Pirates, Peaches and P-values

SOLUTIONS TO THE EXERCISES

VERSION DATE: 28-03-2018

CHAPTER 1 MAPPING DATA (COMPLETE)

1. A

The distribution is skewed to the right; the high score on the right (an outlier maybe?) pulls up the mean, but not the median, which is a bit closer to the peak.

2. A

The trick is to find the quartiles… have you found them?
Since 25% of the dreams lasted 10 minutes or shorter, 10 makes the first quartile bound: $Q_1 = 10$. The third quartile bound is thus formed by the number 25: $Q_3 = 25$. There you go! This means that

$$IQR = Q_3 - Q_1 = 25 - 10 = 15$$
$$1{,}5 * IQR = 1{,}5 * 15 = 22{,}5$$
$$Q_3 + 1{,}5 * IQR = 25 + 22{,}5 = 47{,}5$$

These 47,5 minutes are the upper bound for 'common' dream durations. Because 60 minutes is well above 47,5, we can consider this dream an outlier.

3. C

We can only speak of ratios – as in answer A – if we have a ratio variable; the numbers on the Likert scale do not have a quantitative meaning! Also the distances (intervals) between consecutive numbers are not necessarily equally large. As a result B drops out as well: we're not sure if the difference between 1 and 2 is as large as the difference between 2 and 3. Answer C does apply to an ordinal scale.

4.

a) Sure, a mode is always a go. This is 2.
b) A median is a go from an ordinal variable onward, so bring it on. Count the bar heights: we have 16 scores in total. The median is therefore score number $\frac{N+1}{2} = \frac{17}{2} = 8{,}5$ – in other words, the mean between the eighth and ninth score. Back to the bar chart: the eighth score is a 2 and the ninth a 3. That makes the median 2,5.
c) Rather not: the Likert scale isn't quantitative!
d) Nope. You'd need the mean and you'd have to do a calculation, which doesn't make sense for categorical variables.
e) Nay. The IQR requires a calculation again. Were you thinking of subtracting 3 (partly illogical, partly logical) minus 2 (rather illogical)? And what would the outcome mean? The 50% middle scores are at most '1 logic apart'? That's bollocks. ☹

5.

a) Yeah, could be. It's how this variable behaves in my personal experience: sometimes I meet almost no one, sometimes about five individuals, and quite rarely dozens of people present themselves. This may or may not be the case for other participants as well.
b) Why not? Possibly your dreams are the contrary of mine, dear reader: they may be very crowded, and you're rarely lonely.
c) The distribution is symmetric when participants have usually got, say, about three people around them in their dreams, sometimes less, sometimes more. Not unthinkable at all.
d) What this exercise has had you do is think about the content of statistical information. You have hopefully discovered that we're quite able to make up theories beforehand. This will help us to understand what statistics tell us about the real world – in this case, what the shape of a distribution says. So should you theorise? Yes! Moving back and forth between reality and math is what this book is all about, and anything that supports the process is a good thing. ☺

6.

a) Let's see. The variation indicates how people deviate from the mean. We calculate it by subtracting the mean from all the individual scores, squaring the differences, and then summing all these squares:

$$\sum (X_i - \bar{X})^2$$

However, this new person has experienced his dream once before. He or she thus scores equal to the mean, and as such doesn't deviate from the mean. We will add $\mathbf{0^2 = 0}$ to the current variation.

Try to imagine what this means content-wise as well: this person doesn't vary from the mean, so he or she won't add anything to the extent to which the participants vary together.

b) This is the average variation per person. Since this single person doesn't vary (deviate) from $\bar{X}$, the average deviation will go down. And indeed, to acquire the variance, we divide the variation by a slightly larger number now:

$$s_X^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{N-1}}$$

No longer by $\mathbf{16 - 1 = 15}$, but by $\mathbf{17 - 1 = 16}$. That makes the variance a little smaller.

c) As a result the standard deviation, the square root of the variance, becomes a little smaller as well:

$$s_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{\boldsymbol{N-1}}}$$

7. D

Nout and Emma are both 0,76 standard deviations away from the mean; Emma is below, while Nout is above.

CHAPTER 2 CATEGORICAL RELATIONSHIPS (COMPLETE)

1. C

By looking merely at the univariate distributions, we cannot see how variables are related, precisely because the contingency table can still be filled out in several ways at that stage. Take this situation for instance:

		NUMBER OF ROUTES		
		one	two	
MASSACRED	no	39 (81,25%)	13 (81,25%)	**52 (81,25%)**
	yes	9 (18,75%)	3 (18,75%)	**12 (18,75%)**
		48 (100%)	**16 (100%)**	**64 (100%)**

It gives us no association.
Or this one:

		NUMBER OF ROUTES		
		one	two	
MASSACRED	no	48 (100%)	4 (25%)	**52 (81,25%)**
	yes	0 (0%)	12 (75%)	**12 (18,75%)**
		48 (100%)	**16 (100%)**	**64 (100%)**

Here we have a very strong association.

2. B

Answer A is no fair comparison, because there were many more people who had only one choice to begin with. In that case, it's actually curious that there were as many massacres among them (6) as among the two-route participants. What matters is if both route conditions showed a relatively equal massacre size. So we're interested in the percentages:

		NUMBER OF ROUTES		
		one	two	
MASSACRED	no	42 (87,5%)	10 (62,5%)	**52 (81,25%)**
	yes	6 (12,5%)	6 (37,5%)	**12 (18,75%)**
		48 (100%)	**16 (100%)**	**64 (100%)**

Now we see that relatively more people were killed when they had two routes instead of one. B is correct. C isn't: the ones who escaped did always constitute the majority, but the percentage was 87,5% in the single-route group, and a lower 62,5% in the two-route group. Thus the death percentage still does change depending on the route condition.

3.

I already calculated the column percentages in exercise 2. The row percentages are:

		NUMBER OF ROUTES		
		one	two	
MASSACRED	no	42 (80,77%)	10 (19,23%)	**52 (100%)**
	yes	6 (50%)	6 (50%)	**12 (100%)**
		48 (100%)	**16 (100%)**	**64 (100%)**

The row percentages indicate: 'Among those who were not or were massacred, this many percent had one or two routes.'
The column percentages indicate: 'Among those who had one or two routes, this many percent were not or were massacred.'

Both percentages are informative in their own way. Nevertheless, I slightly prefer percentages that express the causal relationship best. I think the number of routes may influence the chance to get killed, not the other way around. This causal direction is best described by the column percentages.

4.

a) $RD = p_1 - p_0 = 0{,}375 - 0{,}125 = 0{,}25$

b) $RR = \frac{p_1}{p_0} = \frac{0{,}375}{0{,}125} = 3$

c) The risk difference tells us that hesitating in front of two possible routes leads to a 25% increase of the massacres. The relative risk, additionally, tells us that this increase makes the risk of being massacred 3 times as large. I'd say that both statistics are informative in their own right, so I see no problem in reporting them both. Hooper's study isn't retrospective, so no one will complain that we should've used an odds ratio instead.

d) We want to see how the total population may have fared if the risk factor (having two choices) was removed, so we should determine the attributable risk for the total (population attributable risk).

$$AR_t = \frac{p_t - p_0}{p_t} = \frac{0{,}1875 - 0{,}1250}{0{,}1875} = 0{,}3333$$

It seems that the massacre would have gone down in size by one third! This underlines the effectiveness of hesitation pretty boldly. Good to know, dear reader, in case you ever decide to become a sadistic serial killer.

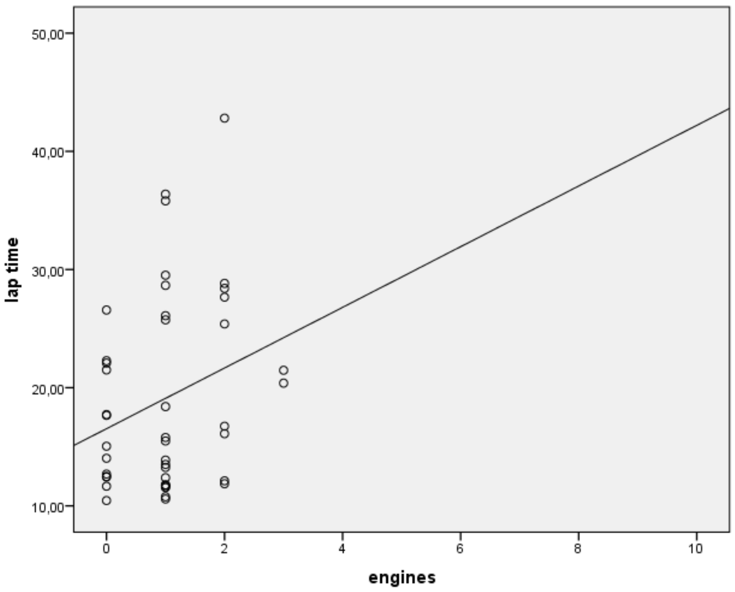
5. C

This kind of statistics is misleading (and dangerous as a result): people who escaped were extremely afraid in large numbers… but who's to say that this didn't apply to people who had their heads sawn off? Until we put these people next to those who got out of the house, we won't be able to see if the escapees were more afraid than the fatal victims. Now that we're only looking at the individuals who got past the exit, the variable MASSACRED doesn't vary; however, relationships can only exist between variables.

CHAPTER 3 QUANTITATIVE RELATIONSHIPS (COMPLETE)

1. A

Can you install half an engine? You simply can't. Nevertheless the number of engines used remains a ratio variable – it certainly isn't ordinal. Are we seeing four subgroups then? Heck no: there should be a third variable which creates these subgroups, but in fact each 'stack' of scores just consists of drivers with a certain natural number of engines on their soapbox (*X*). For the fun of it, have a look below here: this is what the scatterplot would've looked like if the X axis extended to 10 engines.



Doesn't this look like a perfectly normal scatterplot? ☺

2.

The more engines…: TRUE. We can say that there's a positive trend.

The regression line…: FALSE. The correlation coefficient only tells us how strong the relationship is, so, how well the points follow the regression line. It says nothing about the slope of the line (except that it's positive).

Engines have an…: FALSE. The predicted lap time increases if the driver uses more engines… Is that advantageous? No! A lap around the course takes longer! The director had expected a decreasing trend though. The introductory text mentions that the course is extremely winding, which may explain these results. It seems that drivers with engines lose control of their vehicle more easily, and fly off track for example.

Beside the number…: TRUE. The correlation doesn't equal 1, so the number of engines isn't the only thing which influences the lap time. Makes sense if you ask me. Aside from other properties of the soapbox (traction, agility, brakes) the driver's skill is likely to play a large role as well.

3. B

If necessary, refer to the theory on restriction of range (final paragraph). This is the opposite situation: we've currently studied a rather small range of engine numbers, which suppresses the correlation. Possibly this correlation would grow if drivers with even more engines had joined as well.

We're not sure of this by the way. Would the linear trend continue or not? It's also thinkable that yet another extra engine won't make a difference for the lap time at a certain point (and that the line will flatten as a result). Extend the regression line into an area that wasn't measured, and you're **extrapolating**. Extrapolation is always a bit uncertain.

4.

a)

$$b = r_{XY} * \frac{s_Y}{s_X} = 0{,}268 * \frac{8{,}107}{0{,}845} = 2{,}57$$
$$a = \bar{Y} - b\bar{X} = 19{,}10 - 2{,}57 * 1 = 16{,}53$$

In short,

$$\hat{Y} = 16{,}53 + 2{,}57X$$

b) The predicted lap time for a soapbox <u>without engines</u> ($X = 0$); this is 16,53 minutes.

c) The predicted increase of the lap time per extra engine; this is 2,57 minutes.

d) Neither: to determine how <u>strongly</u> the lap time depends on the engine quantity, we'll need the <u>correlation</u>.

5.

In my opinion it does: I don't see another remaining association (such as a curvilinear one).

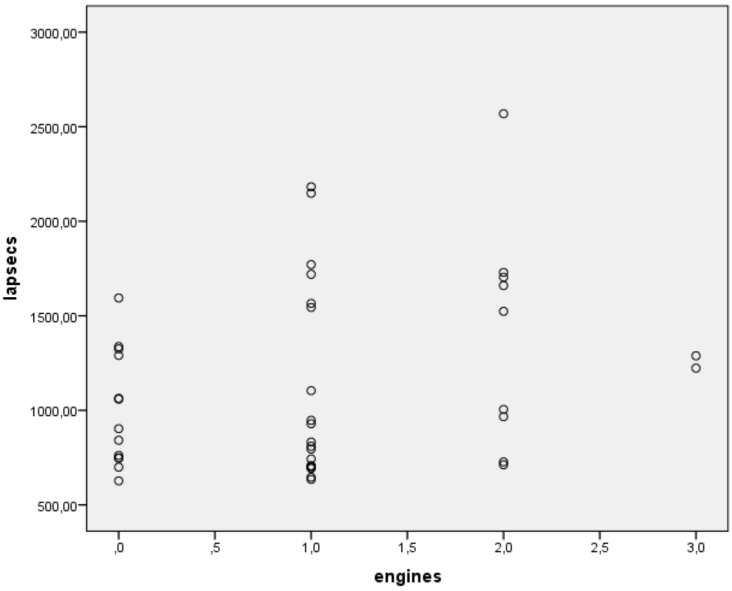
6.

$$R^2 = 0{,}268 = 0{,}072$$

In other words, the variation in the lap times can be partly explained (to an extent of 7,2%) because not every soapbox had an equal number of engines. That's not a lot: 92,8% of the variation must be attributed to other factors, such as the fact that the drivers differ in talent. The question is therefore whether the director should attribute the race results to the engines. If not, the use of engines won't need to become obligatory next year.

7. B

The director has <u>transformed</u> the lap time: from minutes to seconds. We're having a <u>multiplication</u> here, for all the time scores become 60 times as large. Does this mean that the scores move closer together or further apart? No, because they all change to exactly the same extent!



So remember: when we transform variables linearly, their correlation will stay exactly the same.

8. D

No association means a flat regression line ($b = 0$) and no correlation ($r = 0$), so a proportion of explained variation equal to nothing as well ($R^2 = 0$). The regression line now simply stands for the average line. If necessary, see the image at the end of paragraph 3.4, below the heading *Special situations*.

9. C

The two variables mentioned are <u>categorical</u>! We can only calculate a correlation between quantitative variables.

CHAPTER 4 PROBABILITY THEORY (COMPLETE)

1.

a)

		BRIBERY		
		no	yes	
PRIZE	stuffed bear	12	12	**24**
	piece of junk	228	18	**246**
		240	**30**	**270**

b) A matter of reading comprehension. We're looking for

$$P(bribery\ and\ bear) = \frac{12}{270} = 0{,}044$$

c) This is not the same as in b! Now it's about a conditional probability: the probability of bribery, <u>if</u> the child won a stuffed bear. Write down the requested probability correctly and you'll be fine:

$$P(bribery|bear) = \frac{12}{24} = 0{,}5$$

d) No, certainly not: after all there are 12 children with both a stuffed bear and bribing parents.

e) Non-disjoint events are not necessarily independent! So let's see. We have statistical independence if

$$P(A) = P(A|B)$$

Do the data satisfy this condition?

$$P(bribery) = \frac{30}{270} = 0{,}111$$

However,

$$P(bribery|bear) = \frac{12}{24} = 0{,}5$$

… like we already calculated in exercise b. So the probability of bribing parents rises fiercely if we've got a child that won a stuffed bear. The event 'bribery' is strongly dependent on the event 'stuffed bear'.

Another good solution:

$$P(bear) = \frac{24}{270} = 0{,}089$$

However,

$$P(bear|bribery) = \frac{12}{30} = 0{,}4$$

Yes, this last probability was already in the introduction… in other words, the introduction already told us that the probability of a stuffed bear is partially dependent on the parents bribing or not bribing the owner. Not many students notice this… so if you did, good job! ☺

2.

a) Simple.

$$P(junk\ and\ junk) = \frac{246}{270} * \frac{246}{270} = 0{,}830$$

b) A little bit less simple: keep in mind that two orders are possible (see the final theory paragraph)! Either the first kid wins a stuffed bear and the second one a piece of junk, <u>or the other way around</u>.

$$P(bear\ and\ junk) = \frac{24}{270} * \frac{246}{270} * 2 = 0{,}162$$

c) Simple once again.

$$P(bear\ and\ bear) = \frac{24}{270} * \frac{24}{270} = 0{,}008$$

d) This is an event: a set of multiple elementary outcomes. You can draw all kinds of samples of two children (one sample constitutes one EO) who both won a piece of junk! We have just calculated that this will happen in 83,0% of all the EOs.

By the way, the probability distribution is correct, since the sum of the three probabilities equals 1. ☺

e) That's going to be difficult… because the two events will no longer be independent! Suppose that the boy wins a stuffed bear? Then his parents may be more inclined to bribe the owner, so his little sister will have a higher chance of winning a bear as well. We're not sure what the parents will do though. Therefore we can't determine the true probability that the girl will still win a piece of junk.

For this reason, almost all statistical techniques in *Pirates, Peaches and P-values* assume that all participants are independent. Keep this in mind if you ever test participants yourself!

3. C

A is the population; B is a single elementary outcome. C contains all possible elementary outcomes; after all, an EO will always be the complete sample if we draw random samples.

4. B

To repeat the definition: a random variable is a measure that will randomly take on different values if the probability experiment is repeated.

The sample size is fixed – we determine it ourselves – and isn't a result of randomness because of this. If we change the size, we'll also change the probability experiment.

The number of parents who bribed the owner on this day (namely 30) is a **population** value (an appetiser to chapter 5): this number is fixed as well and won't change, no matter how often we keep drawing samples of 10 kids from all the kids who pulled a rope today.

However, the number of kids in the sample who won a stuffed bear will be different each time we repeat the probability experiment. One time we will draw 0 of these lucky devils, next time we'll draw 1, sometimes even 2, and so on.

5.

a) The expected value(!) is

$$E(X) = N * p = 20 * 0{,}17 = 3{,}4$$

So on average we'll draw 3,4 puking passengers in a sample of 20 people.

b) We can calculate the variance using the formula

$$s_X^2 = N * p * (1 - p) = 20 * 0{,}17 * 0{,}83 = 2{,}82$$

Which makes the standard deviation

$$s_X = \sqrt{2{,}82} = 1{,}68$$

Or, the number of vomiting passengers in the sample will deviate from the expected value by 1,68 on average. This is rather much, so the samples we draw will often give a lopsided impression of the population.

6. C

'First collect information,' would be my advice. There's enough of it in the exercise, but we need to write things down a bit more accessibly. In any case the probability of a vomiting person is

$$P(vomit) = 0{,}17$$

And the other percentages, what kind of probabilities are they? Make sure to write them down correctly!

$$P(adult|vomit) = 0{,}70$$

$$P(child\ or\ adolescent|vomit) = 0{,}30$$

After all, if the person is amongst the vomiting passengers (condition), there's a certain chance that he or she is an adult or not.

Now, what do these bits of information tell us? Is the probability of a child equal to 0,3? No: that's only the case if the person in question had to throw up. We don't now if the general probability of a child is the same.

Does $P(vomit|adult)$ equal 0,7 then? Nope, wrong again: whoever chooses answer B must have interpreted that 70% from the introduction 'backwards'.

Can we perhaps derive the probability of an adult passenger who also has to throw up? Using the product rule by chance?

$$P(adult\ and\ vomit) = P(adult) * P(vomit|adult)$$

No, not like this…

$$\boldsymbol{P(vomit\ and\ adult)} = P(vomit) * P(adult|vomit) =$$
$$0{,}17 * 0{,}70 = \mathbf{0{,}119}$$

… ah! Like this! ☺

CHAPTER 5 PROBABILITY DISTRIBUTIONS (COMPLETE)

1.

a) The 68-95-99,7 rule of thumb tells us that 95% of the sausage slices differ at most 2 standard deviations from the mean.

$$11 - 2 * 4{,}5 = 2$$
$$11 + 2 * 4{,}5 = 20$$

Answer: 95% of the slices are between 2 and 20 millimetres thick.

b) Beside the slices of sausage in exercise a, 5% remain. A normal distribution is symmetric, so that 5% is distributed evenly across the left and right tail of the distribution. (Sketch this!) 2,5% of the slices are thinner than 2 millimetres, and 2,5% are thicker than 20 millimetres. In short, the 2,5% thickest slices are 20 millimetres or thicker.

c) This question is beyond the rule of thumb. We'll have to calculate a z-score. This is simply a z-score concerning the <u>population distribution</u>:

$$Z = \frac{X - \mu}{\sigma} = \frac{15 - 11}{4{,}5} = 0{,}89$$

If necessary, <u>draw</u> the distribution and indicate the desired probability! The z-table tells us that

$$P(Z > 0{,}89) = 1 - 0{,}8133 = 0{,}1867$$

So for 18,67% of his slices, Luca risks getting a scolding. Whew, it'd make me rather nervous…

2. A

This percentage is an estimate: a <u>sample value</u>, obtained from a random sample. It's an estimate of the population parameter, the percentage of *all* slices which have been cut too thick. The sample is of course the 5 picked slices themselves, not their thickness.

Population: ALL SLICES OF CHORIZO	**Sample:** THE 5 MEASURED SLICES OF CHORIZO
Parameter: PERCENTAGE OF OVERLY THICK SLICES IN THE POPULATION (40%?)	**Estimate:** PERCENTAGE OF OVERLY THICK SLICES IN THE SAMPLE (40%)

3. C

If necessary, look up the definitions of bias and efficiency again. The sample means are unbiased, because they come from a random sample; if we determined a new mean infinitely often, we'd end up with 11 millimetres on average (μ). However, the two sample means of 8,4 and 13,7 millimetres don't really hit the bull's-eye; they lack accurary. I hope you already felt this coming in exercise 2 – I mean, base a percentage on 5 slices? We can thus call the means inefficient.

4.

a) This sample is too small ($N = 7$); the distribution's shape and the estimates aren't very reliable (efficient). We'd better not make any statements about the population distribution and the sampling distribution of the mean.

b) This sample is pretty large ($N = 70$); the shape of the distribution of sample scores probably approaches the population distribution, which is therefore likely to be symmetric and unimodal[1] as well. It's plausible that the population mean μ_X is approximately equal to 6 centimetres, and the standard deviation σ_X equal to 1,5 centimetres. The sampling distribution of the mean is approximately normal, just like the population distribution. Its expected value equals the population mean, namely about 6, and the standard error is $\sigma_{\bar{X}} = {}^{\sigma_X}/_{\sqrt{N}} \approx {}^{1{,}5}/_{\sqrt{70}} = 0{,}18$. We can see from this standard error how efficient the sample mean is: on average is will deviate from the population mean by a mere 0,18 centimetres.

c) This sample is large as well ($N = 70$); large enough to make statements about the population distribution and the sampling distribution of the mean. Once again the distribution of sample scores probably approaches the population distribution, which is likely to be unimodal and skewed to the left. The population mean μ_X will be about 2 and the standard deviation σ_X about 0,5. However, the sampling distribution isn't skewed like the population distribution, but approximately normal thanks to the large N (central limit theorem)! Expected value and standard error: $\mu_{\bar{X}} \approx 2$, $\sigma_{\bar{X}} \approx {}^{0{,}5}/_{\sqrt{70}} = 0{,}060$. Very tiny standard error! ☺

d) Only the <u>size</u> of the sample matters. The sampling distribution of $\bar{X}$ describes which $\bar{X}$s Nick would get how often, if he drew an infinite amount of samples. This sampling distribution is therefore a theoretical idea: this single sample constitutes one of the results Nick <u>could have obtained</u>. All these possible results are approximately normally distributed due to the large N.

[1] Which isn't necessarily the same as 'normal', but for now I don't mind if you always consider a symmetric distribution as normal. We'll start making some extra nuances in chapter 10. ☺

e) It looks like that at the moment: the average water level has receded from 6 to 2 centimetres, according to Nick's estimations. We can't be sure though, since they remain sample estimates. Can we then never exclude the possibility that this decrease is a pure coincidence? Well… from chapter 6 onward, we're going to give it a try!

5.

a) 1,6 grams: it's about the standard deviation of the distribution of sample scores. The standard deviation of a large sample will probably look much like the population's standard deviation.

b) The keyword is 'average': it's about the probability of a certain sample mean. In other words, that's a probability under the sampling distribution. Sketch it!

The z-score is

$$Z = \frac{32 - 33}{1{,}6/\sqrt{30}} = -3{,}42$$

It follows from the z-table that

$$P(Z < -3{,}42) = 0{,}0003$$

In short, the probability that we will underestimate the average weight of the population of stroopwafels by even a single gram is really low! A sample of 30 waffles is quite reliable.

c) That's right, but we calculate the probability of a sample mean of 32 or lower under the sampling distribution – and that one *is* approximately normal due to the large N (30), in spite of the skewed population distribution.

6. B

Sketch the z-distribution: according to the exercise we may assume that it's symmetric, and the mean is 0 in any case. The z-score of Luca's waffle will then of course lie to the left of the mean, or lower. This waffle is lighter than average.[2] More than 50% of the scores lie to the right of Luca's z-score, so more than half of the stroopwafels (15) are heavier than Luca's. Should we wish to look up how many waffles were heaver, we can do that in the z-table (as long as we consider the distribution to be normal): $P(Z < -0{,}15) = 0{,}4404$. So, Luca's waffle is roughly among the lightest 44%, but definitely not among the lightest 15%! Shame, now he has no excuse to take another one…

[2] Also holds for a skewed distribution.

CHAPTER 6 HYPOTHESIS TESTING (COMPLETE)

Exercises for peaches

1. B

The null hypothesis always needs to contain an equal sign (=). It makes a specific statement about (in this case) what the population mean may be, and we will need to compare our sample result against that fixed value. You can't compare your sample result against a range of possible values for the population mean.
Apparently, Snape attempted to prove that the average person inside the quiet compartment would be more silent (although he didn't expect that this would be the case). This makes this left-sided alternative hypothesis defensible.

2.

The distribution of sample scores contains an outlier – an extremely noisy person who seems to have practised his drum kit aboard the train or something. It's likely that he or she distorts the sample mean. I wouldn't advise a z-test as things stand now.

3. C

The formula for the confidence interval is $\bar{X} \pm Z^* {}^{\sigma}/_{\sqrt{N}}$. Filling it in gives us:

$$39{,}5 - 1{,}96 * {}^{12}/_{\sqrt{36}} = 39{,}5 - 1{,}96 * 2 = 39{,}5 - 3{,}92$$
$$39{,}5 + 1{,}96 * {}^{12}/_{\sqrt{36}} = 39{,}5 + 1{,}96 * 2 = 39{,}5 + 3{,}92$$
$$[\, 35{,}58\, ;\, 43{,}42\,]$$

This result should be interpreted like answer C says. A and B are misunderstandings; the components of the formula pertain to the <u>sampling distribution</u> of $\bar{X}$ (look at the standard error for instance). A talks about the population distribution, while B talks about the distribution of sample scores. This last one can actually be seen right above: it's the histogram! And it clearly shows that there are many scores below 30 and above 50 – easily more than 5% of the sample. With thanks to Marie for this keen observation. ☺
If you're confused about the logic behind the confidence interval, reread the theoretical discussion in chapter 6.

4. B

If we perform the test, we obtain the following:

$$Z = \frac{39{,}5 - 40}{{}^{12}/_{\sqrt{36}}} = \frac{-0{,}5}{2} = -0{,}25$$

Snape went looking for a <u>left-tailed</u> p-value. Note that you'll need the correct answer to exercise 1 in order to determine this bit. On an exam, questions are never allowed to be dependent like this, so no worries.

$$p = P(Z < -0{,}25) = 0{,}4013$$

So apparently, the challenge in the question is to interpret what the p-value says. B is the only correct interpretation. We cannot assign a probability to the null or alternative hypothesis itself; we can only investigate how likely Snape's sample result would be, <u>if</u> the null hypothesis was true. Since a sample like Snape's would be quite regular if the population mean equalled 40 decibels indeed, we have no reason to doubt the null hypothesis.

5. C

The only knowledge we can gain from the confidence interval is that the population mean probably lies within its borders. All values within the confidence interval are equally likely, due to the logic on which it was based; see the theoretical explanation if you want.

6.

a) A Type I error is a false rejection of a null hypothesis that is actually true. Since Rodent did not reject the null hypothesis, such an error wasn't possible here.

b) A Type II error is a failure to reject a null hypothesis that is actually false. This is a possibility. The p-value is not super-high, so perhaps the test should've been significant but wasn't.

c) The power is the probability that a false null hypothesis will be rejected. If the null hypothesis is untrue, greater power might have rendered the test significant after all. However, power as a concept is 'not applicable' in the (theoretical) situation of a true null hypothesis. Short answer to the question: no.

Exercises for pirates

1.

a) $Z = \frac{4{,}4-5}{{}^{3}/_{\sqrt{22}}} = -0{,}94$

Since he should look for the <u>right-tailed</u> p-value, this yields that

$$p = P(Z > -0{,}94) = 1 - P(Z < -0{,}94) = 1 - 0{,}1736 = 0{,}8267$$

Which is totally insignificant…

b) No! You need to state the hypotheses <u>before</u> drawing the sample, otherwise you can adapt your alternative hypothesis to the sample mean you found. That's cheating. Unfortunately the producer has no choice but to draw a new sample if the wishes to perform the same test at this point.

c) $Z = \frac{6,4-5}{3/\sqrt{22}} = 2,19$

Now we should take care to <u>double</u> the p-value from the z-table. It's a two-tailed test!

$$p = 2 * P(Z > 2,19) = 2 * 0,0143 = 0,0286$$

Still significant though. ☺

d) Not entirely: this sample isn't random anymore. The 22 testers are somewhat experienced from their first try. As such, the mean of this sample is a **biased** estimator of the population mean! With thanks to Paula for this brilliant remark. ☺

2. A

The sample mean is the centre of the confidence interval; this sample mean will change with each new sample, so the centre of the confidence interval will change along. The margin of error, however, consists of the critical z-value and the standard error; these are both constant. As a result the margin of error won't change.

3.

a) He could decrease the confidence level; then the critical z-value will go down and the width will lessen. However, in that case he is less confident that the population mean is in his interval. Not such a good idea after all. Better idea: increase the sample size! Then the standard error will go down and the width will decrease.

b) He'd have to up the critical z-value, which makes the confidence level increase. The consequence is that the interval becomes broader though, so that the producer can say with somewhat less accuracy what the population mean will be.

4.

a) $H_0: \mu = 32$
$H_A: \mu > 32$

b) This question requires a bit of recall from chapter 5. The hypothesis test can only be performed if the population distribution of the stroopwafel weights is <u>normal</u>. The sampling distribution needs to be normal after all, because the probabilities we look up in the z-table are based on a normal distribution. The central limit theorem doesn't hold, since the sample is pretty small. Therefore the sampling distribution is only still normal if the population distribution is such as well. (Actually that kite won't get off the ground: have a look back at exercise 5 from chapter 5.)

c) Tip: sketch the sampling distribution!

Cohen's d is estimated to be $\hat{d} = \frac{32,8-32}{1,6} = 0,50$. That's a medium effect, according to the guideline.

The z-score is

$$Z = \frac{32,8-32}{1,6/\sqrt{10}} = 1,58$$

It follows from the z-table that

$$p = P(Z > 1,58) = 1 - P(Z < 1,58) = 1 - 0,9429 = 0,0571$$

So, the test result is *just* not significant. This is probably due to the sample being too small: the test will then have little power. (You'll need to calculate the power yourself later on.) A slightly larger sample – with a smaller standard error – would probably have delivered a significant result.

d) The formula for the confidence interval is $\bar{X} \pm Z^* \sigma/\sqrt{N}$. Filling it in gives us:

$$32,8 - 1,645 * 1,6/\sqrt{10} = 32,8 - 0,83$$
$$32,8 + 1,645 * 1,6/\sqrt{10} = 32,8 + 0,83$$
$$[\,31,97\,;\,33,63\,]$$

In other words, the population mean is probably (with 90% confidence) between the weights of 31,97 and 33,63 grams.

5. C

Quite important: the p-value does not tell us the probability that the null hypothesis is true! That might have been easier and more intuitive, but such a thing isn't possible: the null hypothesis of this study is either true, or not true – there's no probability involved in that. The p-value assumes that the null hypothesis is true, and tells us <u>in that case</u> what would be the (conditional) probability of the sample mean we found. Which leads us to say one of two things:

♦ If the probability is larger than – commonly – 5%, we say: 'Okay, this sample mean wouldn't be so surprising to find if the null hypothesis was correct.'

- If the probability is smaller than or equal to 5%, we say: 'Hey, but then my sample mean would be very special indeed if the null hypothesis is correct!'

In the latter case we reject the null hypothesis.

6.

a) No: only the probability of a Type II error will decrease with a larger sample. The probability of a Type I error always equals the significance level α. Therefore the producer would have to reduce this significane level if he's afraid of Type I errors. The disadvantage: this will cost power! Go ahead and reduce the significance area in the drawing from paragraph 6.5. The critical boundary moves to the right and the green area under the sampling distribution (the power) becomes smaller. So much for mathematics. Does this also make conceptual sense? Yes: use a smaller α and you make it more difficult to reject the null hypothesis, thereby reducing the chance that you'll succeed justly.

b) Not 5%: this probability equals 0. The researcher doesn't reject the null hypothesis, so he can't falsely reject it. Trick question! MWUAHAHA!

c) Also use my elaboration in the final paragraph.
Draw the null distribution and the true sampling distribution next to each other; let the image in paragraph 6.5 inspire you.
The critical sample mean is the sample mean that would lead you to reject the null hypothesis $\mu = 32$. This mean has a p-value of 0,05.
The z- or t-table tells us that the critical z-score is approximately 1,645 (look up the z-score with a p-value of 0,05). $Z^* = \frac{\bar{X}^* - \mu}{\sigma/\sqrt{N}}$ gives $1{,}645 = \frac{\bar{X}^* - 32}{1{,}6/\sqrt{10}}$. If we solve this equation, we get $\bar{X}^* = 1{,}645 * {}^{1{,}6}/_{\sqrt{10}} + 32 = 32{,}83$.

d) So, the null hypothesis will be rejected if $\bar{X}$ exceeds 32,83; the null hypothesis will not be rejected if $\bar{X}$ is smaller than 32,83. What is now the true probability that $\bar{X}$ will be smaller than 32,83? This is the red area under the true sampling distribution. We can calculate it in the old-fashioned way:
The z-score is $Z = \frac{32{,}83 - 33}{1{,}6/\sqrt{10}} = -0{,}34$
It follows from the z-table that $\boldsymbol{P(Type\ II\ error)} = P(Z < -0{,}34) = \mathbf{0{,}3669}$.
This is a considerable probability!

e) The power is equal to the green area under the true sampling distribution: $1 - P(Type\ II\ error) = 1 - 0{,}3669 = \mathbf{0{,}6331}$. This isn't such a great power. A larger sample could fix the problem.

CHAPTER 7-9 T-TESTS (COMPLETE)

Exercises for peaches

1.

a) The null hypothesis should be straightforward:

$$H_0: \mu_{neutral} = \mu_{religious}$$

The alternative hypothesis allows for some discussion. Wilde clearly expected the religious group to be less sexually active, which defends a one-sided alternative. But if you want to be on the safe side, a two-sided test is fine as well.

$$H_0: \mu_{neutral} > \mu_{religious} \quad \text{or} \quad H_0: \mu_{neutral} \neq \mu_{religious}$$

b) $\bar{X}_{neutral} = \frac{2+0+3+6+4}{5} = 3$ and $\bar{X}_{religious} = \frac{2+1+5+10+2}{5} = 4$

This was actually not what Wilde had expected; why is the religious group more active? Let's inspect the sample results in more detail. There are a few, um, curious scores in there. The 10 in the religious group is a clear outlier (in God's name, what happened?), so it's likely to distort the group's mean. If we left it out, the second mean would drop down to 2,5 – which is more in line with expectations.

c) Since you skipped the assumptions check, let's assume equal variances. Cohen's d was

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{s_p} = \frac{|3-4|}{3{,}041} = 0{,}33$$

Between small and medium, around and about.

And the t-value equalled

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{3-4}{3{,}041 * \sqrt{\frac{1}{5} + \frac{1}{5}}} = \frac{-1}{1{,}923} = -0{,}520$$

The degrees of freedom were $df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$, so the p-value found in the Appendix was greater than 0,25. In case you chose to conduct a two-sided test earlier, you doubled the p-value and concluded that it exceeded 0,50. Not significant, in any case – you couldn't reject the null hypothesis.

2.

a) You sure did:

neutral	religious	$d = X_{neutral} - X_{religious}$
2	2	$2 - 2 = 0$
0	1	$0 - 1 = -1$
3	5	$3 - 5 = -2$
6	10	$6 - 10 = -4$
4	2	$4 - 2 = 2$

b) $H_0: \mu_d = 0$

c) It was

$$\bar{X}_d = \frac{0 - 1 - 2 - 4 + 2}{5} = -1$$

Which shows, again, that the neutral condition actually engaged in *less* sexual intercourse on average. The outlier in the religious group may explain this.

d) We now have five matched pairs of student groups, so $N = 5$. Yep!

In that case, T becomes

$$T = \frac{\bar{X}_d - \mu_{d,0}}{s_d/\sqrt{N}} = \frac{-1-0}{2{,}236/\sqrt{5}} = -1{,}000$$

Huh. Nice. ☺ The degrees of freedom were $df = N - 1 = 5 - 1 = 4$. That put the p-value in the Appendix between 0,15 and 0,20. If you opted for a two-sided test, you doubled the p-value and arrived at a result between 0,30 and 0,40 (just double the bounds). Either way, the t-test was not significant, and once again you couldn't reject the null hypothesis.

3.
Mean difference: this is -1, identical for both tests! I would not call that surprising, since you kept comparing the same two conditions.
Standard error: this one differs between the tests! The independent samples t-test looks at the behaviour of the separate sample means, but the paired samples t-test investigates the difference scores of the pairs. Here's another overview:

neutral	religious	$d = X_{neutral} - X_{religious}$
2	2	$2 - 2 = 0$
0	1	$0 - 1 = -1$
3	5	$3 - 5 = -2$
6	10	$6 - 10 = -4$
4	2	$4 - 2 = 2$
$\bar{X}_1 - \bar{X}_2 = -1$ $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1{,}923$		$\bar{X}_d = -1$ ${s_d}/{\sqrt{N}} = 1{,}000$

Degrees of freedom: the independent t-test 'sees' 10 individuals in 2 groups, which makes for 8 degrees of freedom. By contrast, the paired t-test 'sees' 5 pairs. This yields only half as many degrees of freedom (4). It's like having only half as many participants. (If it helps, also think of a repeated measures design: in it, we truly would've had only 5 participants who went to both vacation houses, not 10.)
Power: in spite of having less degrees of freedom, the paired t-test appears to have greater power. The t-value is more extreme (further away from 0) and the p-value is lower. This has to do with the advantages of good matching. Let's look at the last pair for example: we see that the student group in the neutral vacation house had 4 sexually active members, while its partner group in the religious house had only 2. Since the groups are so similar in terms of social backgrounds, IQs and risk-seeking personalities, it's less likely that their sexual activity differed due to such traits. This makes it easier to prove that their difference was more than just pure chance, and in fact resulted from the atmosphere of the vacation houses.[3]

4. C
Matching is all about making sure that the individual participants within each pair are as similar as possible. If they are, the scores in the neutral group will correlate highly with those in the religious group; after all, we try to simulate that we're measuring the same participant twice.

5. A
Only the independent samples t-test assumes equal variances between the groups! The paired samples t-test doesn't. It studies difference scores after all, and there is only one set of difference scores (with one single variance) instead of two. ☺

6. B
This is exactly why we should always look at the sample results, before performing the test. The *Paired Samples Statistics* table shows that the neutral sample was sexually more active than the religious one (compare the means). We now know the direction of the sample difference. The t-test subsequently demonstrates that this difference was significantly large.

Exercises for pirates

1.

a) Make no mistake: this is an investigation of 1 population. The population distribution of the Nova Fyra was already known, so it would cost accuracy (and therefore power) to draw a sample from these old models as well. Research and Development only drew one sample of 30 Photons; that made the appropriate choice a one sample t-test.

b) $H_0 : \mu = 1000$
$H_A : \mu > 1000$

c) Cohen's d was estimated to be $\hat{d} = \frac{1041-1000}{38} = 1{,}08$. So, a large effect!
The t-value was $T = \frac{1041-1000}{{38}/{\sqrt{30}}} = 5{,}909$.
If necessary draw the t- distribution, with the mean (0) and the t-value that was found.
The degrees of freedom amounted to $30 - 1 = 29$.

[3] In ANOVA terms (see chapter 10): in repeated measures and matched pairs designs, the **error variance** is lower.
If the matching procedure sucks and the error variance is hardly reduced, the paired t-test actually has *less* power than the independent t-test, due to its lower amount of degrees of freedom.

It follows from the t-table that $p = P(T > 5{,}909) \ll 0{,}0005$.
This was holy-smokes significant: the null hypothesis could be rejected rock-hard. Good news for the Star Fleet: the Photon was faster than the Nova Fyra!

d) The formula for the confidence interval is $\bar{X} \pm T^{*}\, {}^{s}/_{\sqrt{N}}$. Filling it in gives us (rounded down to whole kilometres per hour):

$$1041 - 2{,}756 * {}^{38}/_{\sqrt{30}} = 1041 - 19$$
$$1041 + 2{,}756 * {}^{38}/_{\sqrt{30}} = 1041 + 19$$
$$[1022; 1060]$$

Or, the population mean was most likely (with 99% confidence) between the speeds of 1022 and 1060 kilometres per hour.

2. A

This was a paired samples t-test, so the groups that were measured had to be dependent. Option A presents us with a design of repeated measures; in option B, different fighter spacecraft were compared. Theoretically the research team might have matched individual Photons and Z-Wings, but I wouldn't really know on what basis. As long as we don't have any more information, it's logical to assume that the Photons and the Z-Wings constituted independent groups. The study in option B thus required an independent samples t-test.

3. B

The paired samples t-test is computationally equivalent to a one sample t-test, performed on the difference scores between two measurements.

4.

a) Certainly: the observer compared two independent samples.

b) Variable quantitative: met.
Independent groups: met.
Normality: in my opinion the histograms (the distributions of sample scores) look fine, dear reader. In case you do have your doubts about this assumption, the samples were large enough to correct for a minor violation.
Equal variances: this was probably met, since the rule of thumb gives us no reason to panic $\left(\frac{40{,}637}{35{,}760} = 1{,}14 < 2\right)$ and Levene's Test is not significant at all ($p = 0{,}715$).
Conclusion: no violations that would be problematic.

c) We may assume equal variances (see question b), so we need $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

$$s_p = \sqrt{\frac{21 * 35{,}760^2 + 18 * 40{,}637^2}{21 + 18}} = 38{,}089$$
$$s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 38{,}089\sqrt{\frac{1}{22} + \frac{1}{19}} = \mathbf{11{,}929}$$

This number agrees with the *Standard Error of the Difference* in the row *Equal variances assumed.*

d) The right number of degrees of freedom is $\boldsymbol{df} = n_1 + n_2 - 2 = 22 + 19 - 2 = \mathbf{39}$. That agrees as well with the output in the row *Equal variances assumed.*

5. B

Namely, he could have resorted to… the confidence interval! The value 0 doesn't lie within this interval (neither in *Equal variances assumed* nor in *Equal variances not assumed* by the way, though the first choice is actually the best one). 0 is thus not a plausible difference between the two population means. The significance level for a 95% confidence interval equals $100\% - 95\% = 5\%$. The null hypothesis $H_0: \mu_{training} = \mu_{control}$, also $H_0: \mu_{training} - \mu_{control} = 0$, should therefore be rejected at this significance level.

6. C

The null hypothesis we can reject: the training demonstrably had an effect, so B drops out. The question is now, did the effect also manifest in the desired direction? Looking once more at the sample scores (Group Statistics), we see that the mean improvement was far lower for the trained pilots: they declined fiercely on average! The pilots on the waiting-list scored much higher (approximately 0; they hardly changed on average, which is logical, since they weren't experimented upon). This difference between the training and control group turned out significant. The observer's conclusion was therefore answer C… 😐

Exercises for peaches

1.

a) $H_0: \mu_{1\,(Fungus\ Fantasticus)} = \mu_{2\,(Rainbow\ Spore)} = \mu_{3\,(Psycho\ Classic)}$
$H_A: not\ all\ \mu\ are\ equal$

b) First we need to calculate the means of all groups, and the Grand Mean too:

$$\bar{Y}_1 = \frac{2+7+6+1}{4} = 4$$
$$\bar{Y}_2 = \frac{10+6+10+10}{4} = 9$$
$$\bar{Y}_3 = \frac{2+2+9+7}{4} = 5$$
$$\bar{Y} = \frac{2+7+6+1+10+8+8+10+4+6+5+5}{12} = 6$$

Now we can begin with the Sums of Squares. Are you not seeing what happens in these calculations? Have another look at the theory!

$$\boldsymbol{SS(Total)} = \sum_{ij} (\boldsymbol{Y_{ij}} - \boldsymbol{\bar{Y}})^2 =$$
$$(2-6)^2 + (7-6)^2 + (6-6)^2 + (1-6)^2 + (10-6)^2 + (6-6)^2 +$$
$$(10-6)^2 + (10-6)^2 + (2-6)^2 + (2-6)^2 + (9-6)^2 + (7-6)^2 =$$
$$16 + 1 + 0 + 25 + 16 + 0 + 16 + 16 + 16 + 16 + 9 + 1 =$$
$$\mathbf{132}$$

$$\boldsymbol{SS(Between)} = \sum_{ij} (\boldsymbol{\bar{Y}_i} - \boldsymbol{\bar{Y}})^2 =$$
$$(4-6)^2 + (4-6)^2 + (4-6)^2 + (4-6)^2 +$$
$$+(9-6)^2 + (9-6)^2 + (9-6)^2 + (9-6)^2 +$$
$$(5-6)^2 + (5-6)^2 + (5-6)^2 + (5-6)^2$$

Take care that <u>for every individual</u> we count the difference between his or her group mean and the overall mean! But, well, since that difference is the same for everyone who's in the same group, we can also just calculate it per group and multiply by the number of persons. (Note that the j, for the participant, then vanishes from under the sum sign.)

$$\boldsymbol{SS(Between)} = \sum_{i} \boldsymbol{n_i} * (\boldsymbol{\bar{Y}_i} - \boldsymbol{\bar{Y}})^2 =$$
$$4 * (4-6)^2 + 4 * (9-6)^2 + 4 * (5-6)^2 =$$
$$4 * 4 + 4 * 9 + 4 * 1 =$$
$$\mathbf{56}$$

$$\boldsymbol{SS(Within)} = \sum_{ij} (\boldsymbol{Y_{ij}} - \boldsymbol{\bar{Y}_i})^2 =$$
$$(2-4)^2 + (7-4)^2 + (6-4)^2 + (1-4)^2 + (10-9)^2 + (6-9)^2 +$$
$$(10-9)^2 + (10-9)^2 + (2-5)^2 + (2-5)^2 + (9-5)^2 + (7-5)^2 =$$
$$4 + 9 + 4 + 9 + 1 + 9 + 1 + 1 + 9 + 9 + 16 + 4 =$$
$$\mathbf{76}$$

There, the worst is behind us. Now the ANOVA table. See the theory for the calculation rules. We look up the p-value in the appendix.

	Sum of Squares	df	Mean Square	F	p
Between Groups	56	2	28	3,32	⟨ 0,05 ; 0,10 ⟩
Within Groups	76	9	8,44		
Total	132	11			

c) Conclusion: not significant. We cannot demonstrate that any kind of mushroom makes for more pleasant hallucinations on average!

2. A

Fewer participants mean less power (see chapter 6) and that's a reasonable explanation for the non-significant ANOVA in exercise 1. Were any assumptions violated? That seems unlikely to me. Assumptions are always about the population, so if we use the SPSS output from the large sample, we can check whether they were met for both ANOVAs. The histograms suggest rather symmetrically distributed populations. In case equal variances had not been met, this wouldn't have mattered for either ANOVA, since all groups are consistently equally large.

Lastly the suggestion of a Bonferroni correction is bollocks: we only apply that to the pairwise comparisons, not to the ANOVA itself. Return to the theory if you can't remember what the Bonferroni correction is for.

3. B

If Bonferroni won't come to us, we'll just have to go to Bonferroni – in other words, correct ourselves. That's easily done by hand! We can divide the significance level α of each test by 3 (after all there are 3 tests) or triple all the p-values. In this case that doesn't change the list of significant results: Rainbow Spore differs significantly from both Fungus Fantasticus and Psycho Classic and that's all.

Now, is Rainbow Spore faring better or worse than the other two mushrooms? I hinted at the *Descriptives* table for this reason. There we see that the testers of Rainbow Spore gave the highest mean to their tripping experience. Next, the pairwise comparisons told us that this mean is also significantly higher than the other two.

4. B

The ANOVA is far from significant ($p = 0{,}214$), so we can't reject the null hypothesis; we fail to demonstrate that the type of mushroom (the treatment) affects the quality of the trip. In that case this quality differs among individuals solely due to other factors.

5. D

See the theory for details. MS(Between) may well be an unbiased estimator of the error variance as well, since it doesn't seem like the null hypothesis is untrue. The group means differ purely due to error effects in that case.

6. A

The null hypothesis can be rejected if MS(Between) is strikingly larger than MS(Within). After all, MS(Between) is an unbiased estimator of the error variance in the population, plus the variance as a result of the group effect; whereas MS(Within) is an unbiased estimator of the error variance only. If MS(Between) is strikingly larger, that suggests that such variance as a result of the group effect really exists. The question is: is the current result striking enough? That turns out not to be the case, for the ANOVA is not significant. An MS(Between) which is slightly more dan 1,5 times as large as MS(Within) (this is the F-ratio: 1,594) turns out to occur rather often if the null hypothesis is true: 21,4% of the time we draw a sample (this is the p-value: 0,214).

Exercises for pirates

1.

a) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

b) The individuals in the 'luck' and the 'greed' condition scored lowest on average (the two groups don't differ that much). The condition 'hard work' did better, and the condition 'passion' achieved the highest mean. An explanation for this can be that a passionate role model motivates participants to do their best: they will think that they can earn a lot of money by having fun at their task. A role model which works hard can be inspiring as well: work hard and you'll advance. It's thinkable that the participants in the first condition were not inspired by a businessman who had been mostly lucky, for in that case the cause of wealth is largely outside your own zone of influence. A greedy role model won't likely inspire either; participants don't want to identify with such a person.

c) Dependent variable quantitative: met.
Independent groups: met.
Normality: most distributions of sample scores are a bit skewed to the right (look at the histograms). However, the skewness and kurtosis of these distributions could be 0 in their respective populations (check the *Explore* table: all skewness and kurtosis values fall between -1 and 1). In that case the samples would come from normally distributed populations after all. Furthermore most tests of normality are insignificant (study the *Tests of Normality* table). There's only some doubt about the population distribution of the 'luck' condition, since the Shapiro-Wilk test is barely significant ($p = 0{,}049$). Should the normality assumption be violated, then that's hardly a problem: each sample contains 25 observations ($n_i = 25$). The ANOVA is thus robust against an eventual violation.
Equal variances: the standard deviations of the four samples differ from each other: the largest one is 69,394 (in the 'hard work' condition) and the smallest one 44,198 (in the 'greed' condition). The ratio is $\frac{69{,}394}{44{,}198} = 1{,}57$. This difference isn't so big in the end: the largest standard deviation is about one and a half times larger than

the smallest one, so not twice as large or even more. Additionally, Levene's Test is not significant ($p = 0{,}115$). We may therefore suppose that this assumption has been met: it's fairly plausible that all these sample standard deviations are estimating the same σ. By the way, each sample contains an equal number of observations (25), so the ANOVA would've been robust against a violation of equal variances.

2. A

MS(Within) is the variance within groups, and describes the degree to which individuals within one and the same group still differ from each other – or, the degree to which individuals deviate from their group mean. The square root of the variance is the standard deviation, or the average deviation per person. Since MS(Within) describes the variance within *all* groups, the quantity applies to the whole sample. For those who want to explore this a bit further: mind that MS(Within) is equal to s_p^2, the pooled variance.
Answer B describes the square root of MS(Total), a quantity which normally isn't in the ANOVA table since we don't need it. $MS(Total) = \frac{SS(Total)}{df(Total)} = \frac{341782{,}990}{99} = 3452{,}35$. The square root of this is the standard deviation of all individuals relative to the general mean: 58,76. That agrees with the *Total Std. Deviation* in the *Descriptives* table! ☺
Answer C gives a rough impression of what MS(Between) implies.

3. C

The test result is really significant indeed, but it's not prudent to use the p-value to determine the size of the effect; the p-value is influenced by the effect size, but by the size of the sample (N) as well. Using a larger sample one obtains more power, so an accompanying statistical test can have a lower p-value – even if the exact same effect is investigated. The measure of effect size η^2, however, is insensitive to the sample size and stands for the proportion of explained variation. $\eta^2 = \frac{SS(Between)}{SS(Total)} = \frac{69814{,}750}{341782{,}990} \approx 0{,}204$. What does this indicate? The fact that not every participant collected the same amount of coins can be explained – at sample level – to the extent of 20,4% by the fact that not every participant had the same role model. That's quite a lot: one single factor already seems responsible for 0,204 of all the variation. We can therefore regard this as a large effect. Cohen's guidelines are:

- 0,01: small effect
- 0,06: medium effect
- 0,14: large effect

These are of course just guidelines.

4. B

Seeing that the ANOVA is significant and the null hypothesis can be rejected, there is demonstrably a group effect. MS(Between) thus becomes an unbiased estimator of the error variance plus the variance as a result of this group effect. MS(Between) therefore estimates more than MS(Within), such that the F-ratio acquires an expected value larger than 1. However, we don't know precisely what the expected value of F is. After all, we've merely estimated both the error variance and the variance as a result of the group effect. Probably the F-value of this ANOVA isn't exactly the F-value we'd get on average if we drew Miyamoto's sample again an infinite amount of times.

5. B

Both A and C are actually saying: 'The probability that the null hypothesis is true is 0,000.' However, there's no such thing as a probability that the null hypothesis from Miyamoto et al. is true[4]; the four groups simply have the same population mean or not. That's the 'truth'; there's no probability involved. The idea of a statistical test is that we start from the null hypothesis: we're sceptical and assume that the groups don't differ in terms of their means. Next we see if the obtained sample would then be *so* special that we can reject this starting point. It turns out that this sample, with such differences between the groups, would hardly ever occur (less than 1 in a 1000 times) if the four population means were the same. That's reason enough to no longer believe that the population means are equal.

6. A

Pay attention: we've got four groups, so we obtain six unique comparisons! Go ahead and count the number of rows in the table (12), and divide it by 2. Or use the little calculation rule $\frac{k(k-1)}{2}$, where k stands for the number of groups: $\frac{4*(4-1)}{2} = \frac{4*3}{2} = 6$. If we sextuple[5] all the p-values, we get the following table (please turn over):

[4] At least not in the frequentist interpretation of the 'probability' concept. There is such a thing in Bayesian statistics, should you find it interesting…

[5] This is an awesome word for 'multiply by six'. Google it: single, double, triple, quadruple, quintuple…

Multiple Comparisons

Dependent Variable: coins

Bonferroni

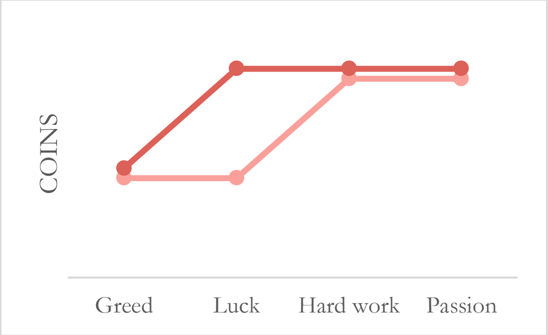
(I) role model	(J) role model	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
luck	hard work	-38,600	15,055	,071	-79,16	1,96
	passion	-59,400*	15,055	,001	-99,96	-18,84
	greed	3,400	15,055	1,000	-37,16	43,96
hard work	luck	38,600	15,055	,071	-1,96	79,16
	passion	-20,800	15,055	1,000	-61,36	19,76
	greed	42,000*	15,055	,038	1,44	82,56
passion	luck	59,400*	15,055	,001	18,84	99,96
	hard work	20,800	15,055	1,000	-19,76	61,36
	greed	62,800*	15,055	,000	22,24	103,36
greed	luck	-3,400	15,055	1,000	-43,96	37,16
	hard work	-42,000*	15,055	,038	-82,56	-1,44
	passion	-62,800*	15,055	,000	-103,36	-22,24

*. The mean difference is significant at the 0.05 level.

(Note: a p-value cannot exceed 1 or 100%.)
After this rather heavy correction, only three unique comparisons remain significant:

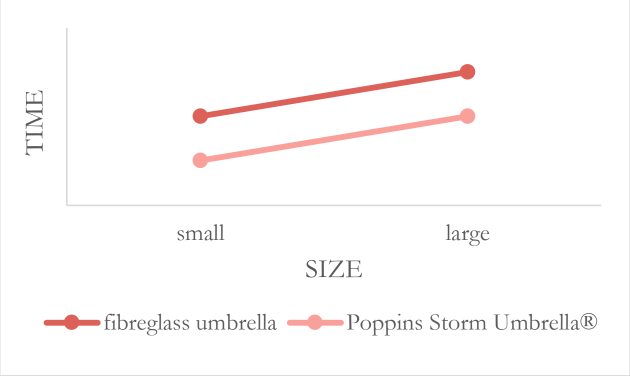
- The difference between ‘luck’ and ‘passion’;
- The difference between ‘hard work’ and ‘greed’;
- The difference between ‘greed’ and ‘passion’.

This result contains something contradictory because, have a go at trying to plot it:

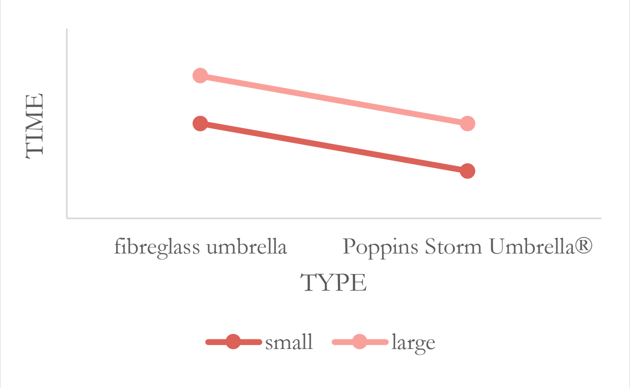


‘Luck’ differs demonstrably from ‘passion’: in that case the light pattern would display the population means best. But ‘luck’ no longer differs demonstrably from ‘hard work’: in that case the dark pattern would be correct. The two patterns are irreconcilable and so a Type I or Type II error must be at work somewhere. Since ‘luck’ and ‘hard work’ did differ significantly *before* the Bonferroni correction, we’re probably making a Type II error in this comparison.

My idea was that people with small umbrellas need less time than people with large umbrellas, and potentially, that the Poppins Storm Umbrella® also makes walking easier than the standard fibreglass umbrella. We could visualise these effects like so (all the dots represent a population mean).



Or with the umbrella's type on the X axis instead:



These are just schematic drawings, mind you. If you have other expectations, that's perfectly fine, as long as you can support them with solid theoretical arguments. ☺
In two-way ANOVA terms, I would expect a main effect of both the umbrellas' SIZE and TYPE, but no interaction.

2.
The best alternative, in my opinion, would be an adjusted Bonferroni correction. With 6 comparisons, we should multiply the p-values by 3. If you don't want that, a 'standard' Bonferroni correction is possible as well, but that has you multiply the p-values by 6.
In both cases (luckily it makes no difference here), the result is that 3 out of 6 comparisons are significant. The large fibreglass umbrella differs from all the other conditions. Looking back at the sample results – which we should actually have done beforehand – we see that participants who use this kind of umbrella are the fastest on average. This actually contradicts my expectations from exercise 1. Surprising…

3.
We need two tables: the sample means are in the *Descriptives* table, and the contrast coefficients can be found in the table with the same name. They're [-1,-1, 1, 1]. Filling in everything yields

$$L = \sum_i c_i \bar{Y}_i = -1 * 15{,}28 - 1 * 12{,}05 + 1 * 16{,}98 + 1 * 16{,}26 = 5{,}91$$

4.

a) Note, once again, that groups with the same contrast coefficient belong together.
Contrast 1 combines the fibreglass umbrella groups (small and large) and sets them against the combined Poppins Storm Umbrella® groups. Clearly, this contrast tests the effect of the umbrella's type.
Contrast 2 contrasts the small umbrella groups with the large umbrella ones, and therefore tests the effect of the umbrella's size.
Contrast 3, finally, curiously combines 'opposite' groups. The upshot is that it tests for interaction – it assesses whether the type effect depends on the umbrella's size, and vice versa. If you wish to understand why these are the correct coefficients for interaction (not mandatory), consider this. You may need some knowledge from chapter 12. If there's no interaction, the difference in average TIME when using small versus large

umbrellas should be the same regardless of the umbrella's type. Or, $H_0\colon \mu_1 - \mu_2 = \mu_3 - \mu_4$. Rewrite that to $H_0\colon \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$ and you've got the coefficients [1,-1,-1, 1].

b) Only the main effects of size and type are significantly large; the interaction effect is not.

c) Yes, this contradicts the pairwise comparisons in several ways. For instance, if the umbrella's size always has an effect, regardless of the type, why isn't the comparison between the small and large Poppins Storm Umbrella® significant? A good explanation may be that the contrast analyses are way more specific and as such more powerful. These three contrasts are also orthogonal (verify it!), so they don't need a Bonferroni correction! This makes them *even more* powerful.

5. A

Polynomial contrasts? Have you considered what those would mean in practice, dear reader? Have a look at the *Means Plot* at the end of the output. Ahem… 'the more the umbrella is large Poppins Storm Umbrella®-ish, the faster or slower people perform…' Complete nonsense. GROUP is a nominal variable! We can only use polynomial contrasts if the independent variable, though categorical, represents a quantitative scale.
Answer C is therefore the worst answer you could've picked. The cubic contrast is significant in statistics, but meaningless in reality. So is the linear term. The quadratic contrast isn't even statistically significant, so B still contains an erroneous statement. A is best, because it's fully correct.

6.

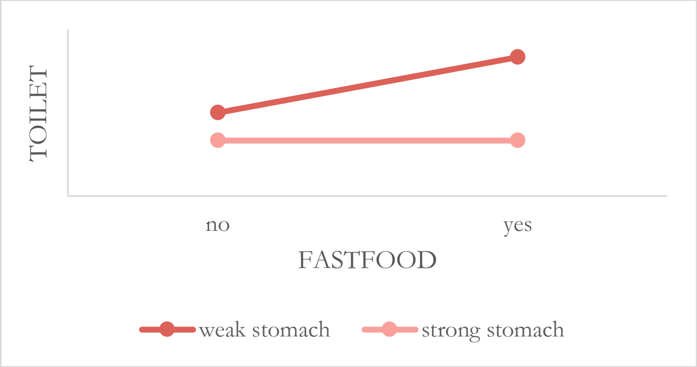
a) I'd pick the 'standard' contrasts (the non-polynomial ones). They're powerful, are very specific and make practical sense.

b) C
After all, the sample results show that users of Poppins' prototypes actually take more time on average… Contrast analysis 1 shows that this difference in performance is significantly large.

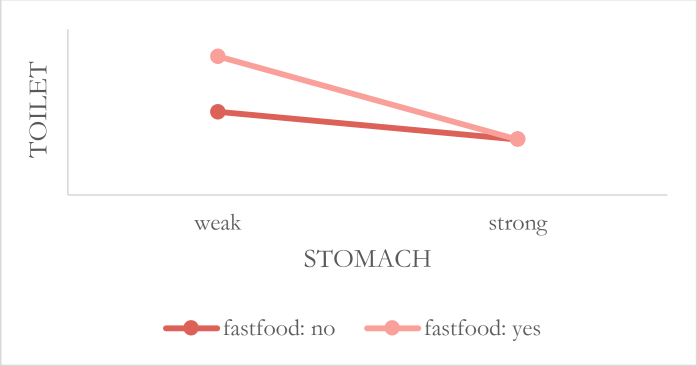
CHAPTER 12 TWO-WAY ANOVA (COMPLETE)

1.

a) Personally I was thinking of an interaction effect:



Gastro patients with a weak stomach may particularly suffer from diarrhoea if they ate fastfood. It may be that eating fastfood doesn't affect patients with a strong stomach.
We can also look at the interaction effect from the other angle:



Patients with a weak stomach need to visit the loo more often on average than patients with a strong stomach, but the difference is greater when the patients ate fastfood.
Again: other ideas, if substantiated, are correct as well.

b) This is a non-orthogonal design, although it's pretty close. That's because the participants aren't distributed across the conditions at an entirely constant ratio:

		STOMACH	
		weak	strong
FASTFOOD	no	19	20
	yes	16	16

This isn't odd; it's not an experiment after all. The participants were observed in their natural environment (ahem), so King & McDonald had no control over the combination group sizes. Anyway, if both ratios had been 19/16 or 20/16, the factors FASTFOOD and STOMACH would've been statistically independent. Now they display a mild association, and the risk of mild confounding exists. We can see these frequencies in the *Descriptive Statistics* table.
To correct for potential confounding, we let SPSS calculate Type III Sums of Squares in the ANOVA. In an orthogonal design, the Sums of Squares of the separate effects don't need to be corrected and so these neatly add up to the Corrected Model, but that's not the case here: $623{,}754 + 140{,}695 + 323{,}719 = 1088{,}168 \neq 1010{,}406$.
Also, we can ask SPSS for Estimated Marginal Means. These group means are unweighted, and as such corrected for potential confounding. Hence they differ lightly from the means in the *Descriptive Statistics* table.

c) Dependent variable quantitative: met.
Independent groups: met.
Normality: TOILET clearly follows a skewed distribution within most groups: each condition featured some gastro patients who spent about half a day longer at the loo than the majority. Multiple normality tests are significant. We should therefore not believe that the assumption of normality has been met.
The question is now if the ANOVA is robust against a violation of this assumption. Two conditions contain only 16 participants. This is a small number. It's sufficient when the sample contains at least 15 observations, *and* isn't extremely skewed, *and* contains no outliers. It is then debatable what you should consider extremely

skewed. I, for one, would give the ANOVA a go. However, I would only trust test results which were clearly very significant, or clearly not significant at all.
Equal variances: this assumption is obviously not met. The largest and the smallest standard deviation differ strongly $\left(\frac{9{,}571}{2{,}903} = 3{,}30 > 2\right)$ and Levene's Test is highly significant ($p = 0{,}001$). Is the ANOVA robust against this violation? I'm not so sure: the groups are not equally large. Your conclusion, dear reader, is allowed to sound: 'No, the ANOVA is not robust and we actually shouldn't perform it.'[6]

2. B
This design is non-orthogonal, so as long as the interaction term is in the model, the main effects should not be interpreted. A speaks of a main effect of the STOMACH and is as such wrong. B defines the interaction effect and this one is significant indeed ($p = 0{,}007$). C, finally, speaks of an association between the two independent variables, STOMACH and FASTFOOD (so not of an interaction effect). This is a rather silly claim which is tested nowhere in the output.
Incidentally, C is true at sample level: the people with a weak stomach ate fastfood slightly more often $\left(\frac{16}{35}\right)$ than the people with a strong stomach $\left(\frac{16}{36}\right)$ (in other words, the design is slightly non-orthogonal). The difference is minuscule though, and there's currently no reason to believe that the same applies to the population.

3. B
In general a non-significant factor had best be removed, since we need to take it into account and that costs power (because it costs a degree of freedom). However, the effect of the stomach is already quite significant and it's likely to stay that way when the factor fastfood is removed.

4. A
In the sample, we do see an effect: the groups who ate fastfood need to visit the loo a bit more often on average (preferably check this model's *3. FASTFOOD* table). As a nice detail, this is also expressed by the Sum of Squares of the factor FASTFOOD: if FASTFOOD had no effect at all, not even in the sample, how strongly would the FASTFOOD groups vary from each other? They would not vary at all: the FASTFOOD Sum of Squares would be zero!
However, the sample effect is too small to demonstrate that there's also an effect in the population ($p = 0{,}089$); for now we'll have to attribute the sample effect to chance.
Note: in a sample, we'll actually always see all effects to some extent, just because it's a sample and its groups will differ due to chance. Wouldn't it be interesting if these 39 patients who ate no fastfood had the *exact* same TOILET average as these 32 fastfood eaters? The statistical test is meant to assess whether the differences were caused by more than mere chance.

5. B
Apologies: the models with a simple effect are of course two t-tests! I replaced the two one-way ANOVAs that were there originally, to make things a bit easier for you. ☺

If you picked answer A, I wonder how you imagined it. If we say that a weak stomach has an effect, we always mean that in relation to something else. People with a weak stomach need to visit the loo more often than… people with a strong stomach of course. In that case, a strong stomach automatically has an effect as well: you need to visit the loo less often than… if you have a weak stomach. If you only look at the weak stomachs, STOMACH is no longer a variable. We can only test the effect of variables, by detecting differences.
Now look more closely at the output. In fact, it tests the effect of FASTFOOD, for the two STOMACH groups separately. Among people with a weak stomach, an effect of FASTFOOD is demonstrated ($p = 0{,}012$); it is not among people with a strong stomach ($p = 0{,}345$). This is consistent with my personal expectations from exercise 1, as well as with the interaction effect we found in exercise 2. The required Bonferroni correction (both p-values times 2) does not change the conclusions.

6. C
Once more: in the full model, look at the interaction term first. If it's not significant, we should remove it in order to interpret the main effects (we may also leave it in when the design is orthogonal). If the interaction term *is* significant, like now, main effects are (often) rendered meaningless. We should study simple effects in such a scenario.

[6] What you could defend is that the two most extreme standard deviations also come from the smallest samples (and are thus weighted less heavily), but that's going a bit too far for a normal statistics course.

CHAPTER 13 ANCOVA (COMPLETE)

1. B

You may very well have chosen answer A at first. If that happened, I hope exercise 2 changed your mind! ☺ Don't worry – as you probably understood, I designed the exercises like this on purpose.

Anyway, even though 'emotionality before' is probably not a confounder (see exercise 2), this one-way ANOVA is still problematic. We can't find an effect of 'music' on 'emotionality', but we have not used the covariate yet – which may be good for the power of the test. The notion of possible interaction (option C) sounds relatively far-fetched, I think.

2.

a) It's an ANOVA, which can only compare groups (categories). The independent variable must therefore be 'music'. The dependent variable is stated in the upper left corner: it's the pre-test, 'emotionality before'. Thus, this ANOVA tests whether the three music groups had the same population mean on this pre-test. It turns out that they probably did, since the test is not significant at all ($p = 0{,}886$).

b) They did not, it would seem.

c) No: draw the confounding triangle. The relationship between 'music' and 'emo_before' is not present. Therefore, if the music groups differ in their emotionality directly after the musical manipulation, this can't be because they already differed beforehand.

d) The introductory text tells us that the participants were allocated to the conditions at random. In other words, this was an experiment! It's quite logical, then, that emotional and cold participants have been allocated pretty evenly. This makes confounding unlikely in experiments.

e) Well, perhaps the covariate could still be useful as a power booster? I certainly think so: if we look at how the individual participants differed in their emotionality before the experiment began, this can explain to a great extent why they differed directly after. The pre-test thus explains a good deal of variation in the 'emotionality' scores. It's a very typical covariate for experiments.

3. A

The normality assumption cannot be checked right now (we have no histograms, skewness-kurtosis statistics and the like). The assumption of equal variances is indeed violated (Levene's Test: $p = 0{,}003$), but the three groups have the same size. As for the additional assumptions of ANCOVA: the scatterplot shows no curvilinear relationships, so I'd say linearity has been met. Furthermore the interaction test in Model 4 is not significant, so non-interaction appears met as well.

4. B

The scatterplot in Model 3 shows the sample result: the three regression lines don't run perfectly parallel in the sample, and so the bs are not exactly equal. In short, light interaction occurs. However, this interaction is not significant (see Model 4).

5. C

We discarded Model 1 earlier because it's lacking in power (too much error variance). Model 4 still contains an interaction term and is as such a SCHMANCOVA; the main effect of 'music' should not be interpreted here. The effect of 'music' is tested validly in Model 5. It comes out significant ($p = 0{,}038$).

6. A

This question is about the way ANCOVA corrects for potential confounding (see the explanation in the book). Make a drawing of this, like in the orc example: put the dependent variable ('emotionality') on the Y axis and the covariate ('emo_before') on the X axis, and draw one regression line per 'music' group (sloping upward in this case). It's pretty much like Model 3 in the Supplement, only the three lines run parallel because we assume no interaction.

Now, the chosen point of reference is currently an emotionality before equal to 51,49 (this is reported below the EMM tables). This is the average emotionality before in the entire sample. If everyone had been 10 points less emotional before the computer task, the ANCOVA model also predicts lower emotionalities after the task (look at the regression lines). All groups then obtain a lower Estimated Marginal Mean. Since the ANCOVA assumes non-interaction, however, the lines run parallel: hence the three Estimated Marginal Means all go down to the same degree. Their mutual differences will thus stay the same.

The idea of this exercise is to show you that the point of reference we choose doesn't matter: we're interested in the differences between the 'music' groups, and these don't change.

Chapter 14 Within-subjects ANOVA (complete)

1. A

'Participant' is a required random factor for a within-subjects design, even when it doesn't affect the dependent variable demonstrably (it simply means that the persons are not very different on Clarkson's test); if we remove it, we'll treat all measures as independent participants, and that's unacceptable. Adding 'participant' as a random factor turns the design into a within-subjects one; SPSS doesn't see that and believes that 'participant' is simply an additional factor. This doesn't matter for the analysis. The only serious problem of *GLM Univariate* is described by answer A: if sphericity is not met, you'll need an epsilon correction. *GLM Univariate* will never perform one.

2. B

Further behind the comma, the p-values of the two tables in question differ after all; the tests are not exactly the same. They do test a similar null hypothesis, which boils down to 'the number of signs has no effect on the death count'.

3. A

The p-value of Mauchly's Test is not the right one: this test assesses whether we've got sphericity (null hypothesis: yes), so it's a step before the ANOVA. It's significant, so the sphericity assumption has clearly been violated. If we now want to test the effect of the number of traffic signs on the death count, we can't use the ordinary univariate test; an epsilon correction is needed, preferably Greenhouse-Geisser. Look inside the *Tests of Within-Subjects Effects* in the row *Greenhouse-Geisser*: the p-value reported here is what we seek. 0,694 is no p-value, but the epsilon estimate by Greenhouse-Geisser, found in the *Mauchly's Test of Sphericity* table.

4. B

Lower-bound is a heavy epsilon correction, but of course that doesn't mean that the accompanying test can never become significant; it's relatively lacking in power, but naturally not completely powerless. The other tests are extremely significant, which suggests the same will go for Lower-bound. You can't be completely sure though.

5. C

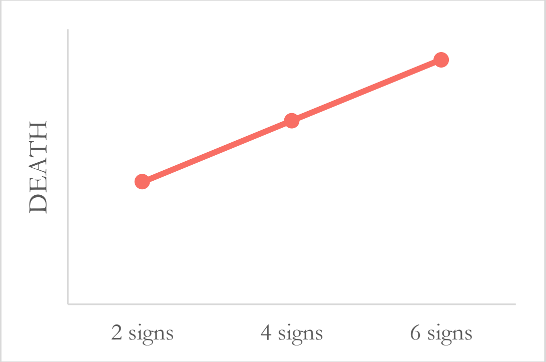
GLM Univariate assumes sphericity. Because of that, the standard errors of the pairwise comparisons are all averaged (the standard error of each comparison t-test is $^{s_d}/_{\sqrt{N}}$, and s_d is assumed to be equal for each set of difference scores). The mean differences remain exactly the same; they just express the sample differences in the death count among the conditions.

With regards to option B: assuming sphericity does give more power (the degrees of freedom are higher), but that doesn't mean that each pairwise comparison will have a lower p-value. This may not be the case if its standard error is much higher than in *GLM Repeated Measures*.

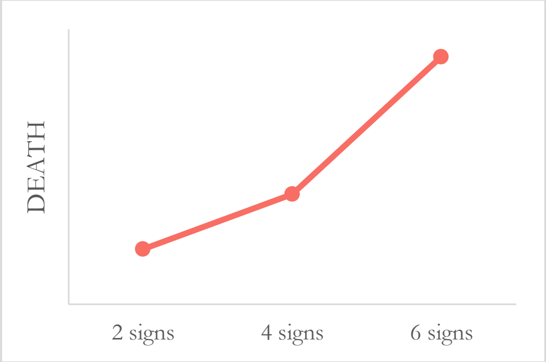
6.

a) Researcher Clarkson was expecting a trend: a quadratic trend, to be precise. The amount of overrun bicyclists and pedestrians would increase quadratically with the number of road signs. Now, the *Tests of Within-Subjects Effects* only allows us to establish if there's a general effect of the traffic signs, not what this effect looks like: the null hypothesis is that all three conditions have the population mean, against the alternative that at least one mean is different. If the amount of collisions increases ever faster with an increasing abundance of road signs, all population means must be different. We cannot demonstrate the latter using the ANOVA alone.

b) The pairwise comparisons are able to tell us which means differ from each other, and are as such more specific than the ANOVA. However, we can only compare two conditions at a time. Hence, this analysis doesn't say that much about Clarkson's exact expectation either. Suppose that condition 2 (4 traffic signs) and condition 3 (6 signs) differ. Then several things may be going on:



Linear



Quadratic

That's why the pairwise comparisons also fail to be specific enough. On top of that, they're less powerful due to a necessary Bonferroni correction.

c) The quadratic trend Clarkson expected *can* be studied directly in the *Tests of Within-Subjects Contrasts* table, and it's demonstrated ($p = 0{,}008$). This contrast is of a higher order than the also significant linear contrast and may take priority as such (see the theory section in the book).

d) If it makes conceptual sense to use a polynomial contrast analysis, always choose this. The categories of the independent variable (signs) actually make for a quantitative scale. That's why I recommend contrast analyses here.

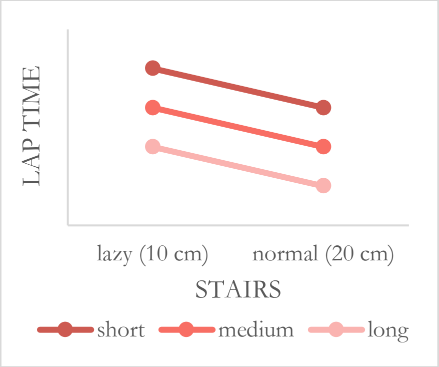
Exercises for peaches

1.

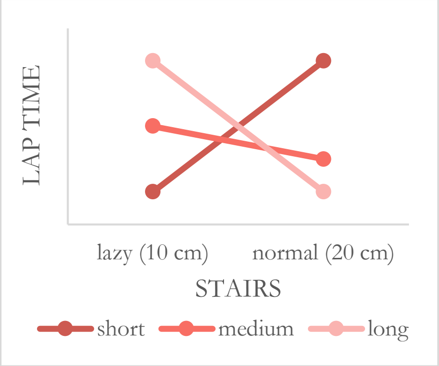
a) The lap time.

b) Leg length: between-subjects factor (length differs between persons). Stairs: within-subjects factor (all persons try both stairs). This means a split-plot design.

c) *Multiple answers possible.* Make a sketch by putting the dependent variable on the Y axis, and the within-subjects factor on the X axis. Now draw a line per between-subjects group.
How about this?



You may believe those lazy stairs are so shitty, everyone is faster on the normal staircase. Naturally, long-legged people are also faster than short-legged ones. If this is your expectation, dear reader, what do you expect to find in the ANOVA? That's right: two significant main effects.
Or maybe you rather thought something like this?



Short-legged persons are faster on the lazy staircase, while long-legged ones benefit most from normal stairs. I'm not quite sure what to do with the medium-legged people here… but anyway, what would you expect in this case? Yes indeed: a significant interaction effect.
Now let's see if you had the truth on your side…

2. B

The assumption of equal covariance matrices states that the between-subjects groups:

- Have the same variance in lap times on the lazy staircase (the first *Levene's Test* evaluates this);
- Have the same variance in lap times on the normal staircase (the second *Levene's Test* evaluates this);
- Have the same covariance (correlation) between lap times on the lazy and normal stairs (neither of the *Levene's Tests* assesses this).

To test if the entire covariance matrices are equal, we need *Box's Test*. While *Box* is not robust against a violation of normality, *Levene* is, as long as the samples are big enough.

3. C

The fact that leg length has 3 levels doesn't matter: this is the between-subjects factor! The multi- and univariate tests only deal with the within-subjects effects, and are thus equivalent when the within-subjects factor has 2 levels. This is the case.

4. C

Pro-tip for exercises like these: draw your own conclusions from the output first! Then check which answer fits with your conclusions. When doing a split-plot ANOVA, we should first study the interaction effect. It's significant. Main effects as described in A and B are now less relevant and may even be meaningless. The simple effects of the stairs differ, and the plot at the end of the Supplement shows us how they differ in the sample: people with short legs are fastest on a lazy staircase, but those with medium and long legs perform better on a normal staircase. This makes it likely that the ideal staircase doesn't exist.

5.
Because we find a significant interaction effect, we shouldn't look at main effects… but the pairwise comparisons test the main effect of leg length! After all, the stairs that the participants climb plays no role in these comparisons. Which makes the pairwise comparisons not informative enough: since there's interaction, the leg effect depends on the type of stairse.

6. B
We've stumbled upon interaction, so an analysis of simple effects still needs to follow. C is about a main effect of the stairs. Pairwise comparisons are unnecessary for that in any case, because stairs has only 2 levels; the ANOVA suffices.

Exercises for pirates

1.

a) Use the Estimated Marginal Means. The difference between the BOTTOM conditions amounts to 4,667 average orders. The corresponding confidence interval is found in the *Pairwise Comparisons* table: in the population, the difference probably lies between 2,999 and 6,334 orders.
b) Compare the means of MOUTH conditions 1 and 3: there's a difference of 3,708 orders, with a confidence interval from 0,453 to 6,964.
c) Confidence intervals would be handy. I used them ☺
d) Yes: we see an interaction effect in the sample, most clearly by looking at the *Profile Plot*.

2. A
Do you remember how the F-ratio of the univariate ANOVA is computed? It's $F(MOUTH) = \frac{MS(MOUTH)}{MS(MOUTH*PERSON)}$. So the quantity we're looking for is actually the error term of this test! Check out the Mean Square at *Error(MOUTH)*, assuming sphericity: it's 21,936.
The tricky thing is that the PERSON factor isn't explicitly mentioned in the output of *GLM Repeated Measures*. But as you can see it still plays a role.

3. A
The first comment is quite correct. The equivalence of the multi- and univariate tests has nothing to do with perfect sphericity though. Have a look back at the Mario case in chapter 14 for instance: even if we assume sphericity, the uni- and multivariate ANOVA give different results. See it like this: if the within-subjects factor has two levels, this has two separate consequences – equivalence of multi- and univariate models, and perfect sphericity for the univariate model.

4. A
'If the MOUTH increases, the ORDERS also increase?' Um, what? MOUTH is a categorical variable! Trend analyses don't make any sense. You can only use them if the independent variable represents a scale. See chapter 14 for details.

5.
First, inspect the interaction effect. It's not significant, so we can check out the main effects. MOUTH and BOTTOM are both highly significant, whether you use the multivariate ANOVAs ($p = 0{,}000$ and $p = 0{,}000$) or the univariate ones with Greenhouse-Geisser correction ($p = 0{,}002$ and $p = 0{,}000$).
BOTTOM has 2 levels, so the ANOVA result tells us enough. It's still useful to check out the Estimated Marginal Means though, to see which condition completes the most ORDERS on average. That's the chili peppers condition, as we established already in exercise 1.
MOUTH has 3 levels, so let's continue to figure out which conditions differ exactly. The Estimated Marginal Means show us that the participants completed the most ORDERS on average when they held chili peppers in their mouths (condition 2), followed by lemon (condition 3), and empty mouths came last (condition 1). The pairwise comparisons test which of these differences are significant. They still need a Bonferroni correction: we should divide the significance level by the number of comparisons, here 3, or multiply the p-values by 3. I'm doing the latter in this explanation. Then empty mouths (1) turn out to differ significantly from chili (2), $p = 0{,}000$, and also from lemon (3), $p = 0{,}087$. However, chili and lemon don't differ significantly, $p = 0{,}357$.
In a scientific report, these results are often supplemented with the mean differences and their confidence intervals.

6. B
The t-tests look at the simple effects of BOTTOM peppers, separately per MOUTH condition.

CHAPTER 18 MANOVA AND DISCRIMINANT ANALYSIS (COMPLETE)

1. D

The sample results are expressed most clearly by the means: we can compare the exotic and the Dutch ENVIRONMENT group on each dependent variable. It appears that the average rating was higher in the exotic group for each dependent, although the difference between the TEXTURE means is negligible.
Second, you should only use a MANOVA when the dependents reasonably correlate:

- If they don't, they can be analysed using separate univariate ANOVAs which require no Bonferroni correction.
- If they correlate extremely highly, they measure practically the same thing. In such a case it may be more advisable to calculate a sum score of the dependents and thus combine them into a single variable… which can then be subjected to a univariate ANOVA. (Note: the theory example on Egyptian dating ignores this option.)

Within each group, there appears to be a fair correlation between several dependents. Not all of them are strong, but I think Nagasaki had reason to give a MANOVA a try (particularly because of the correlations found within the exotic group). Note that *Bartlett's Test of Sphericity*, reported just below the correlations, also seems to confirm that the dependents are related ($p = 0{,}000$).
Third, one MANOVA assumption is the equality of covariance matrices. The exotic and Dutch group should have the same covariance matrix, or correlation matrix, between the dependents at population level:

exotic	(TA)	(TE)	(A)	(S)		Dutch	(TA)	(TE)	(A)	(S)
TASTE (TA)	σ^2_{Ta}	σ_{TaTe}	σ_{TaA}	σ_{TaS}		(TA)	σ^2_{Ta}	σ_{TaTe}	σ_{TaA}	σ_{TaS}
TEXTURE (TE)	σ_{TeTa}	σ^2_{Te}	σ_{TeA}	σ_{TeS}	=	(TE)	σ_{TeTa}	σ^2_{Te}	σ_{TeA}	σ_{TeS}
APPEARANCE (A)	σ_{ATa}	σ_{ATe}	σ^2_{A}	σ_{AS}		(A)	σ_{ATa}	σ_{ATe}	σ^2_{A}	σ_{AS}
SMELL (S)	σ_{STa}	σ_{STe}	σ_{SA}	σ^2_{S}		(S)	σ_{STa}	σ_{STe}	σ_{SA}	σ^2_{S}

We can assess the equality of the variances (the diagonal) by comparing the standard deviations of the two groups, on each separate dependent. These look near-identical on the TASTE variable (2,030 versus 2,037) and very similar on the other variables as well.
As for the equality of the covariances (off-diagonal), we can get an idea by comparing the two within-group correlation matrices. I'd say that the two correlation matrices are not completely the same, with particular differences arising for the correlation between TASTE and APPEARANCE (0,359 versus 0,176) and for the correlation between TEXTURE and SMELL (0,188 versus 0,042). The discrepancies are not huge, but there is some room for doubt here. Luckily, at least all correlations point in the same direction (positive) and the exotic and Dutch group both contained 30 participants – an equal number. This made the MANOVA robust against a violation of equal covariance matrices.

2.

The MANOVA shows a highly significant overall ENVIRONMENT effect (Pillai's trace, $p = 0{,}000$). If we move on to the univariate ANOVAs, we should apply a Bonferroni correction to them. An adjusted Bonferroni correction has you multiply the p-values by 3 (the significant MANOVA already indicates that the groups must differ on at least one dependent, and thus controls for one Type I error). If you don't want that, a basic Bonferroni correction multiplies all p-values by 4. Either way, it then comes out that the TASTE, APPEARANCE and SMELL of the sushi are demonstrably influenced by the ENVIRONMENT in which people eat; the sushi's TEXTURE, however, is not.

3.

a) There are 4 dependents, but only 2 groups. Subtract 1 from the number of groups ($i - 1$). Now we have 4 and 1; the smaller of these values is the amount of orthogonal variates that can be created.

b) Let's check the significance of Wilks' Lambda in the discriminant analysis… and the answer is yes ($p = 0{,}000$).

c) We should investigate several results of the discriminant analysis. The *Functions at Group Centroids* tells us the mean score that the groups obtained on the newly created variate. The exotic group has a high positive mean, and the Dutch group has an identical mean but negative. Second, the *Structure Matrix* shows that all dependents except TEXTURE correlate strongly and positively with the variate. Do these bits of information shed light on the identity of the variate, dear reader? What do you think is the property on which the two ENVIRONMENT groups differ the most? My idea is that it's 'openness to experience': a famous personality trait in psychological research. The discriminant analysis supports this idea: the exotic group is more open than the Dutch one (the group centroids), and higher openness correlates positively with higher taste, appearance and smell ratings for the sushi.

d) I would say that the conclusions described above are exactly what Nagasaki was looking for: he was trying to find out if a conservative attitude to food can be influenced from outside. In other words, participants would have to open up to a new food experience. Putting them in an exotic environment seems to do the trick.

4.
I already discussed this in the context of exercise 3: look at the *Structure Matrix*, which gives us the correlations between the variate and the dependents. TASTE, APPEARANCE and SMELL all show a strong correlation with 'openness(?)', but TEXTURE does not.

5. A
Remember, a variate is a new variable which is directly calculated from the original dependents. It's a linear combination – let's say 'package' – of them. The a_{ij} coefficients for the formula are reported in the *Canonical Discriminant Function Coefficients* table (the footnote says they're *Unstandardized coefficients*). So, the formula is

$$V_1 = -3{,}782 + 0{,}417 * TASTE - 0{,}167 * TEXTURE + 0{,}160 * APPEARANCE + 0{,}197 * SMELL$$

Filling that in for the first participant yields

$$V_1 = -3{,}782 + 0{,}417 * 5 - 0{,}167 * 6 + 0{,}160 * 10 + 0{,}197 * 6 = 0{,}083$$

6.

a) The *Descriptives* table shows the means of both groups on the variate. These are the same as the *Functions of Group Centroids* earlier. It confirms that the *Functions at Group Centroids* table gives us the means of both groups on the variate.

b) The eigenvalue of a variate, λ, is that variate's own contribution to the HE^{-1} matrix of the variates. It represents that variate's ratio of explained over unexplained variation, SS(Between)/SS(Within).

$$\lambda_1 = \frac{SS(Between)}{SS(Within)} = \frac{25{,}035}{58{,}000} = 0{,}432$$

Which fits the value from the discriminant analysis! ☺
Do you recall that this is the highest value we could have obtained? The variate that was created is the quantity on which the groups differ the most. Therefore, the group means on this variate lie as far apart as can be managed with Nagasaki's data. Hence, SS(Between) is maximised and SS(Within) is minimised. Their resulting ratio is the variate's eigenvalue, and contributes to (or forms on its own) the MANOVA's equivalent of an F-ratio.

c) Well, why not? Pillai's trace would be

$$V = \sum_j \frac{\lambda_j}{1 + \lambda_j} = \frac{0{,}432}{1 + 0{,}432} = 0{,}302$$

(There's only one variate in this case, so we didn't really need the sum sign.)

d) This would've added one group to the experiment. 4 dependents, 3 groups minus 1 equals 2; hence, an extra variate would have come into play.

e) The second variate is always the best variate once the first one has been constructed. Therefore, the first variate will always discriminate better between the groups, and have a higher eigenvalue. However, there's no telling what will happen to the first variate's eigenvalue, currently 0,432, when we add a third group with its own data. It could decrease, but also increase. So the answer is that, strictly speaking, we can't say in advance. (Though I'd like to add that, I think, a second variate with a strong eigenvalue is not so likely for this study.)

f) Not at all: the variates are created as independent quantities.

CHAPTER 19-22 CATEGORICAL TESTS (COMPLETE)

1.

a) We now look at two categorical variables: the location where the beachgoers spent most of their time, and whether they wore shoes. A χ²-test for contingency tables is then suited in any case. Both variables are dichotomous, so the contingency table has $(2-1)*(2-1)=1$ degree of freedom. That's why a z-test for 2 proportions is also an option.

b) **Z-test for 2 proportions**

$$H_0: \pi_{beach} = \pi_{beach\ house}$$
$$H_A: \pi_{beach} \neq \pi_{beach\ house}$$

π is here the proportion of people who walked barefoot.
The design assumptions are met (dependent variable dichotomous, independent groups). Normality is automatically violated, but the samples are quite sufficient in size.
So let's perform:

$$p_1 = \frac{121}{187} = 0{,}647$$
$$p_2 = \frac{22}{48} = 0{,}458$$
$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{121 + 22}{187 + 48} = 0{,}609$$
$$Z = \frac{p_1 - p_2}{\sqrt{\pi(1-\pi)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0{,}647 - 0{,}458}{\sqrt{0{,}609 * 0{,}391} * \sqrt{\frac{1}{187} + \frac{1}{48}}} = \frac{0{,}189}{0{,}488 * 0{,}162} = 2{,}39$$

The z-table in the appendix tells us that

$$p = 2 * P(Z > 2{,}39) = 2 * (1 - 0{,}9916) = 2 * 0{,}0084 = 0{,}0168$$

Don't forget to double the p-value if you perform a two-tailed test like me.
Enfin, the outcome is significant: people on the beach demonstrably walk barefoot more often.

χ²-test for contingency tables

$$H_0: no\ association\ location \times footwear$$
$$H_A: yes\ association\ location \times footwear$$

The design assumptions are met (dependent variable categorical, independent groups). We're about to find out if the Expected Counts are large enough, but considering the large samples that's pretty likely beforehand.
So let's perform. First we make the contingency table:

Observed		Spent the day where?		
		at the beach	at the beach house	
Walked around how?	barefoot	121	22	**143**
	in shoes	66	26	**92**
		187	**48**	**235**

We calculate the Expected Counts using the formula $EC = \frac{row\ total * column\ total}{N}$. All of them turn out larger than 5 indeed:

Expected		Spent the day where?		
		at the beach	at the beach house	
Walked around how?	barefoot	113,8	29,2	**143**
	in shoes	73,2	18,8	**92**
		187	**48**	**235**

Now for the test statistic:

$$\chi^2 = \sum \frac{(OC - EC)^2}{EC} =$$
$$\frac{(121 - 113{,}8)^2}{113{,}8} + \frac{(66 - 73{,}2)^2}{73{,}2} + \frac{(22 - 29{,}2)^2}{29{,}2} + \frac{(26 - 18{,}8)^2}{18{,}8} =$$
$$0{,}456 + 0{,}708 + 1{,}775 + 2{,}757 = 5{,}696$$

Check it: $Z^2 = \chi^2$. Yes indeed: the z-test and the χ²-test are in this case exactly the same – or, with a difficult word, data-equivalent. We can thus use them both, in case the contingency table has only 1 degree of freedom! You can read more about the equivalence in bridge chapter 29 of the handbook.
Anyhow, the χ²-table in the appendix tells us (look at the row of 1 degree of freedom) that

$$5{,}412 < \chi^2 < 6{,}635$$
$$0{,}01 < p < 0{,}02$$

So the outcome is at least significant: people on the beach walk demonstrably more often on bare feet. That sounds neater too: you'll wear something inside a beach house, won't you?

c) That's only automatically possible if you chose a z-test!
The confidence interval is

$$p_1 - p_2 \pm Z^* \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$
$$0{,}647 - 0{,}458 \pm 1{,}96 * \sqrt{\frac{0{,}647 * 0{,}353}{187} + \frac{0{,}458 * 0{,}542}{48}}$$
$$0{,}189 \pm 1{,}96 * \sqrt{0{,}00122 + 0{,}00517}$$
$$0{,}189 \pm 0{,}157$$
$$[\,0{,}032\,;0{,}346]$$

So, the difference in proportion of barefoot walkers probably amounts to between 3,2 and 34,6 percent in the population.

2.
We have a single variable with more than 2 categories. The table has $3 - 1 = 2$ degrees of freedom; a χ^2 Goodness of Fit Test is thus the only alternative.
In total there are 235 participants ($N = 235$). We calculate the Expected Counts using the formula $EC_i = N * \pi_i$:

$$EC_{yes} = 235 * 0{,}3 = 70{,}5$$
$$EC_{a\ little} = 235 * 0{,}3 = 70{,}5$$
$$EC_{no} = 235 * 0{,}4 = 94$$

These are all amply large.
Now we can calculate the test statistic:

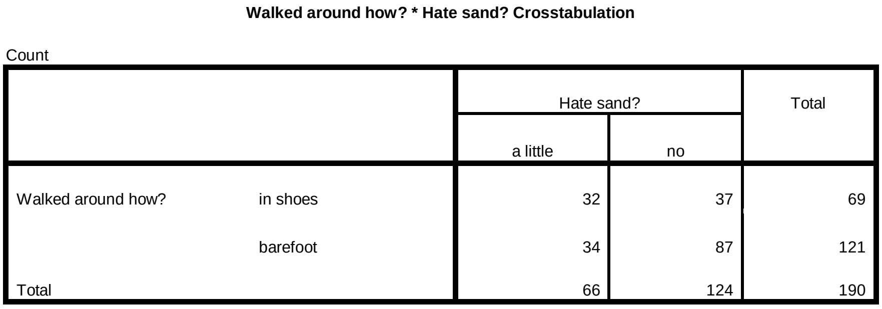
$$\chi^2 = \sum \frac{(OC - EC)^2}{EC} = \frac{(45 - 70{,}5)^2}{70{,}5} + \frac{(66 - 70{,}5)^2}{70{,}5} + \frac{(124 - 94)^2}{94} =$$
$$9{,}22 + 0{,}29 + 9{,}57 = 19{,}08$$

At 2 degrees of freedom this value is fiercely significant: larger than the largest value in the appendix table, so the p-value is even smaller than 0,0005. In Moonshine the beachgoers clearly like sand more often.

3.

a) Since all the Observed Counts are much larger than 5, so will the Expected Counts be (we can thus take a small shortcut). For the test result we look into the Chi-Square Tests table: the Pearson Chi-Square has a p-value of 0,008 and is as such convincingly significant. Depending on their hate of sand, Moonshine's beachgoers don't walk on bare feet equally often.

b) So yes: pairwise comparisons! With Bonferroni correction. The strange thing is that these are not featured in SPSS for contingency tables. That's why nobody does them, although it would of course be good practice… As an advocate of solid statistics (you should not follow the masses!), I will now do the pairwise comparisons anyway:



Walked around how? * Hate sand? Crosstabulation

Count

		Hate sand?		Total
		a little	no	
Walked around how?	in shoes	32	37	69
	barefoot	34	87	121
Total		66	124	190

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,475[a]	1	,011		
Continuity Correction[b]	5,694	1	,017		
Likelihood Ratio	6,392	1	,011		
Fisher's Exact Test				,017	,009
Linear-by-Linear Association	6,441	1	,011		
N of Valid Cases	190				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 23,97.

b. Computed only for a 2x2 table

After Bonferroni correction we get $p = 0{,}033$: significant. People with a little hate of sand walk barefoot less often than people with no hate.

Walked around how? * Hate sand? Crosstabulation

Count

		Hate sand?		Total
		yes	no	
Walked around how?	in shoes	23	37	60
	barefoot	22	87	109
Total		45	124	169

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,525[a]	1	,011		
Continuity Correction[b]	5,629	1	,018		
Likelihood Ratio	6,356	1	,012		
Fisher's Exact Test				,017	,009
Linear-by-Linear Association	6,486	1	,011		
N of Valid Cases	169				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 15,98.

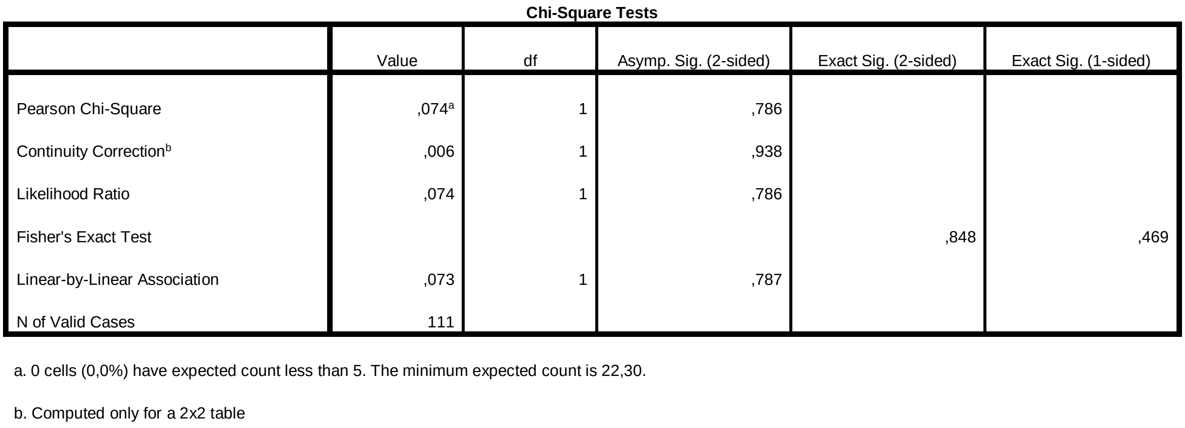
b. Computed only for a 2x2 table

After Bonferroni correction it turns out that $p = 0{,}033$: significant. People with an explicit dislike of sand walk barefoot less often than people without a dislike.

Walked around how? * Hate sand? Crosstabulation

Count

		Hate sand?		Total
		yes	a little	
Walked around how?	in shoes	23	32	55
	barefoot	22	34	56
Total		45	66	111



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,074[a]	1	,786		
Continuity Correction[b]	,006	1	,938		
Likelihood Ratio	,074	1	,786		
Fisher's Exact Test				,848	,469
Linear-by-Linear Association	,073	1	,787		
N of Valid Cases	111				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 22,30.

b. Computed only for a 2x2 table

After Bonferroni correction you've got $p = 1$ (a p-value can't exceed 100%). So, people with a clear and a little hate of sand probably walk barefoot equally often.

4.
The number of bought ice creams is quantitative. A χ^2-test is inadequate for this. She'll need an independent samples t-test or a one-way ANOVA.

5. C
Indeed there are more debutants who use product placement, but the total sample counts a lot more debutants too: as if that makes for a nice and fair comparison! So A is wrong. B gets closer: if the relative frequency distributions differ from each other, that means that the proportion of product placers is unequal between debutants and established authors. This is the case (25% product placers among the debutants versus a mere 12,5% among the established authors), so the two distributions differ… at sample level. The question if whether this difference doesn't take place just due to chance, and if these samples were drawn from two equal population distributions in which the same amount of product placement is employed. That's why B is wrong as well, and that renders C automatically correct. If you want confirmation, you can proceed, though this is not strictly necessary for the exercise.
In that case we test the null hypothesis – with a χ^2-test of course. First we calculate the Expected Counts.

	Debutant	Established author	
Uses no product placement	45 (48)	35 (32)	80
Uses product placement	15 (12)	5 (8)	20
	60	40	100

Next: $\boldsymbol{\chi^2} = \frac{(45-48)^2}{48} + \frac{(15-12)^2}{12} + \frac{(35-32)^2}{32} + \frac{(5-8)^2}{8} = \mathbf{2,34}$
In the χ^2-table we find that at 1 degree of freedom $0{,}10 < P(\chi^2 > 2{,}34) < 0{,}15$. The p-value is thus larger than 0,05 in any case.[7] In other words, the researcher does not reject the null hypothesis, but it's possible that he makes a Type II error. (This is always possible when we don't reject a null hypothesis. Of course, this error is not likely when the p-value is very high.)
Another option would have been to use a z-test for 2 proportions. Its advantage is that the researcher could've performed a one-tailed test. This would still not have been significant though; the p-value (which is half as large) stays too large.

6. A
The Expected Counts of that third author category are far too small (they must be at least 5). Go ahead and try to calculate them!

[7] By the way, SPSS tells us: $p = 0{,}126$.

CHAPTER 23 CONTINGENCY TABLE ANALYSIS (COMPLETE)

1.

a)

Always code ‘no’ as 0, and ‘yes’ as 1. In that case ‘no’ should be on the left or at the top each time, and ‘yes’ should be on the right or at the bottom.

	low SES		high SES	
	no paper towels	paper towels	no paper towels	paper towels
unhappy	49	32	40	26
happy	96	51	91	43
	145	**85**	**131**	**69**

b) Using this properly structured contingency table, we can calculate odds ratios without fear of making errors. To be on the safe side (and avoid confounding) let’s look at the simple effects of PAPER TOWELS.

- Low SES: $OR_{HAPPY*PAPER\,TOWELS} = \frac{A*D}{B*C} = \frac{49*51}{32*96} = 0{,}81$
- High SES: $OR_{HAPPY*PAPER\,TOWELS} = \frac{40*43}{26*91} = 0{,}73$

All in all we see an effect in both groups: the odds ratio is below 1, so there’s a negative association. The use of paper towels appears to lower your odds of being happy. Mind: this may have resulted from sampling error (or from background variables altogether, since this is just an observational study).

c) To be on the safe side (and avoid confounding) let’s look at the simple effects of SES. Take care to pick the right cells from the contingency table. If necessary, cover the columns you’re not using for the moment!

- No paper towels: $OR_{HAPPY*SES} = \frac{A*D}{B*C} = \frac{49*91}{40*96} = 1{,}16$
- Paper towels: $OR_{HAPPY*SES} = \frac{32*43}{26*51} = 1{,}04$

All in all we see an effect in both groups: the odds ratio is slightly above 1, so there’s a positive association. People with a high SES are happy slightly more often. Mind: this may have resulted from sampling error (or from background variables altogether, since this is just an observational study).

d) The fact that the simple effects (the odds ratios) differ each time means that there’s interaction: the effect of PAPER TOWELS on HAPPY depends partly on SES. Automatically, then, the effect of SES on HAPPY depends partly on PAPER TOWELS (interaction is symmetric, as we call it).

2. C

Answer A claims that the interaction effect in the sample is limited. This may be the case, but interaction is wholly unrelated to confounding!

Whether the corrected main effect of PAPER TOWELS is large or small is somewhat up for debate. I don’t find it that big, dear reader. The discussion doesn’t matter though: if we had not corrected for confounding, we might have found a much larger effect of PAPER TOWELS or even none at all! The whole point is whether the correction for confounding changes your results drastically or not. This renders B false as well.

A more direct way to map the risk of confounding is this: make a separate contingency table of the two independent variables. Ignore the dependent variable HAPPY for now.

		SES	
		low	high
PAPER TOWELS	no	145	131
	yes	85	69

The odds ratio turns out to equal

$$OR_{PAPER\,TOWELS*SES} = \frac{145*69}{131*85} = 0{,}90$$

The – negative – relationship is thus not very strong (1 means no relationship). People with a high SES in this sample used paper towels slightly less often, but the balance isn’t that lopsided. Confounding, if any, will therefore be limited. The correct answer is C.

3. A

Answer A and B both pertain to the Mantel-Haenszel Common Odds Ratio Estimate. So the question is: which relationship does this odds ratio express exactly? SES is the covariate and functions as a control variable; it’s mentioned multiple times on the left side of the tables. The effect of the control variable is not tested! For this reason we must be seeing the odds ratio for the relationship between HAPPY and PAPER TOWELS. Answer A is right.

C points us toward the odds ratio we find at ‘SES: low’ in the Risk Estimate table. This is the *Odds Ratio for PAPER TOWELS*. The relationship C suggests is nonsensical, however: if we only study people with a low SES, then SES is no longer a variable, is it? Relationships can only exist between variables. When you say: 'poor people use paper towels more often…' in fact you mean '… than rich people'. In that case you're speaking of rich people as well after all! The odds ratio reported here is in fact the odds ratio for the relationship between HAPPY and PAPER TOWELS, for people with a low SES. And indeed: you calculated the value 0,813 yourself in exercise 1.

4. B

Again SES is the control variable; the effect of this variable cannot be tested. In the *Total* block, poor and rich people are lumped together (we don't add them up really, but rather we simply ignore their SES) and so the variable is in fact even thrown into the rubbish bin. In that case it must be the effect of PAPER TOWELS being tested. Since we no longer take into account the fact that the participants also differed in terms of SES, SES effects might confound this test.

5. B

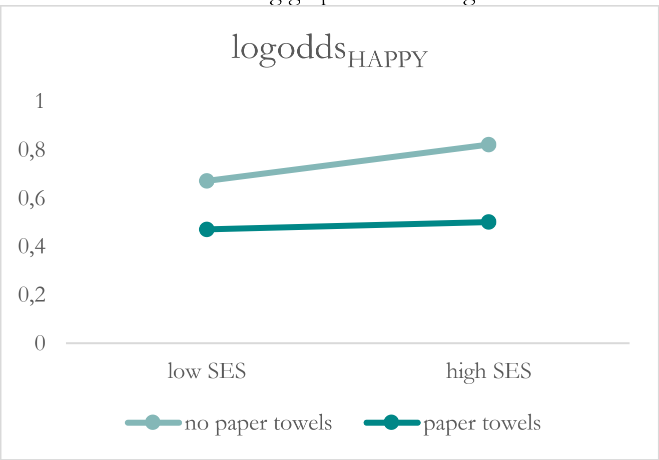
For the last time: the effect of SES isn't tested in this output. A can therefore be dropped. In the sample we do see the interaction effect C mentions, but this effect is not significant (*Tests of Homogeneity of the Odds Ratio*: $p = 0{,}791$). With that B remains, and indeed: the *Tests of Conditional Independence* are not significant ($p = 0{,}222$), so PAPER TOWELS does not demonstrably relate to HAPPY.

6.

Here's the contingency table again, now with all the odds and logodds included:

	low SES		**high SES**	
	no paper towels	**paper towels**	**no paper towels**	**paper towels**
unhappy	49	32	40	26
happy	96	51	91	43
	145	**85**	**131**	**69**
odds	$\left(\frac{96}{49}\right) = 1{,}96$	$\left(\frac{51}{32}\right) = 1{,}59$	$\left(\frac{91}{40}\right) = 2{,}28$	$\left(\frac{43}{26}\right) = 1{,}65$
log *odds*	$\mathbf{0,67}$	$\mathbf{0,47}$	$\mathbf{0,82}$	$\mathbf{0,50}$

We can make the following graph of these logodds:



The slope of the line is consistently equal to the natural logarithm of the odds ratio (which we calculated in exercise 1c):

- No paper towels: $\ln OR = \ln 1{,}16 = 0{,}15 = 0{,}82 - 0{,}67$
- Paper towels: $\ln OR = \ln 1{,}04 = 0{,}03 = 0{,}50 - 0{,}47$

Chapter 24 Simple Regression (complete)

1. B

This person causes the scores to fit less well in general with the (distorted) line, which results into bigger residuals. Such residuals constitute unexplained variation (why does your vocabulary deviate from the prediction?) and the correlation gets smaller. Due to this outlier, then, we can explain the variation in the usage of English words less well.

2. A

The *Model ANOVA* can be used as well to determine if reading Dutch books has an effect; in this simple regression model, it yields the exact same result as the *Coefficients* t-test! And it turns out to be significant ($p = 0{,}029$). So yes, reading Dutch books appears to have an effect. The participants' scores decline indeed (see the slope b_1), but that was supposed to happen actually: the use of English terms and anglicisms decreases!

3.

This design is observational. Ferry Ironisch didn't instruct his participants at random to read few or many Dutch books. So, if those who read more use fewer English terms and anglicisms, this may be related to other factors. The risk of confounding is realistic!

4.

The standard error is

$$s_b = \frac{s_{est}}{\sqrt{\sum_i (X_i - \bar{X})^2}} = \frac{15{,}296}{\sqrt{178{,}74}} = 1{,}144$$

For the 95% confidence interval, we must first look up the critical t-value in the Appendix. At 25 degrees of freedom it turns out to equal 2,060. With that the interval becomes

$$b \pm T^* s_b$$
$$-2{,}656 \pm 2{,}060 * 1{,}144$$
$$-2{,}656 \pm 2{,}357$$
$$[\,-5{,}013\,;\; -0{,}299\,]$$

Save for a minuscule decimal (a rounding error) this agrees with the interval in the SPSS output.

CHAPTER 25 DICHOTOMOUS AND DUMMY VARIABLES (COMPLETE)

1.

a) Group 1 and 4 have a reasonably high sample mean. The atmosphere tends to be clearly lower in the other two groups.

b) The standard deviation(!) is considerably higher within group 2 and 3. It would seem that hosts react a lot more dividedly when the guests take own initiative or make an explicit request: some hosts don't really mind, but other ones are gravely offended by such a move. Silence or a subtle suggestion (a more careful attempt to get some snacks served) are responded to in more similar ways.

c) The dependent variable is quantitative, and the groups are independent. Normality cannot be checked with the current output, but the samples are fairly large ($n = 19$), so as long as no strong skewness or outliers are present, the regression analysis should be robust against a violation. Finally, we already established in question b that the variances don't appear equal. The rule of thumb also gives us the red light: $\frac{largest\ s}{smallest\ s} = \frac{31,291}{14,190} = 2,205 > 2$. However, each sample has the same size. The regression analysis is therefore robust against a violation of equal variances (homoscedasticity).

2.

a) This is reference coding. There's one group that scores 0 on all the dummies, and the other groups score 1 on a single, consistently different, dummy.

b) Yes: the group with all zeroes is the reference group – number 4. This implies that each dummy variable will compare the mean of the silence group with the mean of one other group.

c) I find it logical: silence could be seen as the control condition, in which the participants do nothing to change the catering situation. The other three groups all resort to active behaviour.

d) Up to you, dear reader. One suggestion of my own: perhaps contrast coding might be interesting, for instance, comparing just group 1 and 3 to see if making a request implicit or explicit is the better thing to do.

3. C

Each group can be identified by its scores on the dummies; you can see in the coding scheme that no two groups have the same set of dummy scores. This makes them mathematically distinguishable. If we included a fourth dummy after all, it would be completely **collinear** with the others (see chapter 26-A); this would render the analysis impossible. (SPSS will automatically throw a superfluous dummy out.)

4. C

What is asked for is a comparison between group 2 and 3. However, group 4 is the reference group in the current coding scheme. The dummy variables thus compare group 1, 2 and 3 against group 4; they cannot make any other comparisons.

5.

Even a subtle suggestion…: FALSE. Does group 1 differ significantly from group 4? dummy1 makes the comparison. Its regression coefficient is -2,579 (check the *Coefficients* table), which tells us that the average ATMOSPHERE is slightly worse when a subtle suggestion is made. Yet, this value does not differ significantly from 0 ($p = 0,744$).

Never bring your own…: TRUE. As the *Coefficients* table shows, the regression coefficient of dummy2 is -22,474. The ATMOSPHERE appears to worsen strongly in the case of own initiative. And indeed, this effect is significant ($p = 0,006$).

'I want' never gets…: TRUE. An explicit request lowers the average atmosphere by a point estimate of -20,684, as the b of dummy3 indicates. This decrease is significant ($p = 0,011$).

Speech is silver, silence…: FALSE. This is more or less the same statement as the first one, just phrased differently. Silence makes for a fair atmosphere, but making a subtle suggestion (speech) doesn't seem to hurt.

6. D

There's a Model ANOVA in the output, so Bucket already made sure she had that one provided with the regression analysis. In the background, an ANOVA always uses effect coding, which is different from the reference coding scheme that Bucket employed this time. Her conclusions would not have changed: she'd have found the same differences between the groups, as well as the same p-values when conducting pairwise comparisons. The only possible change is that she might have performed a Bonferroni correction. This would only have been necessary in case she decided to inspect all pairwise comparisons, rather than just the three that are present in her current analysis (merely comparisons with the silence group).

Wanna know more? Have a look at chapter 37, paragraph 37.2, and bridge the gap between ANOVA and regression! ☺

CHAPTER 26-A MULTIPLE REGRESSION: MAIN EFFECTS (COMPLETE)

1.

a) $\hat{Y} = 20{,}769 + 0{,}195 * aggression - 0{,}431 * pressure - 0{,}141 * AAAH$!

b) This is the predicted cortisol concentration (20,769 micrograms per decilitre) for a man who's totally not aggressive (0), has a shower without any water pressure (0) and can turn the knob 0 degrees before his water goes from shivering cold to boiling hot. Would someone like that exist? If so, he had better buy a new shower.

c) If a person scores 1 point higher on the aggression scale, his cortisol concentration is expected to rise by 0,195 micrograms per decilitre. This is the expected rise when we keep pressure and AAAH! constant, so when we assume that this slightly more aggressive person does have the same water pressure in his shower and needs to turn equally far to drastically change the temperature.

2. C

The unstandardised regression coefficients are scale-dependent. Look therefore at the standardised ones instead. There AAAH! turns out to have the regression coefficient which deviates the most from zero.

3. B

The predictor pressure has a tolerance value which almost equals 1: it's thus hardly related to aggression and AAAH! at all. Aggression does relate substantially to the other two predictors, because the tolerance is 0,608, which means the proportion of explained variation in this predictor equals $1 - 0{,}608 = 0{,}392$. This association has to be with AAAH! in that case.

4.

a) Outliers in the Y direction: yes, because at least one *Studentized Residual* is greater than 3.
Outliers in the X direction: no, since the maximum allowed *Centered Leverage Value* is $\frac{3(p+1)}{N} = \frac{3*4}{30} = 0{,}40$ and the biggest *CLV* is 0,25.
Influential cases: yes, because at least one *Cook's Distance* exceeds 1.

b) We've got at least one extreme residual: a person who has a much higher cortisol level than we'd expect on the basis of his aggression, water pressure and AAAH!. Judging from the *Cook's Distance*, this person pulls the regression line toward himself; he distorts the analysis. (Note: I'm suspecting that it's the same individual.)

c) The relationship between pressure and cortisol doesn't appear linear… rather, it's quadratic! That makes much more conceptual sense too: a bit of pressure on the water is pleasant, but a rock-hard jet will enhance stress levels sooner than decrease them. Due to this the relationship between pressure and cortisol hasn't been modelled quite well in the regression analysis. Some distortion is the consequence: the persons in the upper right corner pull the regression line toward themselves.

5. B

Aggression has no demonstrable effect on the cortisol level, so the variation that aggression explains at sample level is probably due to chance. This means that R^2 won't decrease significantly if we exclude aggression. By studying the effects of pressure and AAAH! only, we describe the population (reality) equally well.

6. B

A backward procedure makes you end up at model 2. That's because aggression comes out as non-significant in model 3 ($p > 0{,}10$) and may therefore get out. In model 2 pressure still isn't quite significant, but to avoid Type II errors in backward, it's common to take out predictors only if the p-value is greater than 0,10. Hence I would pick the model with pressure and AAAH! as the final model. The conclusion that follows from this is described in answer B.

CHAPTER 27 LOGISTIC REGRESSION (COMPLETE)

To clarify the solutions, here's a contingency table of the data; you could make it if McDuck's study was your own, since all variables are dichotomous.

	pigeon		headless lark	
	normal window	tinted window	normal window	tinted window
no crash	38	48	19	23
crash	22	12	27	23
	60	**60**	**46**	**46**

1.

a) Yes: the Step and Block chi-square tests are significant ($p = 0{,}049$) in the *Omnibus Tests* table, which indicates that model 1 gives a better view of the population than the empty model 0. Also, in the *Variables in the Equation* table, the Wald test for the regression coefficient (b) of WINDOW is significant ($p = 0{,}050$). Although this test is slightly less reliable than those found in the *Omnibus Tests*, it says the same: the window has an effect.

b) First off, Model 1's result may be a Type I error: the p-value of the WINDOW effect is (near-)equal to the 5% significance level α, which makes the test barely significant. It's plausible that we falsely reject the null hypothesis of no relationship.
In addition, the conclusion which follows from mode 1 may be invalid because of confounding. Perhaps the birds exposed to a normal window were mostly headless larks, while the birds exposed to a tinted window were mostly pigeons. This could lead to confounding. It's not a very likely scenario though, as professor McDuck had the chance to assign each type of bird to each type of window equally often – he could make the predictors nicely orthogonal. That's unless, for various reasons, some birds unexpectedly drop out. ☺
Finally there may be interaction, which model 1 ignores: the effect of the window could depend on the bird species (for instance, headless larks might not see either of the two windows while pigeons see just one).

c) Imagine (or draw) a simple graph with the logodds on the Y axis and WINDOW on the X axis. 0 on the X axis indicates a normal window, 1 a tinted window. The b of WINDOW tells us how strongly the logodds of a crash will rise when we increase the predictor level by 1 point, so when we go from normal to tinted window. In that case the logodds will decrease by 0,556. The logodds of a crash are thus lower with tinted windows, and so the probability of a crash is smaller. This makes the normal window the most dangerous for the birds.
Note that the odds ratio (see the *Exp(B)* value) expresses a negative relationship as well.

2. A

Well, you'd like to look at the model 3 output and use the Step or Block chi-square test from the *Omnibus Tests* table, right? Only the p-value has been wiped. Blimey! Let's use the Wald interaction test from the *Variables in the Equation* table then… crap – that p-value has been wiped too! Is there an alternative way to test for interaction? There's one: the Hosmer and Lemeshow test for model 2. The null hypothesis that this model is complete – and that we therefore need not add a non-linear term or interaction term – cannot be rejected ($p = 0{,}713$). This is why professor McDuck probably didn't find a significant interaction effect in block 3 either.
So what makes the other answers false?
The fact that the interaction term in model 3 isn't zero could be due to chance – sampling error! It's the classic problem: it takes a statistical test to prove that there's a true interaction effect in the population. That makes B a no-go area.
Lastly, it's true in fact that model 3 tests a simple effect of WINDOW (for pigeons) and a simple effect of SPECIES (with a normal window). See the discussion about the meaning of b_1 and b_2 when we include an interaction term, in paragraph 27.5. These simple effects are demonstrated too (the tests are significant). The presence of simple effects doesn't confirm nor debunk interaction though. The question is if the other simple effect of WINDOW respectively SPECIES is the same; this is expressed by b_3, which equals 0 when there's no interaction. C is wrong as well.

3. C

$$\log odds = b_0 + b_1 * WINDOW + b_2 * SPECIES + b_3 * WINDOW * SPECIES$$

When we fill in the bs (using the *Variables in the Equation* table), we obtain:

$$\log odds = -0{,}547 - 0{,}840 * WINDOW + 0{,}898 * SPECIES + 0{,}488 * WINDOW * SPECIES$$

Now we can calculate the logodds of a crash for each situation. The higher the logodds, the greater the probability of a crash.

SITUATION	WINDOW	SPECIES	$\log odds =$	$\log odds =$
Normal window, pigeon	0	0	b_0	$\mathbf{-0,547}$
Tinted window, pigeon	1	0	$b_0 + b_1$	$\mathbf{-1,386}$
Normal window, headless lark	0	1	$b_0 + b_2$	$\mathbf{0,351}$
Tinted window, headless lark	1	1	$b_0 + b_1 + b_2 + b_3$	$\mathbf{0}$

And the winner is… the headless lark with a normal window!

4. B

The natural logarithm of an odds ratio is equal to a b or a combination of bs, so we should seek the correct exponent one and raise e to its power. You can make graphs of the logodds – that clarifies it the most – but it also works with the table above:

- The last two rows pertain to the headless larks;
- The difference between those two rows is a normal versus a tinted window;
- The difference of bs is $b_1 + b_3$. So, this is the (simple) effect of the window for headless larks;
- The exponent of these bs is the odds ratio: $e^{b_1+b_3} = e^{-0,840+0,488} = \mathbf{0,703}$.

<u>Don't</u> take the exponent of b_1 only; this is the (simple) effect of the window for pigeons! You'll end up at answer A then (which is also displayed under *Exp(B)* in the output) and that's incorrect.

5. A

Model 3 is not parsimonious enough, as we established earlier; there's no interaction effect, so D (the definition of interaction) can be scratched. The other answers deal with a main effect of the WINDOW. Check out model 2 (which is significantly better than model 1): the b of WINDOW equals -0,599. That's negative, which means because of the coding that the tinted window delivers the lowest logodds of a crash. Both pigeons and headless larks crash into a tinted window less often than into a normal window.

6. B

The *Mantel-Haenszel Common Odds Ratio Estimate* is a kind of weighted average of the simple effects. By calculating an average, you assume that the simple effects differ purely due to chance and that there's no interaction. Hence, we seek the corrected main effect of BIRD in a model without an interaction term. This is model 2 of course. We find the odds ratio for the relationship between CRASH and BIRD under *Exp(B)*: 3,083.

7.

a) $OR = e^{0,116} = 1,12$

b) They grow 1,12 times as large. I'm just asking for the meaning of the odds ratio! ☺

c) This is a rise of the logodds with $4 * b$, so $OR = e^{4*0,116} = 1,59$.

CHAPTER 29 RELIABILITY (COMPLETE)

1.

a) I think not: I believe that all items measure how the scientist approaches the balance between accessible education on the one hand, and serious and factually sound education on the other hand. Except item IV: it rather measures how much the scientists like their own profession.

b) Yes: have a look at the *Item-Total Statistics*. The *Corrected Item-Total Correlation* of item IV is very low (even negative, which is officially not allowed for a reliability analysis). And if we removed this item, Cronbach's alpha would rise *(Cronbach's Alpha if Item Deleted)*.

c) The items aren't unidimensional (together they measure more than a single personality trait) and so they're not parallel either.

d) The consequence of this is that the items will correlate less on average. We use the average correlation to estimate the reliability of the items, so that value will be an underestimation. Cronbach's alpha will also underestimate the reliability of the questionnaire as a whole. The sum scores are actually more reliable than the analysis suggests.

2. C

Check out the *Item Statistics*. Since the mean of item IV is clearly higher, this item probably doesn't measure the same true score T. The assumption of parallelism is violated in that case; hence argument A.
The standard deviation of item IV is very limited too (argument B). Nearly all scientists must have indicated 4 or 5 points – they (fully) agree with the statement that molecules are immensely cool. Would you expect any different from a chemical scientist? Thus this item fails to discriminate the participants properly; it cannot tell us how they differ in terms of their passion for the field. It turns the item into a pretty useless psychometric instrument.

3. B

Cronbach's alpha *(Reliability Statistics)* is too low; it should be 0,70 at least, and preferably 0,80. The researchers might achieve this by extending the questionnaire with more (parallel) items. The current items – save for number IV – appear pretty parallel and each one correlates nicely with the rest, so I wouldn't call them a train wreck: they're building toward a reliable image of scientists' didactic convictions.

4. C

Study the *Item-Total Statistics: Cronbach's Alpha if Item Deleted*. Without item IV Cronbach's alpha becomes 0,629.
Now we might first calculate the estimated reliability per item by solving the Spearman-Brown formula for Cronbach's alpha, and then fill in this formula again for $k = 10$. I'll do this directly below, but only to show you that there's a faster method – which works as well!
When $k = 4$:

$$\alpha = \frac{k * \bar{r}_{item,item'}}{1 + (k-1)\bar{r}_{item,item'}}$$

$$0{,}629 = \frac{4 * \bar{r}_{item,item'}}{1 + 3 * \bar{r}_{item,item'}}$$

$$0{,}629 * (1 + 3 * \bar{r}_{item,item'}) = 4 * \bar{r}_{item,item'}$$

$$0{,}629 + 1{,}887 * \bar{r}_{item,item'} = 4 * \bar{r}_{item,item'}$$

$$0{,}629 = 2{,}113 * \bar{r}_{item,item'}$$

$$\bar{r}_{item,item'} = \frac{0{,}629}{2{,}113} = 0{,}298$$

When $k = 10$:

$$\alpha = \frac{k * \bar{r}_{item,item'}}{1 + (k-1)\bar{r}_{item,item'}} = \frac{10 * 0{,}298}{1 + 9 * 0{,}298} = \mathbf{0,809}$$

The fastest and easiest method, however, is to state that the number of items (k) would increase from 4 to 10. This means it would get 2,5 times as large $\left(\frac{10}{4}\right)$. Bizarrely enough, you can now simply take the Spearman-Brown formula and fill in $k = 2{,}5$ and the current reliability 0,629!

$$\boldsymbol{\rho_{10\ items}} = \frac{k * \rho_{4\ items}}{1 + (k-1)\rho_{4\ items}} = \frac{2{,}5 * 0{,}629}{1 + 1{,}5 * 0{,}629} = \mathbf{0,809}$$

5. C

The mean sum score of the twenty participants equals 16,45 (*Scale Statistics: Mean*).
A is the wrong answer anyway: Bacarra's higher (measured) score could also have resulted from measurement errors. This is why we should make a **confidence interval** for the true difference between the two sum scores.
The participants' sum scores have a variance of 10,576 (see the *Scale Statistics: Variance*). This measured variance consists of measurement errors to a large extent: Cronbach's alpha is 0,569, so 56,9% of the measured dispersion represents

true differences between the participants, and $100\% - 56{,}9\% = 43{,}1\%$ represents differences due to errors. The error variance of an individual sum score is therefore (estimated to be)

$$\sigma^2_{e_{sum}} = (1 - \rho_{SS'}) * \sigma^2_{sum} = (1 - 0{,}569) * 10{,}576 = 0{,}431 * 10{,}576 = 4{,}558$$

However, the error variance of the <u>difference</u> between two sum scores is even twice as large:

$$\sigma^2_{e_{sum\ difference}} = 2 * 4{,}558 = 9{,}116$$

The estimated standard error of measurement equals the square root of this:

$$SEM = \sqrt{\sigma^2_{e_{sum\ difference}}} = \sqrt{9{,}116} = 3{,}02$$

Assuming that the measurement errors are random (their mean is 0) and normally distributed, we can state that 95% of all the errors made when calculating the difference between two sum scores fall between

$$[\,-2 * 3{,}02\,;\ 2 * 3{,}02\,] = [\,-6{,}04\,;\ 6{,}04\,]$$

This means that the measurement error on the difference between Bacarra's and Spencer's sum score is probably (with 95% confidence) no bigger than 6,04. In short, the true difference can lie anywhere between

$$23 - 19 \pm 6{,}04$$
$$4 \pm 6{,}04$$
$$[\mathbf{-2{,}04};\mathbf{10{,}04}]$$

It appears the true difference can also amount to 0 points. Thus, Hiroshima and his colleagues can't prove that Bacarra is truly more progressive than Spencer. Perhaps this would've been possible with a more reliable questionnaire.

6.

a) Let's use the attenuation formula for this. The reliability of the questionnaire is still estimated at 0,569. Consider the sum scores on the questionnaire to be X and the mean ratings of the handbooks to be Y.

$$\boldsymbol{\rho_{T_X T_Y}} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} = \frac{0{,}32}{\sqrt{0{,}569 * 0{,}88}} = \mathbf{0{,}45}$$

b) Not necessarily: the correlation isn't 1 after all. A positive trend does present itself: handbooks written by a progressive professor tend to be preferred by students in general.

CHAPTER 30 AGREEMENT

1.
For starters,

$$A_o = 1 + 2 + 5 + 3 + 3 = 14$$

If we're smart, we'll limit ourselves to calculating the expected agreements (the diagonal). The respective Expected Counts follow:

- $EC_{red,red} = \frac{5*10}{40} = 1,25$
- $EC_{orange,orange} = \frac{10*5}{40} = 1,25$
- $EC_{green,green} = \frac{5*15}{40} = 1,875$
- $EC_{blue,blue} = \frac{15*5}{40} = 1,875$
- $EC_{purple,purple} = \frac{5*5}{40} = 0,625$

In the contingency table:

Expected Count		ARYOO					
		red	orange	green	blue	purple	
NUTS	red	1,25					**5**
	orange		1,25				**10**
	green			1,875			**5**
	blue				1,875		**15**
	purple					0,625	**5**
		10	**5**	**15**	**5**	**5**	**40**

So the expected agreement equals

$$A_e = 1,25 + 1,25 + 1,875 + 1,875 + 0,625 = 6,875$$

In short, kappa becomes

$$\kappa = \frac{A_o - A_e}{N - A_e} = \frac{14 - 6,875}{40 - 6,875} = 0,22$$

2. C
After all, kappa is downright pathetic. ☺

3. C
The theme colour is a nominal variable: we can easily put the colours in a different order, and then a weighted kappa would keep changing depending on that order.

4.
a) Follow the formulas in the theoretical sections, and you'll end up with the following weight coefficients:

		ARYOO (*i*)				
		Small change (1)	**Modest (2)**	**Hefty sum (3)**	**Fortune (4)**	
NUTS (*j*)	**Small change (1)**	1	0,67	0,33	0	
	Modest (2)	0,67	1	0,67	0,33	
	Hefty sum (3)	0,33	0,67	1	0,67	
	Fortune (4)	0	0,33	0,67	1	

So if all went well, you'll see the agreement decrease in uniform steps (that's the definition of linear!) from 1 to 0.
Were those all the hideous calculations? Largely:

$$A_o = (10 + 4 + 5 + 9) * 1 + (0 + 0 + 5 + 5 + 0 + 0) * 0,67 + (0 + 1 + 0 + 1) * 0,33 + (0 + 0) * 0 = 28 * 1 + 10 * 0,67 + 2 * 0,33 + 0 * 0 = 35,33$$

$$A_e = (3,75 + 1,25 + 1,25 + 3,75) * 1 + (1,25 + 1,25 + 3,75 + 3,75 + 1,25 + 1,25) * 0,67 + (1,25 + 3,75 + 3,75 + 1,25) * 0,33 + (3,75 + 3,75) * 0 = 10 * 1 + 12,50 * 0,67 + 10 * 0,33 + 7,50 * 0 = 21,67$$

$$\kappa = \frac{35{,}33 - 21{,}67}{40 - 21{,}67} = 0{,}75$$

Not bad!

b) Pff… 😐 er, I mean – OF COURSE we can! Super awesome, dude!

		ARYOO (*i*)				
		Small change (1)	**Modest (2)**	**Hefty sum (3)**	**Fortune (4)**	
NUTS (*j*)	**Small change (1)**	1	0,89	0,56	0	
	Modest (2)	0,89	1	0,89	0,56	
	Hefty sum (3)	0,56	0,89	1	0,89	
	Fortune (4)	0	0,56	0,89	1	

CHAPTER 31 MODERN PSYCHOMETRICS (COMPLETE)

1. B

The assumption of unidimensionality (A) applies to the classical *and* modern models. Modern models for polytomous and quantitative items are in development as well, so C isn't right either (but *Pirates, Peaches and P-values* only discusses dichotomous items in the context of modern item response theory). Answer B is about the assumption of parallel items: the classical model assumes this, while modern models don't.

2. C

Different as result from different reliabilities; the error variance then differs per item. This violates the assumption of parallelism. Items that aren't parallel will correlate less strongly. Since we use that correlation to measure the reliability of the items, their reliability will be underestimated – and with it, the reliability of the complete questionnaire.
Different bs imply a violation of parallelism as well: items with different difficulty levels are not parallel. This means that both statements are correct.

3. A

The item with the rightmost ICC has the highest difficulty level: the highest b. We can estimate the difficulty of an item from the item p-value: the proportion of ones (the proportion of persons who were correct, who filled out 'yes' or 'agree'). It's found in the *Item Statistics: Mean*. The higher that proportion, the easier the item and so the lower the b. We're looking for the hardest item, however, so for the one with the smallest item p-value. This is item I.

4. A

The item with the steepest ICC has the highest a. Since the a of an item is estimated directly from its item-rest correlation, we should see which has the highest *Corrected Item-Total Correlation* (see *Item-Total Statistics*). This turns out to be item VI.

5. A

First of all: why not B? Didn't we say once in chapter 29 that participants should score diversely on an item? After all, if everyone scores the same, the item can't distinguish anyone. The problem is that the dispersion on item VIII (expressed by the biggest standard deviation) could also be largely due to measurement errors. In that case an item with less dispersion, largely due to true differences, would be more useful.
The item with the highest discriminatory capacity, according to classical models, is the item with the highest reliability. A reliable item, which consistently gives similar results for the same participant, should be able to determine which participants score high and which score low. Thus we seek the item with the highest item-rest correlation, and that is item VI, as we established in exercise 4 already.
Although the difficulty level (b) has something to do with the informativity of an item as well, we can't make an isolated statement about that. See exercise 6.

6.

a)

Item	θ	a	b	$a(\theta - b)$	P	$I(\theta)$
I	0	1,5	2	-3	0,05	**0,1069**
II	0	0,5	0	0	0,50	**0,0625**
III	0	0,75	2	-1,5	0,18	**0,0830**
V	0	0,75	2	-1,5	0,18	**0,0830**
VI	0	1,5	0	0	0,50	**0,5625**
VII	0	0,75	2	-1,5	0,18	**0,0830**
VIII	0	1,5	-1	1,5	0,82	**0,3321**

b) Yes, the answer is confirmed! But this might not've been the case if, say, Hiroshima had wanted to separate extremely progressive teachers from the rest. In that case he should've administered more difficult items to the scientists. (Classical test theory, applied in Supplement 31, ignores the effect of different θs.)

c) Add up all the item informations: $I_{test}(\theta = 0) = 1{,}313$.

d) The information of such a test would be equal to $I_{test}(\theta = 0) = 3 * 0{,}0625 + 3 * 0{,}5625 = 1{,}875$. That makes a homogeneous test more informative than Hiroshima's current one in this case.

CHAPTER 32 FACTOR ANALYSIS (COMPLETE)

1. B

Kaiser criterion: 3 factors
Scree plot: 3 factors (C is therefore wrong)
Low residual correlations: isn't explicitly presented in the output
Maximum Likelihood: incomplete. We also need the Goodness of Fit test for 2 factors and perhaps even 1 factor, to see which model is the most parsimonious *and* isn't rejected. 3 factors are okay, but can we do with less as well? Thus we can't say for sure that all criteria agree.

2. A

Don't look at the *Initial Eigenvalue* in the *Total Variance Explained* table! It features the eigenvalues of PCA. Those of PFA *(Extraction Sums of Squared Loadings)* have been wiped by me, mwuahaha. In that case we'll just have to calculate Factor 3's eigenvalue by squaring all the loadings and adding them up. You'll find them in the *Factor Matrix* (use the unrotated solution!):

$$eigenvalue_{F3} = \sum \lambda_j^2 =$$
$$0{,}484^2 + 0{,}095^2 + 0{,}593^2 + (-0{,}033)^2 + (-0{,}278)^2 +$$
$$0{,}052^2 + (-0{,}321)^2 + (-0{,}285)^2 + 0{,}209^2 = \mathbf{0{,}904}$$

The K1 criterion should be applied by means of the PCA eigenvalues. For PCA, Factor 3 would've had a sufficiently high eigenvalue, namely 1,279.

3. A

The communality describes how well the variance in the separate items is explained; however, here we seek an explanation of the covariance or correlation between item 4 and 6. That's why we should investigate how well that correlation of 0,699 is reproduced by the factors. Use the *Factor Matrix* to look up all the loadings.

$$r_{reprod} = \lambda_{4,1}\lambda_{6,1} + \lambda_{4,2}\lambda_{6,2} + \lambda_{4,3}\lambda_{6,3} =$$
$$0{,}674 * 0{,}652 + (-0{,}552) * (-0{,}426) + (-0{,}033) * 0{,}052 = 0{,}673$$

The residual correlation is the difference between the real correlation between item 4 and 6, and the reproduced one:

$$\boldsymbol{r_{res}} = r - r_{reprod} = 0{,}699 - 0{,}673 = \mathbf{0{,}026}$$

This is what A says and it's a small value indeed! A model with three factors explains the relationship between stress and sarcasm very well.

4. C

This is about the communality of item 5. We can look it up: it's 0,906. So, a good 90% of the dispersion in the individual scores at item 5 is explained by the factors. With that $1 - 0{,}906 = 0{,}094$ of unexplained dispersion remains: the unicity, a minor 10%. This is very little! ☺

5.

- The structure has become simple, which it wasn't before rotation

Before the solution was rotated, many items still loaded highly on several factors *(Factor Matrix)*. This made it pretty difficult to decide which factor each item measures exactly. The simple structure that arises after oblique rotation *(Pattern Matrix)* is much clearer: now all the items consistently have a single factor on which they highly load. This statement is correct.

- The item communalities have increased

No: rotation doesn't improve the factor model whatsoever. Together, the factors still explain the exact same amount of variance in the items. See the theoretical explanation ($\boldsymbol{communality} = a^2$). This statement is false.

- The reproduced correlations between the items have become more accurate

No again: the model's goodness of fit remains identical when we rotate ($r_{reprod} = a_1 a_2 \cos\alpha$). This statement is false.

- The factors can now correlate, which they couldn't before

Oblique rotation allows the factors to correlate (orthogonal rotation doesn't). That makes this statement correct.

6.

a) What follows are just my interpretations – feel free to think of better ones yourself! Check the *Pattern Matrix*:

- ♦ Item 1 (joy) and 3 (enthusiasm) load highest on Factor 3. This factor appears to be something like the skill to express strong positive emotions.
- ♦ Item 5 (despair), 7 (anger) and 8 (sorrow) load highest on Factor 2. Clearly we're looking at the skill to express strong negative emotions.
- ♦ Finally, item 2 (hesitation), 4 (stress), 6 (sarcasm) and 9 (compassion) load highest on Factor 1. What these items share, I'd say, is that they all require a bit less drama than the others. How about calling their common ground the skill to express subtle emotions?

b) The *Factor Correlation Matrix* indicates that some factors are somewhat related, but not a lot. We see, for instance, that the correlation between Factor 1 and 2 equals 0,285: a good subtle actor is sometimes a good drama queen as well (but surely not always). The correlation between Factor 2 (dramatic acting) and 3 (happy acting) is about the same: 0,312. However, the correlation between Factor 1 and 3 turns out to be almost 0 (namely 0,052). It can't be predicted if a good happy actor will also be able to pull off some subtle scenes – perhaps because these require more seriousness.

c) The disadvantage: the use of fewer items makes a measure less reliable! I would therefore not recommend it.

CHAPTER 33 POWER ANALYSIS

1.

a) Let's see here. First we can calculate Cohen's d:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma} = \frac{|3 - 30|}{3} = 9$$

Ahem… a fairly large effect it seems. Next, we can do the sample size calculation…

$$n_1 = n_2 = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 * \frac{2}{d^2} = (2{,}576 + 3{,}091)^2 * \frac{2}{9^2} = 0{,}793 \rightarrow 1$$

So yeah, the outcome is that you'd need less than one cat and dog to prove this with 99,9% power. If you're going to do a t-test (which is the praxis), add 2 to this result.

b) It's to make sure that the test is robust against a violation of normality.

c) Nah, that won't be necessary. We can use matching to improve the power and thereby reduce the required sample size, but the power is already absurdly high even if the sample sizes are minimal.

d) Yes: it's likely that two types of cats are more similar than cats and dogs, so the effect size will be smaller.

2.

a) First we need Cohen's d again, starting from the standard deviation:

$$\bar{\pi} = \frac{\pi_1 + \pi_2}{2} = \frac{0{,}05 + 0{,}35}{2} = 0{,}20$$

$$\sigma = \sqrt{\bar{\pi}(1 - \bar{\pi})} = \sqrt{0{,}20 * 0{,}80} = 0{,}40$$

$$d = \frac{|\pi_1 - \pi_2|}{\sigma} = \frac{|0{,}05 - 0{,}35|}{0{,}40} = 0{,}75$$

Now the sample size calculation:

$$n_1 = n_2 = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 * \frac{2}{d^2} = (1{,}645 + 1{,}282)^2 * \frac{2}{0{,}75^2} = 30{,}46 \rightarrow 31$$

b) When we compare the first and the last group, the effect size is rather big, so this actually benefits the power. The difference between the other groups will be smaller, and therefore the power of certain pairwise comparisons will be lower – unless more participants are used than in exercise a.
The significance level is higher than normal, which also benefits the power (but makes Type I errors more likely).
The level of measurement is dichotomous, which makes for relatively low power.
Finally, comparing four groups will require a heavy Bonferroni correction (the significance levels of the pairwise comparisons must be reduced by a factor 6), which also lowers the power.

DISCLAIMER
Modifications and errors reserved.
If you think you've found an error,
please contact vince@pppwaarden.nl
or facebook.com/pppvalues.
We'll be happy to fix it! ☺