per group n_i, we conclude our testing here. If this information was available to us, we could run some contrasts to find which particular starting level, beginner, intermediate or advanced, provided the biggest advantage on this particular exam.

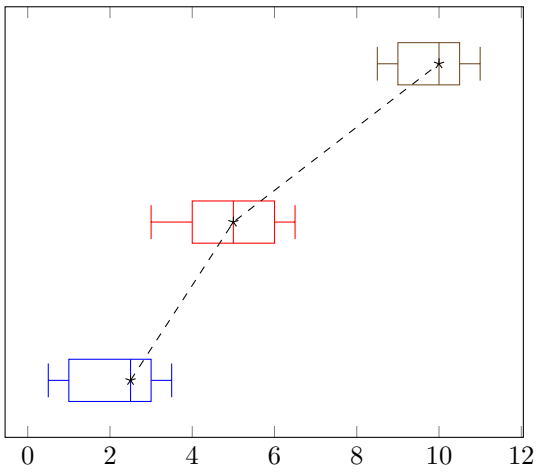

Figure 2: The box plots display the variance **within** groups, which is noticeably small, and the dashed line displays the variance **between** groups, which is noticeably large.

	SS	df	MS	F	sig.
G	91.467	2	45.733	4.467	0.021
E	276.400	27	10.237		
T	367.867	29	12.685		

Table 2: ANOVA one-way table (completed)

Two-way

Fun time!!!
So, not only do you have two or more groups (I), but we now consider that there might be multiple effects (A, B, $C, \dots$) and their interactions ($A \times B$, $A \times B \times C, \dots$). We will start with a simple example, and progress from there:
Perhaps a farmer wants to investigate the effects of manure (A) and nitrogen-based fertiliser (B), and also the combination of both ($A \times B$), on the yield of their corn. The output of an ANOVA II test is given below:

	SS	df	MS	F	sig.
A	15.842		15.842		0.029
B	17.298		17.298		0.035
A × B	3.872		3.872		0.273
E	48.000		3.000		
T	85.012	19	4.474		

Table 3: My caption

Can you fill in the table and draw some conclusions?

Summarise your findings: There is a pretty obvious relationship between the columns "SS" and "MS", so we can assume that the degrees of freedom for the factors and the interaction effect are 1. Meaning: that the factors A and B have two categories each, such as high and low levels (of manure and fertiliser). So, then we know that there are in total $2 \times 2 = 4$ groups: high levels of both, high levels of manure v.s. low levels of fertiliser, low levels of manure v.s. high levels of fertiliser, and low levels of both. Further, we can examine the calculation of df_E:

$$\mathrm{df}_E = \underbrace{(n-1) - (I-1) - (J-1) - \overbrace{(I-1)(J-1)}^{IJ-I-J+1}}_{\mathrm{df}_T - (\mathrm{df}_A + \mathrm{df}_B + \mathrm{df}_{A\times B})} = n - IJ. \tag{32}$$

We already have $I = 2 = J$, and $n = 19 + 1$, so $\mathrm{df}_E = 20 - 2 \times 2 = 16$. We remember how to calculate the F

statistic for each factor and interaction:

$$F = \frac{\text{variance between}}{\text{variance within}} = \begin{cases} \dfrac{\text{MSA}}{\text{MSE}}, & \text{when testing if } A \text{ has a significant effect;} \\ \dfrac{\text{MSB}}{\text{MSE}}, & \text{when testing if } B \text{ has a significant effect;} \\ \dfrac{\text{MSAB}}{\text{MSE}}, & \text{when testing if an interaction of } A \text{ and } B \text{ has a significant effect.} \end{cases} \tag{33}$$

When we are talking about 'significant effect', we want to know if the variation between the groups is relatively equal to the variation within the groups. So, in our table we can input $F_A = 15.842/3 = 5.281$, $F_B = 17.298/3 = 5.77$ and $F_{A\times B} = 3.872/3 = 1.291$. So, given that our F statistic for factors A and B are well above 1, we can can conclude that the levels (each) of manure and fertiliser has a profound affect on corn yield. Further, this is evident by the p-values both being under our accepted 5% level. Now it becomes a little harder to discern the right answer concerning the existence on an interaction effect: we have that the F value is above 1.2 and perhaps would think of rejecting H_0, however we must consider the p-value! Repetitions of this study would yield more extreme F values for this interaction effect more than a quarter of time ($\mathbb{P}\left(F > f_{(1,16\,|\,0.05)}\right) = 0.273 > 0.25$), so we may safely conclude that there is no interaction effect. You can see what (graphically) denotes an effect in the slides from lecture 5 (slide 26-31).
If you were to advise the farmer, what advice would you give?

Previously we found that factors A and B both had a main effect, but there was no significant interaction effect. Now we are interested in conducting a basic two-way ANOVA without the interaction effect. In order to do this, we include the data from the interaction effect into the error terms:

$$\text{SSE}_{\text{new}} = \text{SSE} + \text{SSAB} \qquad \text{df}_{\text{E, new}} = \text{df}_{\text{E}} + \text{df}_{\text{A}\times\text{B}} \tag{34}$$

$$\implies \text{MSE}_{\text{new}} = \frac{\text{SSE}_{\text{new}}}{\text{df}_{\text{E, new}}} \tag{35}$$

Something important to remember for the exam: the "spread of means" is not equal to the "mean spread of the data". The former refers to the variance between means (MSG) and the latter, the mean variance (MSE). The professor will use language like this in order to confuse you!!

Contrasts

Why do we perform contrasts? Simply put, it is to reduce the overall statistical error: if α% for each test and you have I groups then you will need to perform $I \times (I-1)/2$ tests, meaning that

$$\begin{aligned} \text{overall error rate} &= \mathbb{P}\,(\text{at least one false rejection}) && (36)\\ &= \mathbb{P}\left(\text{at least one Type I error} \,\middle|\, \frac{I\times(I-1)}{2} \text{ tests}\right) && (37)\\ &= 1 - \mathbb{P}\left(\text{no Type I errors} \,\middle|\, \frac{I\times(I-1)}{2} \text{ tests}\right) && (38)\\ &\approx 1 - (1-\alpha)^{I\times(I-1)/2} && (39) \end{aligned}$$

So, for a simple ANOVA I with 5 groups at 1% significance level, this means we need to perform $5\times4/2 = 10$ tests resulting in an error rate of $1 - 0.99^{10} = 0.0956$, i.e. 9.6% of a Type I error, which is higher than our accepted 5% level (chance capitalisation). It is possible to plan for this by introducing contrasts **prior to undertaking the test**.
Assume, again, $I = 5$ and we want to know which group accounts for the largest deviation of the data. **Remember**, $SST = SSG + SSE$ - this means that the total variance in the data set is either due to the variance between groups or within the groups themselves. If we have an $F > 1$ (and $p < \alpha$ given the sample size of each group is "large enough"), we know that it is attributable to variance between groups and now we want to know which particular group/s are different. We have that,

$$\bar{x}_1 > \bar{x}_2 \qquad \text{Is group 1 different to group 2?} \qquad \begin{matrix} H_{01} : \mu_1 - \mu_2 = 0 \\ H_{a1} : \mu_1 - \mu_2 > 0 \end{matrix} \tag{40}$$

$$\text{coefficients: } (1, -1, 0, 0, 0) \tag{41}$$

$$\begin{matrix}\bar{x}_1 < \bar{x}_3 \\ \bar{x}_2 < \bar{x}_3\end{matrix} \quad \text{Are groups 1 and 2 different to group 3?} \quad \begin{matrix}H_{02} : \frac{\mu_1+\mu_2}{2} - \mu_3 = 0 \\ H_{a2} : \frac{\mu_1+\mu_2}{2} - \mu_3 < 0\end{matrix} \tag{42}$$

$$\text{coefficients: } \left(\frac{1}{2}, \frac{1}{2}, -1, 0, 0\right) \tag{43}$$

$$\bar{x}_5 > \max\{\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4\} \quad \text{Is group 5 the most different?} \quad \begin{matrix}H_{03} : \frac{\mu_1+\mu_2+\mu_3+\mu_4}{4} - \mu\mu_5 = 0 \\ H_{a3} : \frac{\mu\mu_1+\mu_2+\mu_3+\mu_4}{4} - \mu_5 < 0\end{matrix} \tag{44}$$

$$\text{coefficients: } \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -1\right) \tag{45}$$

We don't know the true values of $\mu_1, \ldots, \mu_5$, so we estimate with $\bar{x}_1, \ldots, \bar{x}_5$ to produce an estimated contrast value c and it's associated t statistic: $c = \sum_{i=1}^{I} a_i \bar{y}_i$ and $t = c/\text{SE}_c \sim t\,(n - I)$ (under H_0). Let's look at an example:

Group	$\bar{y}_i$	s_i	n_i
1	33.31	3.63	9
2	28.72	2.91	15
3	32.46	3.98	9
Total	30.99	3.3	33

Table 4: My caption

Notice that group 2 has the lowest mean and sd, and the largest sample size? This means that we can infer already that group 2 is the most different. First, we should calculate the pooled variance:

$$s_p = \sqrt{\frac{\sum_{i=1}^{3}(n_i - 1)s_i^2}{\sum_{i=1}^{3}(n_i - 1)}} \tag{46}$$

$$= \sqrt{\frac{8 \times 3.63^2 + 14 \times 2.91^2 + 8 \times 3.98^2}{33 - 3}} \tag{47}$$

$$= \sqrt{\frac{350.69}{30}} \tag{48}$$

$$= \sqrt{11.69} \tag{49}$$

$$= 3.42. \tag{50}$$

So we test,

$$H_{01} : \mu_2 = \mu_1 \tag{51}$$

$$H_{a1} : \mu_2 < \mu_1 \tag{52}$$

$$\implies c = -1 \times \bar{y}_1 + 1 \times \bar{y}_2 + 0 \times \bar{y}_3 = -4.59. \tag{53}$$

$$\implies \text{SE}_c = s_p \sqrt{\sum_{i=1}^{3} \frac{a_i^2}{n_i}} \tag{54}$$

$$= 3.42 \times \sqrt{\frac{1}{9} + \frac{1}{15}} \tag{55}$$

$$= 3.42 \times \sqrt{0.178} \tag{56}$$

$$= 3.42 \times 0.422 \tag{57}$$

$$= 1.44. \tag{58}$$

$$\implies t = \frac{c}{\text{SE}_c} \tag{59}$$

$$= \frac{-4.59}{1.44} \tag{60}$$

$$= -3.1875 \sim t\,(33 - 3)_{\alpha/2=0.025}\,. \tag{61}$$

Our critical t^* value for 32 df and (two-tailed) 5% significance is -2.042, which is larger than our t-statistic. So we must conclude to reject H_0 in favour of H_a; you can read the p-value from a table as around 0.003. If we were to construct a confidence interval about c:

$$\text{CI} = (c - t^* \times \text{SE}_c, c + t^{**} \times \text{SE}_c) \tag{62}$$

$$= (-3.1875 - 2.042 \times 1.44, -3.1875 + 2.042 \times 1.44) \tag{63}$$
$$= (-6.13, -0.25) \tag{64}$$

Note that the entire confidence interval is located to the left of zero? So, we certainly reject the null hypothesis, as even when it is assumed, we do not have zero contained in the interval!

Confidence intervals

The general equation for confidence intervals involving multiple comparisons is given by

$$\mathrm{CI}_{ij} = (\bar{y}_i - \bar{y}_j) \pm t^{**} \times s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{65}$$

For this point of the course, you have two options for your critical t^{**}:

1. **Bonferroni:** adjust the significance of each test to ensure that overall error rate is less than the specified α:

$$k = \text{the number of tests} \tag{66}$$
$$= \frac{I \times (I-1)}{2}, \qquad \text{where } I \text{ is the number of groups.} \tag{67}$$

   Let α^* be the significance of **each** of the k tests, then

$$\alpha^* = \frac{\alpha}{k} \qquad \Longrightarrow t^{**} = t^*_{\substack{1-\alpha/2k \\ \nu=\mathrm{df}_E}}. \tag{68}$$

2. **Least significant differences (LSD):** we use this for $I = 3$ groups, otherwise use Bonferroni. This is because LSD method does not alter the significance, but only each individual test is improved, and $I = k$ for 3 groups.

$$t^{**} = t^*_{\substack{1-\alpha/2 \\ \nu=\mathrm{df}_E}} \tag{69}$$

Important:

$$p_{\text{Bonferroni}} = \mathbb{P}\left(|T| > t_{1-\alpha/2k}\right) = \frac{\mathbb{P}\left(|T| > t_{1-\alpha/2}\right)}{k} = \frac{p_{\text{LSD}}}{k} \tag{70}$$

The p-values are scaled with respect to the number of tests performed.

Kruskal-Wallis procedure

ANOVA assumptions are normality, homoskedasticity and independence, and if they are severely violated then we turn to the Kruskal-Wallis procedure (non-parametric ANOVA). The null hypothesis is similar to ANOVA, however the violation of assumptions leads us to a new direction: "the **distribution** is the same in all groups". The alternative is that the scores in some groups are systematically larger.
Begin by ordering all n scores from lowest to highest, assigning rank 1 to the lowest. If some scores are equal, they receive the mean of the ranked score, e.g. there are two scores of 9 and it is the fifth lowest score (taking up spaces 5 and 6) then each of the '9"s receive a rank of $(5+6)/2 = 5.5$.
Now that each score has a rank, you need to separate the ranked scores back into their respective groups. Sum the rankings in each group to yield R_i. Furthermore,

1. if the sample sizes n_i are small (<5), then we use the following test statistic in ANOVA I:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{n(n+1)} \mathrm{SSG}_{\text{ranks}}. \tag{71}$$

2. if the sample sizes are not too small, say $n_i \geq 5$, then use the following approximation: $H \sim \chi^2(I-1)$.

For example, you are investigating the effects of exercise on depression. So you have three groups: no exercise, 20 minutes of jogging per day, and 60 minutes of jogging per day. In order to simplify, we assume that each participant is equivalently depressed, then at the end of the month you ask each participant how depressed they feel as a score out of one hundred, where 1 is totally miserable and 100 is ecstatically happy.
The method of testing is self-recorded and the scores (ordinal) given are non-parametric, so in order to draw conclusion we will need to use the Kruskal-Wallis procedure.

Table 5: This table contains the self-recorded ratings of 1 to 100 from 27 depressed people, where 1 is totally miserable and 100 is ecstatically happy. The ratings were ask for after performing daily jogging for a month, in groups of 20 and 60 minutes per day. The third group is the control group, who were asked to not exercise every day.

$i = 1, 2, 3$; $n_i = 8$	**no exercise**	**20 min/day**	**60 min/day**
	23	22	59
	26	27	66
	51	39	38
	49	29	49
	58	46	56
	37	48	60
	29	49	56
	44	65	62
$\bar{\mathbf{x}}_\mathbf{i}$	**39.63**	**40.63**	**55.75**
$\mathbf{s_i}$	**12.85**	**14.23**	**8.73**

Looking at Table 5, the minimum rating given was 22 and the maximum was 66, so 22 is given rank 1 and 66 is given the lowest rank. Tied scores get the average of the rankings they would've received. I've summarised the next few steps in the following ranked table:

Table 6: This table contains the rankings of the scores given, as well as their sums per group.

$\mathbf{i = 1, 2, 3}$; $\mathbf{n_i = 8}$	**no exercise**	**20 min/day**	**60 min/day**
	2	1	20
	3	4	24
	16	9	8
	14	5.5	14
	19	11	17.5
	7	12	21
	5.5	14	17.5
	10	23	22
$\bar{\mathbf{x}}_\mathbf{i}$	**39.63**	**40.63**	**55.75**
$\mathbf{s_i}$	**12.85**	**14.23**	**8.73**
$\mathbf{R_i}$	**76.5**	**79.5**	**144**

You can see in Table 5 that there is a rating of 49/100 in each column, which is the 14th, 15th and 16th lowest ranking. So they each get the ranking of $(14 + 15 + 16)/3 = 14$. As each $n_i = 8 > 5$, we assume that the H test statistic is approximated by the χ^2 distribution with 2 degrees of freedom. Calculating H:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(n+1) \tag{72}$$

$$= \frac{12}{24(24+1)} \sum_{i=1}^{3} \frac{R_i^2}{8} - 3(24+1) \tag{73}$$

$$= \frac{12}{24 \times 25} \left(\frac{76.5^2}{8} + \frac{79.5^2}{8} + \frac{144^2}{8} \right) - 3 \times 25 \tag{74}$$

$$\approx 7.271 \sim \chi^2(2) \tag{75}$$

Looking at the p-values for a $\chi^2(2)$, we can surmise that the probability of obtaining an H statistic more extreme than what we calculated is between 5% and 2.5%:

$$0.025 \leq \mathbb{P}\left(H > 7.271\right) \leq 0.05 \tag{76}$$

Given the small p-value for the H statistic, we can conclude that there is some difference between the three groups, i.e. the effects of exercise on depression is evident. Looking back at Table 6, you may notice that

the mean of group 3 is relatively larger than the others, and it's standard deviation is comparatively smaller. Meaning that we may infer that group three is the "different" group: 60 minutes of medium exercise (e.g. jogging) has the greatest impact on depression (in terms of improvement), compared to 20 minutes or no exercise (check this with the next section on effect size ☺).

Post-hoc methods

Post-hoc multiple comparisons: tests, using confidence intervals, for differences between **all** pairs of means:

$$\text{CI}_{ij} \text{ for comparing the means of groups } i \text{ and } j\text{: } (\bar{y}_i - \bar{y}_j) \pm t^{**}_{\text{df}_E} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{77}$$

Note the use of t^{**}? This refers you to the section on LSD and Bonferroni!
All this is about, is two main things: do your CI's overlap; is your assumed value inside the CI.

1. So, if we are comparing groups (comparing means), then we would like our assumed value $\mu_i - \mu_j = 0$ (H_0 : all groups are the same, i.e. all means are the same) to be inside the confidence interval. If it does not, we might need to look at our test for any errors (e.g. low effect size, small sample size, chance captilisation). However, if there are no errors present (e.g. large effect size, $n \geq 30$, Bonferroni method used, testing procedure is widely regarded as effective, etc.) then you would **reject** H_0. Why? You assume something, then you construct your test in such a way that there is no possible room for error, however the result shows that $(1-\alpha)\%$ of all samples lead to an interval that does not cover the unknown parameter. So our assumed parameter must be incorrect ($\mu_i - \mu_j \neq 0$). If our samples collected are large enough, the CLT allows us to conclude that true parameter lies somewhere closer to the estimated value $a = \bar{y}_i - \bar{y}_j$.

2. What does it mean if the CI's overlap? Referring back to Figure 2, you can see the boxplots represent variation within the groups, and that the red and blue boxplots overlap whilst the other box stands alone. Rather than adding another figure here, imagine that the boxplots are displaying 95% confidence intervals. So you have that the confidence intervals of groups one and two overlap, whilst group three's confidence interval is completely separated. This means that it is likely that the population means for groups one and two are equal, and the pop.mean of group three is likely greater. In relation to (77), if more than two CI's overlap sequentially (e.g. the upper bound of the CI for $\mu_1 - \mu_2$ is contained in the CI for comparing $\mu_2 - \mu_3$, and the upper bound for the CI comparing $\mu_2 - \mu_3$ is contained in the CI for comparing $\mu_3 - \mu_4$, ...), then you might assume that there it is likely that the groups are all (relatively) the same. How much they overlap is indicative of how likely it is that the groups are all the same. For example, if the upper bound of the CI for $\mu_1 - \mu_2$ is more than the estimate for comparing $\mu_2 - \mu_3$, namely the midpoint $\bar{y}_2 - \bar{y}_3$, then you might assume that these groups all differ by the same amount (more than relatively). If the CI's (almost) coexist, then you can be sure that these groups certainly differ by the same amount - if zero is included in all the CI's, then the groups are the same (no difference).

Effect size

What is "effect size" in the relation of statistics? It's an objective and standardised way to determine if there is indeed an observable effect in the data. The usual method is by way of ratio, so if the ratio is on the large end of the scale then you can say that the researcher will notice an effect by the factors/groups in the data. Another way to think about it, is to regard effect size as the statistical 'yard stick'; "relative to the size of my hand, how big is this ant? So, will I notice it (see it) crawling onto my hand?".
Whilst writing this, I came across a really great website that explains effect size in lay-mans terms. I think it is helpful to read, in order to give a 'layman's response' to a psychological research question. It may also help you to visualise effect size: `https://www.theanalysisfactor.com/effect-size/`.

Cohen's d

Cohen's d is used to measure the standardised difference between means in

1. one random sample drawn from a normally distributed population $\left(y \sim \mathcal{N}\left(\mu, \sigma^2\right)\right)$;

$$d = \frac{\bar{y} - \mu}{\sigma} \tag{78}$$

   This is a z-score, and the old "68-95-99.7" rule gives an indication of how we might view this: if you have a z-score of less than 1, then you know that your estimate based on the data, $\bar{y}$, lies in the middle 68%.

Let's consider only $|d|$ and assume $|d| \leq 0.2$ (small effect size - see slide 15 from lecture 2) - checking the z-table we conclude that there is less than 16% of the population, which is less extreme than our sample estimate ($\mathbb{P}(|D| < 0.2) = 0.15852$). The key point here is that this is a **standardised score**, so you can use it to measure the small thing across different groups of populations. The simplest example is comparing the group means for different classes within a school for the same test. If you know the results for the test is normally distributed with mean μ and standard deviation σ, you can compare whether the means from different classes are noticeably different from what you expect (μ).

2. two random samples drawn from normally distributed populations

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s_p}. \tag{79}$$

N.B.: the use of pooled variance implies that we assume homoskedasticity; if this is true, then s_p is the best approximation for the true standard deviation σ.
Similar to the above, however we assume that $\mu_1 = \mu_2$ and so $\mu_1 - \mu_2 = 0$ is "invisible" in the equation. This is almost a t-statistic (larger standard error due to the omission of $\sqrt{1/n_1 + 1/n_2}$ in the denominator), so it gives the standardised difference of sample means between two (assumed equal) groups. This will tell you if the standardised difference (effect size) is is small (≤ 0.2), medium (≈ 0.5) or large (≥ 0.8).

Eta squared

η^2 is the proportion of the total sample variance explained by the effect $(A, B, A \times B, \dots)$.

$$\eta^2 = \frac{\mathrm{SS}_{\text{effect}}}{\mathrm{SST}} \tag{80}$$

Again, it's basically a ratio: if the variation in the data is due mostly to the effect, then this ratio is near (or more) 0.14. In ANOVA II, "effect" can be factor A, interaction effect $A \times B$, etc. In ANOVA I, the only "effect" we consider is group variation ($\mathrm{SS}_{\text{effect}}$ =SSG) so $\eta^2 = R^2$, where $R^2 = [\mathrm{Cov}(\hat{y}, y)]/s_{\hat{y}}^2 s_y^2$ is the squared ratio of covariance and variance of the data y and the predicted model $\hat{y}$ - **percentage of variance explained by the model** (see the section on regression). So for ANOVA I (not considering regression models), $\eta^2 \times 100$ is the percentage of variation in the data as explained by the variation between groups.
Advantages:

- effects are additive for balanced groups, i.e. $n_1 = n_2 = \cdots = n_I$:

$$\sum_{\text{all effects}} \mathrm{SS}_{\text{effect}} = \mathrm{SSM} \tag{81}$$

Disadvantages:

- η^2 depends on the number and size of the remaining effects. So if you have a lot of factors/interactions, or if one factor/interaction contributes the greatest, then you might have an η^2 which is small, despite the effect size being (actually) medium, for instance:

$$\begin{aligned}
\eta^2 &= \frac{\mathrm{SS_A}}{\mathrm{SST}} && (82)\\
&= \frac{\mathrm{SS_A}}{\sum_{\text{all effects}} \mathrm{SS}_{\text{effect}} + \mathrm{SS_E}} && (83)\\
&= \frac{\mathrm{SS_A}}{\mathrm{SS_A} + \mathrm{SS_B} + \mathrm{SS_C} + \cdots + \mathrm{SS_{A \times B}} + \cdots + \mathrm{SS_{A \times B \times C}} + \cdots + \mathrm{SS_E}} && (84)\\
& && (85)
\end{aligned}$$

- η^2 does not estimate the proportion of variance accounted for in the **population** - it is a **biased estimator**. It always overestimates the explained variance in the population, despite being 'good' for the sample.

Partial eta squared

This tries to solve the disadvantages of η^2, by restricting the ratio to only the effect you are interested in. The disproportional affect of adding effects to the denominator are cancelled out, however there is still the same problem of it being a biased estimator. Additionally, like η^2, $\eta_p^2 = R^2$ in ANOVA I.

$$\eta_p^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{effect}} + \text{SSE}} \tag{86}$$

Advantages:

- η_p^2 does not depend on the remaining effects, like η^2 does. This is because the denominator omits the explained variation from other effects, and focuses solely on the ratio of explained variation (of the effect in question) vs. the combination of explained and unexplained, but restricted to a particular effect.

Disadvantages:

- Effects are no longer additive for balanced designs (see the section on η^2 for further explanation of this).
- η_p^2 only estimates the proportion of variance accounted for in the sample, and not in the population. Thus, it overestimates the population effects.

N.B.: SPSS only outputs η_p^2, so in the exam if you see an SPSS output with η^2, you can be sure which it is ☺.

Omega squared

ω^2 aims to **exactly** estimate population effects, rather than sample effects, so it is an unbiased estimator.

$$\omega^2 = \frac{\text{SS}_{\text{effect}} - \text{df}_{\text{effect}} \times \text{MSE}}{\text{MSE} + \text{SST}} \tag{87}$$

Advantages:

- ω^2 does not overestimate population effects, like η^2 and η_p^2 do.

Disadvantages:

- Effects are no longer additive for balanced designs (see the section on η^2 for further explanation of this). Furthermore, ω^2 can be negative! What does this mean?

$$\omega^2 < 0 \tag{88}$$

$$\iff \text{SS}_{\text{effect}} - \text{df}_{\text{effect}} \times \text{MSE} < 0 \tag{89}$$

I should note here that the symbol $\iff$ means 'if and only if' and denotes equivalency. So in this case, 'ω^2 is negative' is equivalent to:

$$\text{SS}_{\text{effect}} < \text{df}_{\text{effect}} \times \text{MSE} \tag{90}$$

$$\iff \frac{\text{SS}_{\text{effect}}}{\text{df}_{\text{effect}}} = \text{MS}_{\text{effect}} < \text{MSE} \tag{91}$$

$$\iff F = \frac{\text{MS}_{\text{effect}}}{\text{MSE}} < 1. \tag{92}$$

I found a great website which is written in a fun and bitchy tone about the benefits of which effect size to use: `http://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html`. I hope you have a laugh, too!

Power

We will run quickly over statistical power: **power is how well your test works**. You, as a psychologist/statistician, want to ensure that you only make necessary changes to society and do so at the right time. What is the right time? When the data says you have to!
Consider that you are running a test on whether a new psychological disorder should be included in the new DSM. You structure your test as a hypothesis test at a significance level of $\alpha = 0.015$ (one-tailed, two-sample t-test), and find a p-value which is smaller than your acceptance level.

$$H_0 : \mu = \mu_1 - \mu_2 = 0 \tag{93}$$

$$H_a : \mu > 0 \tag{94}$$

If you have already rejected your null hypothesis, that means your sample statistic t is more extreme than the accepted value for a particular significance level **and** degrees of freedom:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \tag{95}$$

$$\implies \bar{x}_1 - \bar{x}_2 > \mu_0 + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{96}$$

Suppose you later find out a closer estimate that $\mu_0 = 0$ for the difference in means: $\mu_a > 0$. What is probability that you correctly reject H_0, given that you know H_a?

$$\implies \mathbb{P}\left(\bar{x}_1 - \bar{x}_2 > \mu_0 + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \,\middle|\, \mu_a\right) = \mathbb{P}\left(\bar{x}_1 - \bar{x}_2 - \mu_a > \mu_0 - \mu_a + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \tag{97}$$

$$= \mathbb{P}\left(\frac{\bar{x}_1 - \bar{x}_2 - \mu_a}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > \frac{\mu_0 - \mu_a + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) \tag{98}$$

$$= \mathbb{P}\left(T > \frac{\mu_0 - \mu_a}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + t^*_{\substack{\alpha=0.015 \\ \nu=n_1+n_2-2}}\right) \tag{99}$$

$$= 1 - \beta = \text{power}. \tag{100}$$

You can find the power of the test using any population, if you know μ_a (or a value which is extremely close to it - can never be 100% certain in science!).

Correlation and regression

Remember! correlation and regression are two different things, despite being heavily interrelated. Correlation is the measurement of the association between two (or more) variables, whereas regression uses the information provided by the association to predict or extrapolate data.

Pearson's r

Pearson's ρ is the measure of a linear association **in a population** between two random variables X and Y, given by

$$\rho_{X,Y} = \frac{\text{Cov}\,(X, Y)}{\sigma_X \times \sigma_Y}, \tag{101}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively. Researchers do not always have the luxury of knowing the magnitude or direction of a linear relationship in a population, so they must estimate it using sample data. Pearson's r approximates ρ, so that we may draw inferences about the population.

$$r_{x,y} = \frac{\text{Cov}\,(x, y)}{s_x \times s_y} \tag{102}$$

We use subscript x, y to denote the variables which r is measuring the relationship of. We substitute in the equation for sample covariance:

$$= \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}}{s_x \times s_y} \tag{103}$$

We can rearrange the denominators so that the standard deviations for x and y are grouped with their respective expressions:

$$= \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right) \times \left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1} \tag{104}$$

You may notice the familiar expressions for calculating the z scores for x and y for a particular i:

$$= \frac{\sum_{i=1}^{n} z_{x_i} \times z_{y_i}}{n-1} \tag{105}$$

So, we have that Pearson's r is the expected value of the product of z-scores, where the degrees of freedom is $n-1$. If we want to see this equation in another way, we can return to the first expression, and elaborate:

$$r_{x,y} = \frac{\text{Cov}(x,y)}{s_x \times s_y} \tag{106}$$

$$= \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{n-1}}{s_x \times s_y} \tag{107}$$

$$= \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} \tag{108}$$

The $n-1$ in all of the denominators equate out to 1 ('cancel out'):

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{109}$$

Next, we expand the brackets in the denominator:

$$= \frac{\sum_{i=1}^{n}(x_iy_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{110}$$

Using that $n\bar{x} = \sum x_i$, and similarly for y, results in the following expression for r:

$$= \frac{\sum_{i=1}^{n}x_iy_i - n\times\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{111}$$

r is bounded between -1 and 1, where an $r=1$ represents a positive correlation between x and y: $x \propto y$ they are proportional, e.g. $y = a+bx$ (a, b are constants with $b>0$); alternatively, $r=-1$ represents a negative correlation between x and y: $\frac{1}{x} \propto y$ they are inversely proportional, e.g. $y = a-bx$.
If $r \approx 0$, then there is no linear relationship between your variables x_i and y_i: **they are linearly independent**.

Simple linear regression (SLR)

Now that you know what Pearson's r is, you can begin learning about simple linear regression and how they interrelate. You collect a data sample from a population, and after calculating Pearson's r for your sample you conclude that x and y are correlated (later you will want to know if they are correlated in the population - Fisher Z transformation). Now, you wish to predict future outcomes associated with your data sample, and so you construct a model

$$\underbrace{y_i}_{\text{data}} = \underbrace{\beta_0 + \beta_1 x_i}_{\substack{\mu_{y_i} \\ \text{model}}} + \underbrace{\epsilon_i}_{\text{error}} \tag{112}$$

for your **population** which predicts the value y_i given your input x_i, based on the constants β_0 and β_1. Given your input x, the y values are normally distributed with population mean μ_y and variance σ^2, and the error term $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ is independent of x_i. This is an incredibly important assumption, as you do not reasonably expect that your model can capture/test all possible contributions to the data. What you cannot test/capture is ϵ_i - typical examples may be infant life experiences or innate ability. Also, you expect that your model is 'good', so you assume that the mean of ϵ_i is zero.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{113}$$

$$= \mu_{y_i} + \epsilon_i \sim \mathcal{N}(\mu_{y_i}, \sigma^2) \tag{114}$$

Notice that the mean of y_i also has subscript i? This is because the mean for a particular value y_i is dependent on the input x_i, so it is really $\mu_{y \text{ given } x_i}$. If we rearrange this expression in favour of ϵ_i, the population error term, we find that it is normally distributed with mean zero and the same variance as y:

$$\implies y_i - \mu_{y_i} = \epsilon_i \sim \mathcal{N}(0,\sigma^2). \tag{115}$$

You estimate β_0 and β_1 (population coefficients) using your sample, where b_0 and b_1 are the **ordinary least squares estimates** (OLS estimates) of β_0 and β_1. They are called OLS estimators because they minimise the sum of squared errors between the actual data y_i and the predicted data $\hat{y}_i$.

$$\min\left(\sum_{i=1}^{n} e_i^2\right) = \min\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right) \tag{116}$$

$$= \min \left[\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i) \right)^2 \right] \tag{117}$$

The b_0 and b_1 which satisfy this are:

$$b_0 = \bar{y} - b_1 \bar{x}; \qquad b_1 = r_{x,y} \times \frac{s_y}{s_x} = \frac{\mathrm{Cov}\,(x, y)}{s_x^2} \tag{118}$$

where $r_{x,y}$ is the Pearson's r for your data sample, s_y is the standard deviation of the sample for output y, s_x is the standard deviation of the sample for the input x, and $\mathrm{Cov}\,(x, y)$ is the covariance of x and y.
b_0 is your intercept, i.e. the value of y_i when $x_i = 0$, so this just tells you the minimum of the range for your predicted output. Under H_0, you assume β_0 to be zero, which is what you will test: is my sample coefficient b_0 significantly different from zero? Note that if your b_0 is forced to be equal to zero, then the regression line does not run through the centre mass point of $(x_i, y_i) = (\bar{x}, \bar{y})$:

$$\begin{aligned} \hat{y}_i &= b_0 + b_1 x_i && (119)\\ &= (\bar{y} - b_1 \bar{x}) + b_1 x_i && (120)\\ &= \bar{y} + b_1 (x_i - \bar{x}) && (121)\\ &= \bar{y}. && (122) \end{aligned}$$

If we have a predicted output $\hat{y}_i$ equal to the sample mean, there are two possibilities; either,

$$\begin{cases} b_1 = 0: & \text{This implies that } r_{x,y} = 0, \text{ and so } x \text{ and } y \text{ are} \\ & \text{linearly independent.} \\ x_i = \bar{x}: & \text{This implies that the regression line passes} \\ & \text{through the point } (\bar{x}, \bar{y}) \text{ if and only if } b_0 \text{ is} \\ & \text{not equal to zero.} \end{cases} \tag{123}$$

It is also cool to note:

$$\begin{aligned} \hat{y}_i &= (\bar{y} - b_1 \bar{x}) + b_1 x_i && (124)\\ &= \bar{y} + b_1 (x_i - \bar{x}) && (125)\\ &= \bar{y} + \left(r_{x,y} \times \frac{s_y}{s_x} \right) \times (x_i - \bar{x}) && (126) \end{aligned}$$

$$\implies \underbrace{\frac{\hat{y}_i - \bar{y}}{s_y}}_{z\text{-score for } \hat{y}_i} = r_{x,y} \times \underbrace{\frac{x_i - \bar{x}}{s_x}}_{z\text{-score for } x_i} \tag{127}$$

b_1 is your slope coefficient: when my x_i increases, does my y_i increase or decrease and at what rate? You can see from (127) that $r_{x,y}$ is the slope of the regression line of the standardized data points. If your variables are strongly linearly correlated, such as $r = 1$, then $b_1 \approx {}^{s_Y}/_{s_X}$ which is the ratio of standard deviations between your independent and dependent variables; if $r = -1$, then $b_1 \approx -{}^{s_Y}/_{s_X}$, meaning that $\hat{y}_i$ will decrease as x_i increases, respective to the rate ${}^{s_Y}/_{s_X}$. It is important to note that the sign of b_1 (i.e. negative or positive) is strictly governed by the sign of the covariance as s_y and s_x are always strictly greater than zero (equal to zero if and only you have a single data point for your sample). So r not only tells you about the observed relationship in the sample, but what to expect of your predictive regression line.
So, now you have your prediction model $\hat{y}_i = b_0 + b_1 x_i + e_i$, you can calculate the (squared) **standard error of the estimate**, which estimates σ^2, the variance of the population y:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n} e_i^2}{\underbrace{n - 2}_{\mathrm{df}}} && (128)\\ &= \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2} && (129) \end{aligned}$$

If you are asked, "what is the percentage of explained variance ..." then you know that they are asking you for R^2 (or adjusted R^2). If the regression model is linear, then $R^2 = r^2$ and the percentage of explained variance is $R^2 \times 100\%$.

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{residuals}}}{\mathrm{SS}_{\mathrm{total}}} \tag{130}$$

Recall that,

$$\mathrm{SS}_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{131}$$

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{132}$$

$$= \mathrm{SS}_{\text{regression}} + \mathrm{SS}_{\text{residuals}}. \tag{133}$$

$$\implies R^2 = \frac{\mathrm{SS}_{\text{total}} - \mathrm{SS}_{\text{residuals}}}{\mathrm{SS}_{\text{total}}} \tag{134}$$

$$= \frac{\mathrm{SS}_{\text{regression}}}{\mathrm{SS}_{\text{total}}} \tag{135}$$

$$= \frac{\text{explained variance}}{\text{total variance}} \tag{136}$$

If you have more than one independent variable, e.g. $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$, then you consider adjusting for the increase in parameters:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p-1} \tag{137}$$

$$= 1 - \frac{\mathrm{SS}_{\text{residuals}}}{\mathrm{SS}_{\text{total}}} \cdot \frac{n-1}{n-p-1} \tag{138}$$

$$= 1 - \frac{\mathrm{SS}_{\text{residuals}}/(n-p-1)}{\mathrm{SS}_{\text{total}}/(n-1)} \tag{139}$$

Comparing the equation for R^2 and $\bar{R}^2$ (adjusted R^2), the 'adjustment' is the degrees of freedom for the residuals. R^2 assumes that $\mathrm{df}_{\text{residuals}} = n-1$, which leads to a biased estimate for the population variance of the residuals. By adjusting the degrees of freedom to $n-p-1$, where p is the number of independent variables (not including constant β_0), we gain an unbiased estimate. This will be further explored during multivariate regression.

If you are asked, "what is the variability about the line of regression", then you know that they are asking you for the standard error of the estimate $s = \sqrt{\sum e_i^2/(n-2)}$.
If you are asked to construct a $(1-\alpha)\,\%$ CI for the population regression parameter β_0 (or β_1), then you know that they are asking you for a t-statistic based CI (see Table 7):

Table 7

	β_0	β_1
Estimate	$b_0 = \bar{y} - b_1\bar{x}$	$b_1 = r_{x,y} \times \frac{s_y}{s_x}$
Mean of estimate	$\mu_{b_0} = \beta_0$	$\mu_{b_1} = \beta_1$
CI	$b_0 \pm t^*_{n-2} \times \mathrm{SE}_{b_0}$, where SE_{b_0} approximates σ_{b_0} and is computed by SPSS	$b_1 \pm t^*_{n-2} \times \mathrm{SE}_{b_1}$, where SE_{b_1} approximates σ_{b_1} and is computed by SPSS
Test	H_0: $\beta_0 = 0$ vs H_a: $\beta_0 \neq 0$	H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$
Statistic	$t = \frac{b_0 - \beta_0^{=0,\text{ under } H_0}}{\mathrm{SE}_{b_0}} \sim t(n-2)$	$t = \frac{b_1 - \beta_1^{=0,\text{ under } H_0}}{\mathrm{SE}_{b_1}} \sim t(n-2)$
Assumptions	$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2_{b_0}\right)$	$b_1 \sim \mathcal{N}\left(\beta_1, \sigma^2_{b_1}\right)$
	$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$	

In general, we are most interested in the results of testing done on the true parameter β_1: if we do not reject H_0, then we conclude that x and y do **not** have a linearly correlated relationship in the population (could be independent variables or another relationship is more appropriate, e.g. semi-logarithmic).

SLR pop-quiz

1. Why do we perform (linear) regression?
2. Based on an SPSS output, what inferences can we make?
3. When might we examine the residual plot, and what inferences might we draw from it?
4. What is the "point of centre mass" on a scatter plot?
5. What tests might we perform on the output of a (linear) regression?
6. What is R^2, and why/when do we adjust it?
7. If we compute the Pearson's correlation coefficient, what can we infer from this statistic? Think about:
   - scatterplot (x_i, y_i);
   - residual plot $(y_i, \hat{y}_i)$;
   - roles in regression equation $\hat{y}_i = b_0 + b_1 x_i$.

Fisher Z-transformation

In the previous sections, we looked at the population model and tested goodness of fit for a linear model. In order to further investigate the existance of a linear relationship between x and y in the population, we look distinctly at the population correlation coefficient, Pearson's ρ, and its sample counterpart, Pearson's r. The following t statistic is used to test H_0: $\rho = 0$ (linear independence) against H_a: $\rho \neq 0$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \tag{140}$$

The problem we face is when we reject the null hypothesis, and consider population correlation coefficients which are not equal to zero - H_0: $\rho = a$ where a is a non-zero constant against H_a: $\rho \neq a$. Under this assumption, r is not normally distributed, so how do we construct a confidence interval around r for ρ? The answer is the Fisher Z transformation to r_z (approximately normal):

$$r_z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \sim \mathcal{N}\left(\rho_z = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \sigma^2_{r_z} = \frac{1}{n-3}\right) \tag{141}$$

This says that our transformed sample correlation coefficient, r_z, is normally distributed with mean ρ_z, which is the transformed population correlation coefficient (just insert ρ in place of r in the equation for r_z), and standard deviation $1/\sqrt{n-3}$.

$$Z = \frac{r_z - \rho_z}{1/\sqrt{n-3}} \sim \mathcal{N}(0,1) \tag{142}$$

If we want to find out if r_z is significant, i.e. the probability of achieving a more extreme value, we can do a z-test:

$$\mathbb{P}\left(|Z| > \frac{r_z - \rho_z}{1/\sqrt{n-3}}\right) = p \leftarrow \text{look this up from the } z\text{-table.} \tag{143}$$

The above z-test is two tailed, because our alternative hypothesis is that $\rho \neq a$, so $\rho < a$ or $\rho > a$. Now we can construct a $(1-\alpha)\%$ confidence interval about r_z for ρ_z, using the critical z-value.

$$\implies \text{CI for } \rho_z: \quad r_z \pm z^* \frac{1}{\sqrt{n-3}} \tag{144}$$

So we have an upper and lower bound now for our confidence interval for ρ_z:

$$\text{LB}_{\rho_z} = r_z - z^* \frac{1}{\sqrt{n-3}} \qquad \text{UB}_{\rho_z} = r_z + z^* \frac{1}{\sqrt{n-3}} \tag{145}$$

This does not tell us anything about ρ, only ρ_z, so in order to have a confidence interval for ρ we can transform r_z back to r using:

$$r_z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \tag{146}$$

$$\implies 2r_z = \log\left(\frac{1+r}{1-r}\right) \tag{147}$$

$$\implies e^{2r_z} = \frac{1+r}{1-r} \tag{148}$$

$$\implies r = \frac{e^{2r_z}-1}{e^{2r_z}+1} \tag{149}$$

The above is the Inverse Fisher Z transformation. We can now construct a confidence interval for our population correlation parameter.

$$\implies \text{CI for } \rho: \quad \left(\frac{e^{2\text{LB}_{\rho_z}}-1}{e^{2\text{LB}_{\rho_z}}+1}, \frac{e^{2\text{UB}_{\rho_z}}-1}{e^{2\text{UB}_{\rho_z}}+1}\right) \tag{150}$$

What do we know about confidence intervals? Well if we have constructed them well (effect size, sample size n, etc.) and the hypothesised population parameter is not contained in the interval **then we reject the null hypothesis**. So if we get a CI for ρ, and ρ_0 is not included in the interval, then we must reject $\rho = \rho_0$, and look for an alternative value for ρ.

$$\begin{array}{ccc} r & \Rightarrow & b_0 \text{ and } b_1 \\ \Downarrow & & \\ r_z & \Rightarrow & \text{CI for } \rho_z \\ \Downarrow & & \Downarrow \\ p\text{-value for } r_z & & \text{CI for } \rho \end{array} \tag{151}$$

What, when and why (assumptions)

The usual question that statisticians ask is, "what formula do I need to use for this?" and hopefully Tables 8 to 10 will help to clear that up.

Usually, we start by talking about some random variable y which has an approximately normal population distribution, i.e. $y \sim \mathcal{N}\left(\mu, \sigma^2\right)$. If the population of y is **not** distributed normally, the Central Limit Theorem allows us to conclude that the sampling distribution of the mean of y (many many samples) is! So even if the distribution of the population of y is skewed, or has no observable pattern, the CLT states that the mean of all possible samples (i.e. large n) has an observable pattern, namely normal. So if y has mean μ and standard deviation σ, then $\bar{y} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$ is the sampling distribution of the mean of y.

Table 8

Number of groups (I)	Assumptions	Name	Statistic	Confidence interval
1	Normality CLT μ known σ known independence	z-score standardised score	$z = \frac{y-\mu}{\sigma} \sim \mathcal{N}(0,1)$	
1	CLT μ known σ **unknown** independence	One-sample independent t-test for estimating population mean	$t = \frac{\bar{y}-\mu}{s/\sqrt{n}} \sim t(n-1)$	CI for μ: $\bar{y} \pm t^* \frac{s}{\sqrt{n}}$
2	Normality CLT independence μ_1, μ_2 unknown σ_1, σ_2 unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 \approx \sigma_2$ $H_0 : \mu = \mu_1 - \mu_2 = \mu_0 = 0$	Two-sample independent t-test for comparing means (assumed equal variance)	$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n-2)$	CI for $\mu = \mu_1 - \mu_2$: $(\bar{y}_1 - \bar{y}_2) \pm t^* s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
2	Normality CLT **dependence** μ_1, μ_2 unknown σ_1, σ_2 unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 \approx \sigma_2$ Construct a new variable $d_i = y_i - x_i$ (difference) from the dependent sample (x_i, y_i) Equal sample sizes $n_x = n = n_y$ $H_0 : \mu_d = \mu_0 = 0$	Two-sample **dependent** t-test for comparing means (**paired data**, e.g. before and after a treatment)	$\left.\begin{array}{l} \bar{d} = \frac{\sum_{i=1}^n (y_i - x_i)}{n} = \bar{y} - \bar{x} \\ s_d^2 = \frac{\sum_{i=1}^n d_i^2 - \bar{d}^2/n}{n-1} \end{array}\right\} \Longrightarrow t = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} \sim t(n-1)$	CI for $\mu_d = \mu_y - \mu_x$: $\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$
2	Normality CLT independence μ_1, μ_2 unknown σ_1, σ_2 unknown **Homogeneity of variances violated** $\sigma_1 \neq \sigma_2$ $H_0 : \mu = \mu_1 - \mu_2 = \mu_0 = 0$	Two-sample independent t-test for comparing means (assumed unequal variance)	$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t(k)$, where k is approximated by a computer	CI for $\mu = \mu_1 - \mu_2$: $(\bar{y}_1 - \bar{y}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Table 9

ANOVA I	
Number of groups	3+
Assumptions	Normality CLT independence μ_i, $i = 1, 2, \ldots, I$, unknown σ_i, $i = 1, 2, \ldots, I$, unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 = \sigma_2 = \cdots = \sigma_I$ $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ all of the groups are the same H_a: one of the groups is different from the rest
Name	ANOVA I "is the observed variance in the data attributable to the variation between the groups, or within the groups" Compare the means of 3 or more groups; if 2 groups, then t-test suffices
Statistic	$i = 1, 2, \ldots, I$ indexes the **group number** $j = 1, 2, \ldots, n_i$ indexes the **score** within a group $\mathrm{SST} = \mathrm{SSG} + \mathrm{SSE}$ $\mathrm{df}_T = \mathrm{df}_G + \mathrm{df}_E$ $\mathrm{MST} = \dfrac{\mathrm{SST}}{\mathrm{df}_T} = \mathrm{Var}\, y$ $\mathrm{SST} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ $\mathrm{df} = n - 1$ $\mathrm{MSG} = \dfrac{\mathrm{SSG}}{\mathrm{df}_G}$ $\mathrm{SSG} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2$ $\mathrm{df}_G = I - 1$ $\mathrm{MSE} = \dfrac{\mathrm{SSE}}{\mathrm{df}_E} = s_p^2$ $\mathrm{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{I} (n_i - 1)\, s_i^2$ $\mathrm{df}_E = n - I$ $\Longrightarrow F = \dfrac{MSG}{MSE} \sim F(\mathrm{df}_G, \mathrm{df}_E)$
Confidence interval	CI for mean of group i: (**homoskedasticity met**): $\bar{y}_i \pm t^*_{n-1} \frac{s_p}{\sqrt{n_i}}$ (**homoskedasticity violated**): $\bar{y}_i \pm t^*_{n_i-1} \frac{s_i}{\sqrt{n_i}}$

Table 10

ANOVA II	
Number of groups	3+ or 2+ factors
Assumptions	Normality CLT independence μ_i, $i = 1, 2, \ldots, I$, unknown σ_i, $i = 1, 2, \ldots, I$, unknown **Homoskedasticity**: Homogeneity of variances $\sigma_1 = \sigma_2 = \cdots = \sigma_I$ $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ all of the groups are the same H_a: one of the groups is different from the rest
Name	ANOVA II "is the observed variance in the data attributable to the variation between the groups (across factors), or within the groups" Compare the means of 3 or more groups, where group membership is defined by 2 factors, A and B; if 2 groups, then t-test suffices
Statistic	$i = 1, 2, \ldots, I$ indexes **factor** A $j = 1, 2, \ldots, J$ indexes **factor** B $k = 1, 2, \ldots, n$ indexes the individual score $\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$ $\text{df}_T = \text{df}_A + \text{df}_B + \text{df}_{A\times B} + \text{df}_E$ $\text{MST} = \dfrac{\text{SST}}{\text{df}_T} = \text{Var}\, y$ $\text{SST} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ijk} - \bar{y})^2$ $\text{df} = n - 1$ $\text{MSA} = \dfrac{\text{SSA}}{\text{df}_A}$ $\text{SSA} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} n \times J \times (\bar{y}_i - \bar{y})^2$ $\text{df}_A = I - 1$ $\text{MSB} = \dfrac{\text{SSB}}{\text{df}_B}$ $\text{SSB} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{J} n \times I \times (\bar{y}_j - \bar{y})^2$ $\text{df}_B = J - 1$ $\text{MSAB} = \dfrac{\text{SSAB}}{\text{df}_{A\times B}}$ $\text{SSAB} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$ $\text{df}_{A\times B} = (I - 1) \times (J - 1)$ $\text{MSE} = \dfrac{\text{SSE}}{\text{df}_E} = s_p^2$ $\text{SSE} = \sum_{k=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ijk} - \bar{y})^2$ $\text{df}_E = n - I \times J$ $\Longrightarrow \begin{cases} F_A = \dfrac{MSA}{MSE} \sim F(\text{df}_A, \text{df}_E) \\ F_B = \dfrac{MSB}{MSE} \sim F(\text{df}_B, \text{df}_E) \\ F_{A\times B} = \dfrac{MSAB}{MSE} \sim F(\text{df}_{A\times B}, \text{df}_E) \end{cases}$
Confidence interval	CI_{ij} for comparing the means of groups i and j: $(\bar{y}_i - \bar{y}_j) \pm t^{**}_{\text{df}_E} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ (See the section on post-hoc methods)

Exercises

Regression - ANOVA analysis

1. The "Healthy Breakfast" dataset contains, among other variables, the Consumer Reports ratings of 77 cereals, the number of grams of sugar contained in each serving, and the number of grams of fat contained in each serving.

   Considering "Sugars" as the explanatory variable and "Rating" as the response variable generated the following regression line:

   Rating = 59.3 - 2.40 Sugars

Source	DF	SS	MS	F	p
Regression	1	8654.7	8654.7	102.35	0.000
Error	75	6342.1	84.6		
Total	76	14996.8	194.76		

Table 11: Analysis of Variance - rating ~ sugar

   As a simple linear regression model, we previously considered "Sugars" as the explanatory variable and "Rating" as the response variable.

   The regression line generated by the inclusion of "Sugars" and "Fat" is the following:

   Rating = 61.1 - 2.21 Sugars - 3.07 Fat

Source	DF	SS	MS	F	p
Regression	2	9325.3	4662.6	60.84	0.000
Error	74	5671.5	76.6		
Total	76	14996.8	194.76		
Source	DF	Seq SS			
Sugars	1	8654.7			
Fat	1	670.5			

Table 12: Analysis of Variance - rating ~ sugar + fat

   (a) Define the population regression model using table 12. If two cereals have the same fat content but different sugar content, what can you say about the rating?

   (b) What does VIF stand for? Compute the VIF using table 12.

   (c) What does VAF stand for? Compute the VAF using tables 11 and 12.

   (d) How do the ANOVA results change when "FAT" is added as a second explanatory variable?

   (e) Formulate appropriate hypotheses, make a decision and explain your reasoning.

   Solution.

   (a) The population regression model used in table 12 is rating = β_0 + β_1 sugars + β_2 fat + ε, where β_j's are approximated by b_j's such that $\vec{b} = (61.1, -2.21, -3.07)^T$. If variable fat is kept constant, then the marginal difference in rating is -2.21 per gram of sugar. This says that the rating of the breakfast cereal will decrease by 2.21 points per additional gram of sugar, under the condition that fat content is kept constant.

   (b) VIF stands for variance inflation factor, and is given by $\text{VIF}_j = 1/\left(1 - R_j^2\right)$, where R_j^2 is the coefficient of determination of the regression equation $X_j = \alpha_0 + \alpha_1 X_{-j} + \delta$ (regress the explanatory variables on the others). The square root of the VIF indicates how much larger the standard error increases compared to if that variable had 0 correlation to other predictor variables in the model. For example, if the variance inflation factor of a predictor variable were 5.27 ($\sqrt{5.27} = 2.3$), this means that the standard error for the coefficient of that predictor variable is 2.3 times larger than if that predictor variable had 0 correlation with the other predictor variables. Rule of thumb: $VIF_j > 10$ indicates multicollinearity in the model, i.e. explanatory variables are dependent on each other.

   It is not possible to compute the VIF using table 12, as we need the partial SS information.

   (c) VAF stands from variance accounted for and is given by the R^2 coefficient for linear regression, where $R_2 = SSR/SST$. From table 11, R^2 = 8654.7/14996.8 = 0.577, and from table 12, R^2 = 9325.3/14996.8 = 0.622.

(d) Comparing the VAF's tells us that the model is improved with the addition of fat to the model. This is further shown by the column Seq SS, which shows that fat reduces the SSE by 670.5, which in turn reduces the MSE, indicating less deviation between the observed and fitted values.

(e) $H_0 : \beta_2 = 0$ and $H_A : \beta_2 \neq 0$. F is significant with $p < 0.05$, i.e. reject H_0 in favour of H_A and conclude that fat is in the population model.

2. Answer the following questions using the tables and graphs below.

Table 13: Descriptive Statistics

	sales	adverts	airplay	attract
Valid	200	200	200	200
Missing	0	0	0	0
Mean	193.200	614.412	27.500	6.770
Std. Error of Mean	5.706	34.341	0.868	0.099
Std. Deviation	80.699	485.655	12.270	1.395
Minimum	10.000	9.104	0.000	1.000
Maximum	360.000	2271.860	63.000	10.000

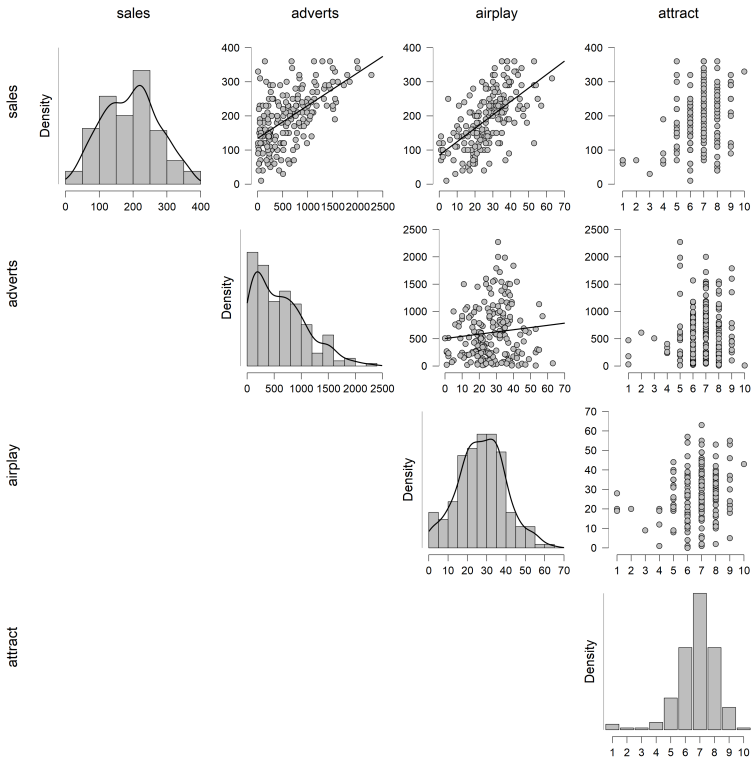

Figure 3

Table 14: Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p
0	0.578	0.335	0.331	65.991	0.335	99.587	1	198	< .001
1	0.815	0.665	0.660	47.087	0.330	96.447	2	196	< .001

Table 15: Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI Lower	95% CI Upper	Collinearity Statistics Tolerance	Collinearity Statistics VIF
0	(Intercept)	134.140	7.537		17.799	< .001	119.278	149.002		
	adverts	0.096	0.010	0.578	9.979	< .001	0.077	0.115	1.000	1.000
1	(Intercept)	-26.613	17.350		-1.534	0.127	-60.830	7.604		
	adverts	0.085	0.007	0.511	12.261	< .001	0.071	0.099	0.986	1.015
	airplay	3.367	0.278	0.512	12.123	< .001	2.820	3.915	0.959	1.043
	attract	11.086	2.438	0.192	4.548	< .001	6.279	15.894	0.963	1.038

Table 16: ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	433687.833	1	433687.833	99.587	< .001
	Residual	862264.167	198	4354.870		
	Total	1.296e+6	199			
1	Regression	861377.418	3	287125.806	129.498	< .001
	Residual	434574.582	196	2217.217		
	Total	1.296e+6	199			

(a) What is the population regression equations for Model 0 and 1?

(b) Describe the regression equations you wrote above, in words (1-2 sentences each model). What is the point of the comparison?

(c) Summarise the findings of table 13 and compare with fig. 3.

(d) Write the null and alternative hypothesis based on the regression equations you wrote in part (a). Now, describe these hypothese in words (do not refer to beta coefficients, just use plain language - like you're informing a friend). What does table 14 inform you about your hypotheses?

(e) Under Model 0, what is the expected output if the explanatory variable input has value 600? Compare this with output with the output from Model 1 under the same conditions. Explain the difference in your results.

(f) Explain the fourth and seventh columns of table 14.

(g) Table 15 provides you with the VIF for both models. Interpret the results without making too many references to the exact value of the VIF, i.e. what do these values mean?

(h) Provide the standardised regression equations for both models.

(i) Use table 16 to make your decision about your hypotheses. Explain your reasoning.

Solution.

(a) Model 0: sales $= \beta_0 + \beta_1$ adverts $+ \varepsilon$.

Model 1: sales $= \beta_0 + \beta_1$ adverts $+ \beta_2$ airplay $+ \beta_3$ attract $+ \varepsilon$.

(b) In Model 0, we propose that the album sales depends only only the number of advertisements for the given album.

In Model 1, we propose that the album sales depends not only on the number of advertisements, but also on the radio airplay time