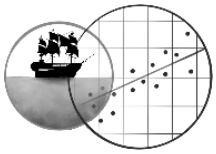


VERSION DATE: 21-11-2020

This document with solutions is currently in development. Please check my website regularly for the latest version. If you urgently need a set of solutions that I haven't written yet, send me an e-mail at vince@pppwaarden.nl.

CHAPTER 1 CHARTING DATA (COMPLETE)

1.A EXERCISES FOR PEACHES

1. C

You can pick answer A only in the case of ratio variables. However, the numbers on the Likert scale have no quantitative meaning! The distances (intervals) between consecutive numbers may not be equal either. This causes B to drop out as well: you don't know if two phones with 1 and 2 points respectively differ to the same extent as two phones with 2 and 3 points. Answer C does work: it applies to ordinal scales.

2.

Make sure that you're able to read the graph. The x-axis shows the phone model score (1 to 5 points); the y-axis indicates the frequency of each score (how many phones fell into each category).

a) Sure, a mode is always a go. It's clearly 4 (rather expensive and advanced). People often spend quite the money on their mobile devices, don't they…
b) A median is possible from ordinal variables onward, so bring it on. As indicated, we have 16 scores in total (you can confirm this by counting the bar heights). The median is therefore score number $\frac{N+1}{2} = \frac{17}{2} = 8{,}5$ – in other words, we should take the average of the eighth and ninth phone. Note that 8th and 9th are the ranks (or positions) of the scores, not their values. If we turn back to the bar chart and start counting, we discover that the eighth score is a 3 and the ninth a 4. That makes the median 3,5.
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) This seems plausible at first. After all, we can easily find Q_1 and Q_3; they turn out to be 2 and 4, respectively. But when we try to subtract them from each other, things get weird. 4 (rather expensive and advanced) minus 2 (rather cheap and basic) equals 2… but 2 what, exactly? 😐 This value has no clear meaning. After all, you need at least an interval variable to do this kind of calculation(!): the difference between subsequent categories should be constant! 2 might have a different meaning if the quartiles had been 3 and 5 (this difference could, conceptually, be smaller or larger than the difference between 2 and 4). In short: it makes no sense to compute an IQR.

3.

a) Feel free to discuss this a bit with your fellow students. Personally, I would expect a distribution that's skewed to the right. This kind of accident should happen to most people only once or never; only a few individuals are clumsy enough to drop their phone multiple times.
b) My first guess would just be a unimodal distribution. However, it *is* possible that we're going to see a divide between genders. Perhaps some men are confident that they can pee one-handedly and check their messages in the meantime – which has to go wrong every so often… In that case, men might have their own peak in the distribution, further to the right (or at least their own tail).
c) Should we theorise? Of course we should! That will help us understand what's going on once we see the data. Moving back and forth between reality and maths is what this book is all about, and anything that supports the process is a good thing. ☺

4. A

The distribution is skewed to the right, mainly due to one high score (an outlier perhaps?). This right skewness pulls up the mean, but not so much the median, which is a bit closer to the peak. Check FIGURE 1.10 (page 22) in the *PPP* book if you need to refresh the general rules.

5. A

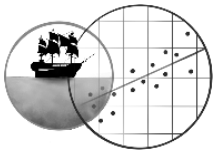
The trick is to find the quartiles… Did you find them?
Since 25% of the respondents had caused 0 accidents, this is the first quartile: $Q_1 = 0$. The third quartile is formed by the value 2: $Q_3 = 2$. There you go! This means that

$$IQR = Q_3 - Q_1 = 2 - 0 = 2$$
$$1{,}5 * IQR = 1{,}5 * 2 = 3$$

We suspect an upward outlier, so let's just calculate the upper bound of the acceptable range:

$$Q_3 + 1{,}5 * IQR = 2 + 3 = 5$$

The idea is that scores above 5 are considered outliers. The person who dropped his phone 7 times clearly fits the bill.



6.

a) Mode: if we add one person who scored 1, the most frequent number of accidents is still 0. The mode therefore stays the same.
Median: the new participant scored equal to the median, so we're just adding a person in the middle. This obviously keeps the median (the middle score) the same as before. You can also calculate it again and confirm that I'm right.
Mean: doesn't change, for the same reason as the median. If you add an average person, the average will not move.
Conclusion: all quantities keep their original values.

b) Well, the standard deviation represents the average deviation from the mean. This new person scored equal to the mean, so she didn't deviate from it at all. As a result, the average deviation should get a little smaller.

c) The formula is as follows:

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

The sum of squares, $\sum(X_i - \bar{X})^2$, is calculated by subtracting the mean from all the individual scores, squaring the differences, and then summing all these squared values. $(5-1)^2$, $(2-1)^2$, and so on. Now we're going to add the new participant. She scored equal to the mean, which implies that she adds $(1-1)^2 = 0$ to the sum of squares.
However, we do have an extra person now: N rises from 16 to 17. This means that the denominator of the formula gets slightly larger, while the numerator stays the same. The outcome, the standard deviation, decreases as a result.

7. D

Both Yoeri and Nout are only 0,76 standard deviations away from the mean (which is 0 on a z-scale); Nout is below, while Yoeri is above.

1.B EXERCISES FOR PIRATES

1.

a) They expected that most people would steal a small sum of money, and that relatively few would have the guts to go for large amounts.

b) There were still quite a lot of Corellians who grabbed huge amounts of money from the safe. Had the distribution been more heavily skewed to the right, then big thieves would have been rarer.

2.

a) The calculation is

$$z = \frac{X - \bar{X}}{s_X} = \frac{35 - 74{,}87}{40{,}728} = -0{,}9789$$

b) You can clearly see that Henry S.'s raw score is higher than Jane M.'s. Linear transformations are not going to change that at all: S. must have a higher z-score than M. You can even calculate it:

$$z = \frac{X - \bar{X}}{s_X} = \frac{50 - 74{,}87}{40{,}728} = -0{,}6106$$

The transformation to euros is just a detour which does nothing: you can convert an amount of dollars to yen, rupees and francs, but it will still be worth the same amount in euros. In the same way, the z-score will stay the same. I was simply trying to throw you off. ☺

3.

a) Let's get the interquartile range first:

$$IQR = Q_3 - Q_1 = 100 - 49 = 51$$

This times 1,5 is:

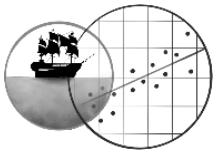
$$1{,}5 * IQR = 76{,}5$$

Now let's compute the acceptable range:

$$Q_1 - 76{,}5 = 49 - 76{,}5 = -27{,}5$$
$$Q_3 + 76{,}5 = 100 + 76{,}5 = 176{,}5$$

Scores that fall outside this range count as outliers. Since the minimum is 0, there are no extremely low scores. The maximum is 215, however, which exceeds the upper bound. We seem to have some upward outliers.

b) The histogram doesn't suggest that the scores above 176,5 are really outliers; they don't stand apart from the rest of the distribution, which flows rather smoothly. What this means is that the 1,5*IQR criterion is not a holy one. It's a useful tool to detect potential outliers, but in the end it's just that – something to help you. The final decision is always up to educated humans, not to maths. ☺ Chapter 20 in **Parrrt 2** of *Pirates, Peaches and P-values* will pick up the story on outliers again.



4. B

Answer A sounds appealing but is wrong. The median is resistant to outliers, sure, but this only means that the value of the maximum (or minimum) score has no impact on what the middle score will be. The girl could've stolen 200 or 500 – this would not have changed the median of 67. But when we take the girl out, the location of the middle score will certainly change, because the number of scores drops by one!

Let's imagine that B is correct. What would that look like? There were 268 participants originally, which is an even number, so the median must've been the average of the middle two scores:

…… 67,67, ……

Now a person who scored above the median was excluded, leaving 267 participants. This must mean that the middle of the data set shifts slightly down, to the first 67. This sounds like a good explanation!

Answer C can only be called gibberish. If we remove a score, it's perfectly possible that the median changes while the mode stays the same.

5.

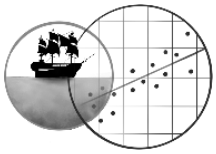
I'm not sure myself if there is a definitive answer to this question. Which is fine! ☺

At least I don't think the folder letter is of interval level; King only lists the rank number of the letters, but these rank numbers (1st letter, 2nd letter) don't imbue the letters A, B et cetera with a quantitative meaning.

Nguyen makes an interesting point; the order of the Roman alphabet has been fixed since ancient times. This suggests that we have an ordinal variable here. However, my personal view is that this order is kind of arbitrary; why did the Romans decide to put M before N? They could have chosen a different order and no one would've blinked an eye. I would therefore call the letters nominal. But you're allowed to disagree with me, dear reader.

6. C

The chief characteristic of linear transformations is that they leave the shape of the distribution completely intact. You can see this in the explanation in book section 1.6. Answer A is false because the distribution becomes *more* skewed, which never happens with linear transformations (the degree of skewness stays the same). Answer B describes something that also occurs when we centre (linear transformation!): the mean shifts, but the standard deviation doesn't. Also, did you check the formula properly? You can clearly see that this is a quadratic transformation. ☺



CHAPTER 2 CATEGORICAL RELATIONSHIPS (COMPLETE)

2.A EXERCISES FOR PEACHES

1. B

Answer A makes no fair comparison, because there were many more people who had only one choice to begin with. What matters is if both route conditions showed the same proportion of massacres. So we're interested in the relative frequencies, or 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.

2. C

By looking only at the univariate distributions, we cannot see how variables are related, precisely because the contingency table can still be filled up 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%)**

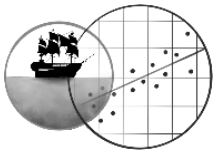
It gives us no association.[1]
Or this one:

		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 – the strongest association that is possible with the current marginals, in fact.[2]

[1] "Vincenzo, how did you make this one?"
I used some knowledge from chapter 16 on kappa. To calculate this measure of agreement, we need to make a contingency table that shows no association. Take the formula for the expected counts and presto.

[2] "And how did you construct this one, Vincenzo?"
I made the association as strong as possible by filling in the biggest possible number. The largest marginal value is 52. However, you cannot fit that anywhere in the empty cells, since it won't match up with the other marginals. The second largest marginal is 48. We can fit this one in the upper left corner. The rest follows automatically.



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 how many subjects who were not massacred had one route (80,77%) and how many had two routes (19,23%). They also show how many subjects that *were* massacred had one route (50%) or two (50%).
The column percentages tell us how many of the subjects with one route were not killed (87,5%) and how many were (12,5%). They also describe how many of the subjects with two routes were not killted (62,5%) and how many were (37,5%).
Both sets of percentages are informative in their own way. Nevertheless, I lightly prefer percentages that best express the causal relationship. 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 might 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): the people who escaped were extremely afraid in large numbers… but perhaps this also applied to the people who had their heads sawn off. Unless we ask them as well, we haven't the slightest indication that fear helps you survive (and unfortunately, there is no scientifically proven way to ask the dead). As long as we only look at the individuals who got past the exit, the variable MASSACRED doesn't vary; however, relationships can only exist between variables.

2.B EXERCISES FOR PIRATES

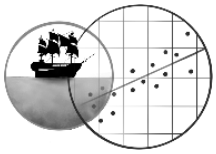
1.
Not really. It is the case that $\frac{707}{1127} \triangleq 62{,}7\%$ of all criminals were humans, but perhaps this race formed the biggest population to begin with. In other words, 62,7% of all innocent citizens could be humans as well. The dwarves would only have a point if the humans were overrepresented in the criminal records… but we don't know that yet. See exercise 2.

2.

a) Here are the data including relative frequencies:

		RACE			
		human	elf	dwarf	
CRIME	no	8552 (92,4%)	1377 (96,0%)	3083 (89,5%)	**13 012 (92,0%)**
	yes	707 (7,6%)	58 (4,0%)	362 (10,5%)	**1127 (8,0%)**
		9259 (100%)	**1435 (100%)**	**3445 (100%)**	**14 139 (100%)**

So yes, there was an association. The most criminal race was actually the dwarves (they had the highest percentage of juveniles).



b) No. This is a purely observational study. A third variable may be responsible for this apparent relationship between race and crime rate (as you are about to see).

3.
A single measure is not possible; risk differences, risk ratios and odds ratios are only suitable for 2x2 tables. To put it differently for this example, they can only compare two races at a time.
What we could do is calculate multiple ones. Here are some risk differences:

- Humans versus elves: $RD = 0{,}040 - 0{,}076 = -0{,}036$
  So based on these data, the risk of being criminal was 3,6% lower for elves than for humans.
- Humans versus dwarves: $RD = 0{,}105 - 0{,}076 = 0{,}029$
  So dwarves were 2,9% more likely to be criminal than humans.
- Elves versus dwarves: $RD = 0{,}105 - 0{,}040 = 0{,}065$
  So the crime rate among dwarves was 6,5% higher than among elves.

Likewise, you can compute 3 relative risks or 3 odds ratios if preferred.

4.

a) Let's do it:

POOR		RACE			
		human	elf	dwarf	
CRIME	no	4641 (90,2%)	428 (90,9%)	2710 (90,3%)	**7779 (90,2%)**
	yes	507 (9,8%)	43 (9,1%)	292 (9,7%)	**842 (9,8%)**
		5148 (100%)	**471 (100%)**	**3002 (100%)**	**8621 (100%)**

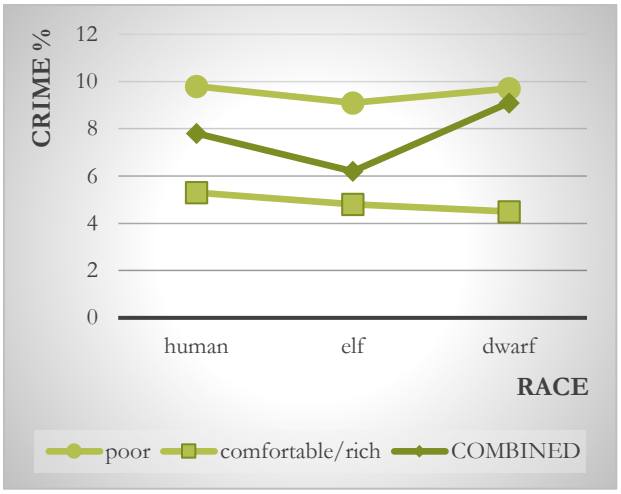
COMFORTABLE/RICH		RACE			
		human	elf	dwarf	
CRIME	no	3892 (94,7%)	918 (95,2%)	423 (95,5%)	**5233 (94,8%)**
	yes	219 (5,3%)	46 (4,8%)	20 (4,5%)	**285 (5,2%)**
		4111 (100%)	**964 (100%)**	**443 (100%)**	**5518 (100%)**

Well… that's weird. 😐 Now the humans were the most criminal race again (slightly)!

b) That is an interesting question:

		RACE			
		human	elf	dwarf	
WEALTH	poor	5148 (55,6%)	471 (32,8%)	3002 (87,1%)	**8621 (61,0%)**
	comfy/rich	4111 (44,4%)	964 (67,2%)	443 (12,9%)	**5518 (39,0%)**
		9259 (100%)	**1435 (100%)**	**3445 (100%)**	**14 139 (100%)**

As you can see, a large majority of the dwarves were poor. By contrast, elves were usually middle class or aristocracy. It was roughly 50-50 for the humans. This can solve the apparent contradiction between exercise 2 and 4. It's easiest to explain if I draw a graph of the crime percentages for every race and socioeconomic group:



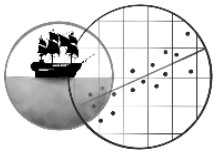
See what happens? The wealth of humans was 50-50, so their overall crime rate ended up halfway between the poor and wealthy group. The overall crime percentage of the elves was pulled down toward the rich group, because elves were mostly rich… and the dwarves looked more criminal overall because they were mostly poor – their overall percentage was pulled up.
So in fact, the citizens' WEALTH was a **confounder**. It had a strong impact on their chance to become criminal (since many poor people had to steal just to survive). But since the dwarves were much poorer than the humans, it looked as if this race was innately more likely to be criminal… whereas the three races are actually pretty similar when you look inside each WEALTH group separately. The confounding was so heavy that even the *wrong* race arose as the most criminal one (the dwarves, whereas it was actually the humans in both the poor and the rich group). This is when Simpon's paradox occurs: when a third variable causes confounding that is so extreme, it doesn't simply distort but even *reverses* the relationship we see between *X* and *Y* (in this case RACE and CRIME).

Another detail (only read it if you're interested or if you noticed something weird).
If you add the poor and rich group together manually in order to get the table from exercise 2, you will discover that I made some mistakes in it. Great. 😐 The total numbers are actually supposed to look like this:

		RACE			
		human	**elf**	**dwarf**	
CRIME	**no**	8533 (92,2%)	1346 (96,0%)	3133 (90,9%)	**13 012 (92,0%)**
	yes	726 (7,8%)	89 (6,2%)	312 (9,1%)	**1127 (8,0%)**
		9259 (100%)	**1435 (100%)**	**3445 (100%)**	**14 139 (100%)**

This makes a bit more sense because the overall crime rates now fall between the separate rates of the poor and rich groups (see the graph I made). Luckily, the story and conclusions of these exercises stay the same! ☺

5.
This is a 'food for thought' question that I should not answer in detail for you. ☺ Just think about the different outcomes that exercise 1, 2 and 4 have presented. Isn't there often a deeper truth beneath numbers that look clear and definitive?



CHAPTER 3 QUANTITATIVE RELATIONSHIPS

3.A EXERCISES FOR PEACHES

1. A

In my view, this scatterplot shows no problems. Note that you cannot drive on half-wheels, so it's not strange that the points constitute only whole numbers on the x-axis. We can still count the number of wheels, though, so this variable is quantitative (ratio, to be precise). The same goes for the race time in minutes. There are no clear subgroups either. The relationship between WHEELS and RACE TIME does not look very strong, but insofar it exists, a straight line can describe it well enough – we don't need a curve.

2.

The more wheels…: TRUE. We can say that there's a positive trend.
Race times increase…: 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).
Mounting more wheels…: FALSE. The predicted race time increases if the drivers use more wheels… Is that advantageous? No – it's a race. They want to drive as fast as possible, so in the shortest amount of time! ☺ Perhaps wheels make the soapboxes less maneuverable; they have to avoid a lot of obstacles after all. This would cost time.
Beside the number…: TRUE. The correlation doesn't equal 1, so the wheels aren't the only thing that influences the race 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 at the end of section 3.3 in the book. This is the opposite situation: we've currently studied a rather small range of wheels, which suppresses the correlation. Possibly this correlation would grow if drivers with even more wheels had joined as well.
We cannot be sure of this, by the way. Would the linear trend continue or not? It's also thinkable that yet another extra wheel won't make a difference for the race time at a certain point (so the line may flatten as a result). If you extend the regression line into an area that wasn't measured, you **extrapolate**. Extrapolation is always uncertain.

4.

a)

$$b = r_{XY} * \frac{s_Y}{s_X} = 0{,}268 * \frac{8{,}107}{0{,}845} = 2{,}57$$
$$a = \bar{Y} - b\bar{X} = 19{,}10 - 2{,}57 * 4 = 8{,}82$$

In short,

$$\hat{Y} = 8{,}82 + 2{,}57X$$

b) 8,82 is the predicted race time (in minutes) of a soapbox with no wheels ($X = 0$). That sounds rather hypothetical. ☺
c) 2,57 is the predicted increase of the race time when the number of wheels rises by 1. (This means that 2 additional wheels would increase the time by $2{,}57 * 2 = 5{,}14$ minutes, for example.)
d) Neither: to determine how strongly the race time depends on the wheel quantity, we need the correlation.

5.

It's helpful to calculate R^2 first:

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

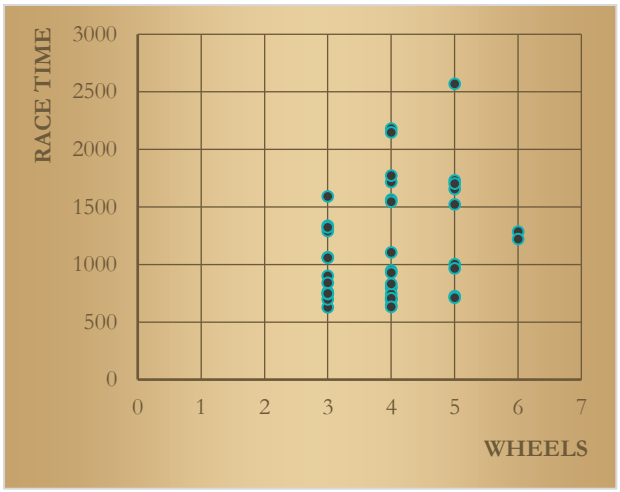
In other words, 7,2% of the variation in the race times seems to be related to the fact that some soapboxes used more wheels than others. That's not a lot: 92,8% of the variation in the race times must be attributed to other factors, such as the fact that the drivers differ in talent. The question is therefore whether the director should credit the wheels with the race results.

6. C

The residuals constitute the unexplained variation, and this is equal to $1 - R^2 = 1 - 0{,}072 = 0{,}928$.
Simple once you see it, right? ☺

7. B

The director has transformed the race time: from minutes to seconds. We have a multiplication here, since 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! Here's the scatterplot when we express the race times in seconds:



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

8. D

This exercise allows us to make a nice summary of all the things that are going on when no association presents itself. If it helps, look back at the scatterplot in FIGURE 3.8b. First of all, it means no correlation, so $r = 0$. Of course $R^2 = 0$ as well in that case, which confirms answer A: we cannot explain any variation in the race times. Third, what's the regression line like? It must be flat: $b = 0$. That confirms answer C as well, so in fact all statements must be correct (you should pick D). But why would B be right?

Well, let's look at the regression equation:

$$\hat{Y} = a + bX$$

We have just said that $b = 0$, so we might as well take the last element out:

$$\hat{Y} = a + 0 * X = a$$

Result: the predicted race time ($\hat{Y}$) is going to be the same for all drivers, namely a. After all, the wheels (X) cannot help us predict race times. Lastly, we should ask ourselves what the value of a could be. If your x-variable is useless, but you still want to predict someone's race time as accurately as possible, then what's the best prediction you can make? Think about this for a second.

Done thinking? It would be the *average* race time! And indeed, $a = \bar{Y}$ when we have no association. This also follows from its formula since b equals zero: $\boldsymbol{a = \bar{Y} - b\bar{X}} = \bar{Y} - 0 * \bar{X} = \boldsymbol{\bar{Y}}$.

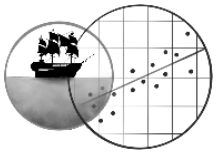
Conclusions:

NO ASSOCIATION	
1	$r = 0$
2	$R^2 = 0$
3	$b = 0$
4	$\hat{Y} = a = \bar{Y}$

Now answer B makes sense as well. The residuals form the distance of individual points from the regression line, but the regression line is just the mean line in this scenario. (That's another way of saying that all deviations remain fully unexplained.)

9. C

The two variables in question are categorical! We can only calculate a correlation between quantitative variables.



3.B EXERCISES FOR PIRATES

1.

a) Take care that PANCAKES describes the number of pancakes the guests have already eaten. The APPETITE that they still have left should depend on that. So APPETITE is the dependent variable, which should go on the y-axis. PANCAKES is the independent variable, which should go on the x-axis.

b) Everyone has to indicate their appetite again after every pancake, so all guests appear in the scatterplot multiple times. In other words, this is a within-subjects design. So far we have only conducted regression for between-subjects designs, when each point is a different participant. This matters only a little for describing the sample relationship between X and Y (but it becomes very important when you want to statistically test it in later chapters).

c) A negative one: the APPETITE should go down the more PANCAKES you eat.

2.

a) It's the same scatterplot, just flipped (X and Y axes trade places). The points follow the line equally well.

b) Let's make an overview:

	ANALYSIS I	ANALYSIS II
a	Predicted appetite ($\hat{Y}$) for a person who hasn't eaten any pancakes yet ($X = 0$): 68,04 appetite points	Predicted number of pancakes that a person has eaten when their appetite is down to 0: 5,00 pancakes
b	Predicted change in a person's appetite when they eat 1 additional pancake: a decrease by 11,07 appetite points	Predicted difference in how many pancakes a person has eaten when their appetite is 1 point higher: 0,06 pancakes less

c) C
The first analysis is much more logical and easier to interpret (PANCAKES should influence APPETITE).

3.

a) Nope. We already established this in exercise for peaches 7. The z-score is a linear transformation as well.

b) Z-scores have a mean of 0 and a standard deviation of 1. This lets us calculate the new a and b:

$$b = r * \frac{s_{Z_Y}}{s_{Z_X}} = -0{,}809 * \frac{1}{1} = -0{,}809$$

Lookie here: when we standardise, the slope becomes equal to the correlation coefficient! $\boldsymbol{b = r_{XY}}$

$$a = \bar{Y} - b\bar{X} = 0 - (-0{,}809) * 0 = 0$$

And look: when we standardise, the intercept always becomes 0 (so it can be dropped)! $\boldsymbol{a = 0}$

c) Note again that z-scores have a mean of 0 and a standard deviation of 1. So the slope b tells us that if X increases by 1 unit i.e. 1 standard deviation, we predict that Y increases by -0,809 standard deviations.
(We can ignore the intercept since it's always 0.)

4.

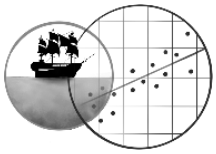
On second thought, the relationship between PANCAKES and APPETITE appears quadratic: we can draw a downward curve, i.e. the left half of a U-shape. This makes more sense than a straight line, in fact. Your APPETITE does not forever continue to go down when you eat more PANCAKES; rather, it will be very low at some point and stay there when you keep eating. Statistically, we could add a quadratic term to the regression equation to make this curvilinear relationship possible: $\hat{Y} = a + bX + cX^2$. See section 3.6.I for some more information.

5. A

If the relationship between PANCAKES and APPETITE is actually quadratic and not linear, and we take this into account, we'll see that their relationship is actually stronger than we originally thought. After all, we'll draw a curve through the plot that fits the points even better than a straight line. This will make the residuals (the unexplained variation) smaller, which causes R^2 (the proportion of explained variation) to go up.

6. B

The subgroup in question would be another group of guests who eat normal pancakes. If their APPETITE curve goes down more quickly, this will suggest that Gulliver's pancakes achieve what he advertises.
(This is just a bonus, but in fact, Gulliver is therefore looking for quadratic moderation/interaction – a topic we'll get back to in **Parrrt 2** of *Pirates, Peaches and P-values*.)
A and C may sound tempting but don't get you anywhere. Gulliver needs to make a comparison with a control group.



CHAPTER 4 PROBABILITY THEORY (COMPLETE)

1.

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

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

The product rule works as well, although it's more tedious:

$$P(bribe\ and\ bear) = P(bribe) * P(bear|bribe) = \frac{30}{270} * \frac{12}{30} = \frac{12}{270} = 0{,}044$$

Or

$$P(bear\ and\ bribe) = P(bear) * P(bribe|bear) = \frac{24}{270} * \frac{12}{24} = \frac{12}{270} = 0{,}044$$

b) This probability is different from the one in question a! It's a conditional one: the probability of bribery, if the child won a stuffed bear. Write down the requested probability correctly and you'll be fine:

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

c) It was the one from question b. This probability would have told the official that kids who won a bear were clear suspects (50% of their parents had bribed). Had he known this, he would have been able to catch parents in the act more easily. The probability from question a was not really useful, because the stuffed bears did not interest him as such.

d) We have statistical independence if

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

Do the data satisfy this condition?

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

However,

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

… as we already calculated in exercise b. So it was much more likely that the parents had bribed if the child won a stuffed bear. The event 'bribery' was strongly dependent on the event 'stuffed bear'.
Another good solution:

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

However,

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

Which means that the probability that the child would win a bear was 8,9% in general, but it rose fiercely (to 40%) if the owner was bribed. (These probabilities express the causal relationship between the variables a bit more neatly.)

2.

a) Use the product rule. It's the probability of a bribe *and* another bribe:

$$P(2\ bribes) = \frac{30}{270} * \frac{30}{270} = \frac{1}{81} = 0{,}012$$

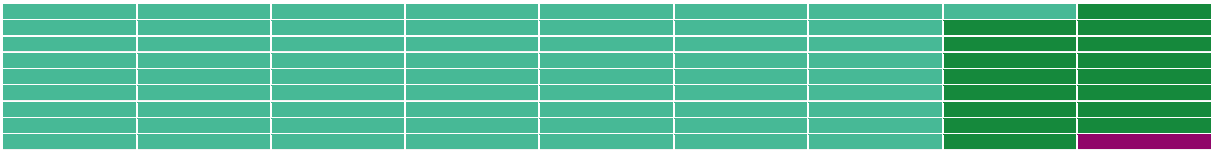
b) Be a bit careful with this one. There are two possible orders: the first kid could have bribing parents while the second one did not, or the other way around. It helps to make a tree diagram like in section 4.5. So,

$$P(1\ bribe) = \left(\frac{30}{270} * \frac{240}{270}\right) + \left(\frac{240}{270} * \frac{30}{270}\right) = \frac{16}{81} = 0{,}198$$

c) This probability is simpler once again (no bribe *and* no bribe):

$$P(0\ bribes) = \frac{240}{270} * \frac{240}{270} = \frac{64}{81} = 0{,}790$$

d) This is an event: a set of multiple elementary outcomes. Consider what one elementary outcome is: it's a sample of 2 children. We have 270 options for the first kid we draw, and 270 for the second one (assuming that we draw with replacement). So there are $270^2 = 72900$ possible combinations (72900 elementary outcomes). It's too much work to write them all down individually, but we can say that 79,0% of all the possible samples contain two children who both stuck to fair play. Likewise, just one of the two kids had bribing parents in 19,8% of all possible samples, and both kids did in 1,2% of the samples. Remember that probability distributions always describe events.
By the way, the probability distribution is correct, since the sum of the three probabilities equals 1. ☺



Here's a visualisation of the sample space.

e) I'm not so sure about that. Suppose that the boy won a stuffed bear? Then his parents might be more inclined to bribe the owner, so his little sister would have a higher chance of winning a bear as well. We're not sure what the parents would do though, so we would be unable to determine the probability that the girl (the second child) would have bribing parents. In other words, the events would not be independent.
So keep this in mind if you ever test participants yourself: they *must* be independent. Almost all statistical techniques in *Pirates, Peaches and P-values* assume that this is the case.

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 when we draw random samples.

4. B

The event in A requires you to 'stretch out' the product rule. No bear *and* no bear *and* no bear *and*… in other words, multiply the probability that a random child won no bear with itself twenty times:

$$P(0\ out\ of\ 20\ winners) = \left(\frac{246}{270}\right)^{20} = 0{,}155$$

The event in B asks that you to the same: a bear *and* no bear *and* no bear *and*… but don't forget that there were 20 different orders, because the child who won the bear could be in 20 different positions:

$$P(1\ out\ of\ 20\ winners) = \left(\frac{24}{270}\right)^{1} * \left(\frac{246}{270}\right)^{19} * 20 = 0{,}303$$

So event B is almost twice as likely. (You may have expected that already.)
Answer C is false: in a uniform probability model, all elementary outcomes are equally likely (which is the case with random sampling), but certainly not all events! The likelihood of each event depends on how many elementary outcomes it contains.

5.

Note that a random variable is a quantity whose value depends on chance (the elementary outcome).

- The size of the sample is not a random variable, because the researcher controls it. If you change N, you also change the probability experiment.
- The number of children in the sample that won a stuffed bear is a random variable: it could be anything between 0 and 20.
- The same goes for the number of red-haired children in the sample. This is a random variable (also because, I will admit, I chose it a bit randomly).
- The number of parents who bribed the owner on that day is not a random variable. After all, it's a fixed quantity that belongs to the population (a parameter; see chapter 5).

6.

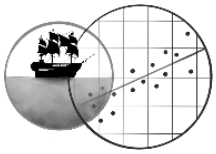
a) When $N = 2$, you can still calculate the expected value by taking the different possible numbers of bribes and multiplying them by the probabilities you calculated in exercise 2, before adding them up ($E(X) = \sum_i X_i * P(X_i)$):

		X		$P(X)$
		0	*	$\frac{64}{81}$
	+	1	*	$\frac{16}{81}$
	+	2	*	$\frac{1}{81}$
$E(X) =$		$\frac{18}{81} = 0{,}222$		

It's easier, however, to use the shortened formula for binomial distributions:

$$E(X) = N * p = 2 * \frac{30}{270} = \frac{2}{9} = 0{,}222$$

This amounts to $\frac{0{,}222}{2} \triangleq 11{,}1\%$ of the total sample. Which is logical, since $\frac{30}{270} \triangleq 11{,}1\%$ of the population bribed.



When $N = 20$, making a table like above is probably way too much work. I would definitely use the simplified formula here:

$$E(X) = N * p = 20 * \frac{30}{270} = \frac{2}{9} = 2{,}22$$

Which again amounts to $\frac{2{,}22}{20} \widehat{=} 11{,}1\%$ of the total sample.

b) When $N = 2$, we can calculate the variance in two ways again. The general formula is $\sigma_X^2 = \sum_i (X_i - \mu_X)^2 * P(X_i)$:

	$(X - \mu_X)^2$		$P(X)$
	$\left(0 - \frac{2}{9}\right)^2$	*	$\frac{64}{81}$
+	$\left(1 - \frac{2}{9}\right)^2$	*	$\frac{16}{81}$
+	$\left(2 - \frac{2}{9}\right)^2$	*	$\frac{1}{81}$
$\boldsymbol{\sigma_X^2 =}$	$\boldsymbol{\frac{16}{81} = 0{,}198}$		

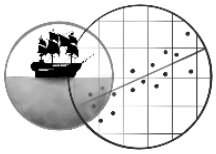
The simplified formula for binomial distributions works a tad faster:

$$\sigma_X^2 = N * p * (1 - p) = 20 * \frac{30}{270} * \frac{240}{270} = 0{,}198$$

The standard deviation is therefore $\sigma_X = \sqrt{0{,}198} = 0{,}444$, and this amounts to $\frac{0{,}444}{2} \widehat{=} 22{,}2\%$ of the total sample.
When $N = 20$, definitely use the simplified formula:

$$\sigma_X^2 = N * p * (1 - p) = 20 * \frac{30}{270} * \frac{240}{270} = 1{,}98$$

This makes the standard deviation $\sigma_X = \sqrt{1{,}98} = 1{,}41$, which amounts to only $\frac{1{,}41}{20} \widehat{=} 7{,}0\%$ of the total sample.
Conclusion: on average, a larger sample deviates less from the population than a smaller sample does!



CHAPTER 5 PROBABILITY DISTRIBUTIONS (COMPLETE)

1.

a) The telephone number rule (68-95-99,7) tells us that 68% of the pizzas are no more than 1 standard deviation away from the mean. So,

$$\mu \pm \sigma$$
$$1{,}1 - 0{,}4 = 0{,}7$$
$$1{,}1 + 0{,}4 = 1{,}5$$

Answer: the middle 68% of the pizza population are between 0,7 and 1,5 millimetres thin.
In addition, the middle 95% is at most 2 standard deviations away from the mean, or between 0,3 and 1,9:

$$\mu \pm 2\sigma$$
$$1{,}1 - 0{,}8 = 0{,}3$$
$$1{,}1 + 0{,}8 = 1{,}9$$

b) A normal distribution is symmetric (sketch it!). So now that we indicated the 95% middle pizzas in exercise a, 5% remain, which are divided evenly across the left and right tail of the distribution. 2,5% of the pizzas are thinner than 0,3 millimetres, and 2,5% are thicker than 1,9 millimetres. In short, the 2,5% thickest pizzas are 1,9 millimetres or thicker.

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

$$z = \frac{X - \mu}{\sigma} = \frac{1{,}2 - 1{,}1}{0{,}4} = \frac{0{,}1}{0{,}4} = 0{,}25$$

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

$$P(z > 0{,}25) = 1 - 0{,}5987 = 0{,}4013$$

So Nick complains 40,13% of the time. That sounds really annoying.

d) A very similar problem to question c. But this time 2 pizzas are drawn, and we look at their mean thickness. The probability that we would get a sample mean of at least 1,4 millimetres must be sought under the sampling distribution. This changes the z-formula slightly, because the standard deviation of a sampling distribution is the standard error:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1{,}4 - 1{,}1}{0{,}4/\sqrt{2}} = 1{,}06$$

Are you afraid that you might mix up the right formula for z when you're trying to determine a probability? Then a quick-and-dirty solution is so always use the formula for the sampling distribution of the mean (with $\sqrt{N}$ included). After all, when $N = 1$, the sampling distribution of $\bar{X}$ is identical to the population distribution anyway, and you will divide by the square root of 1 in the z-formula… which does nothing.
Now then, according to the z-table,

$$P(z > 1{,}06) = 1 - 0{,}8554 = 0{,}1446$$

So there's a 14,46% chance that Nick will stumble upon a sample of this kind.

2. D

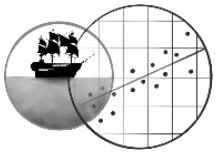
This percentage is an estimate: a quantity obtained from a random sample. It's an estimate of the population parameter, the percentage of *all* pizzas that are not thin enough in Nick's view. The sample are of course the 2 pizzas that Nick picked, not their thickness. Let's tabulate that:

POPULATION:	SAMPLE:
all of Luca's pizzas	the 2 pizzas meant for Nick and his niece
PARAMETER:	ESTIMATE:
the sample percentage of pizzas that are thinner than 1,2 millimetres (also 0%?)	the sample percentage of pizzas that are thinner than 1,2 millimetres (0%)

So Nick mentions an estimate, but he presents it like it's a parameter. Can he do that? See exercise 3…

3. B

Let's review the definitions of bias and efficiency. An estimate is unbiased if its expected value equals the parameter, so in this case, if the average of all the sample means that you could draw is equal to the population mean ($\mu_{\bar{X}} = \mu$). Chapter 5 has taught us that this is the case whenever the sample is randomly drawn (and that sample size doesn't influence the principle). So if Luca wants to say that the sample mean of 1,4 millimetres is biased, he can only mean that the sample was not random… in other words, that he was about to bake extra thick pizzas for Nick on purpose. That can't have been what he meant. ☺
No, what Luca probably means is that the sample mean is unreliable – that it might easily have been 1,0 millimetres (for example) in another random sample of two pizzas. This $\bar{X}$ of 1,4 does not have to represent the population mean because $\bar{X}$ fluctuates a lot across samples. In other words, its standard error is too large. And that's exactly how we



define an inefficient estimator. (Note also your answer to question 1d: a sample mean of at least 1,4 millimetres will pop up in about 14% of all samples, which is pretty often when you think about it. Nick just got a bit unlucky.)
The central limit theorem only says that when the sample is large enough, the shape of the sampling distribution of $\bar{X}$ becomes approximately normal, even if the population distribution is not. That doesn't seem related to this question. Besides, exercise 1 says that the population distribution is already normal, so for a normal sampling distribution, we would not have needed the central limit theorem anyway in this case.

4.

a) This sample is very small ($N = 7$). The book chapter has shown you that samples of this size can easily give a distorted picture of the population distribution: the shape and the estimates from FIGURE 5.19 aren't very reliable (efficient). So we shouldn't make any strong claims about the population distribution. It *could* look like the given distribution of sample scores, but it could just as well be vastly different.
b) This sample is very large ($N = 230$); the distribution of sample scores should resemble the population pretty well. The shape appears almost normal, so I think the population distribution may well be roughly normal as well. It says next to the histogram that the sample mean equals $\bar{X} = 7{,}96$, and the population mean μ should be similar, so approximately 8 grams. The sample standard deviation equals $s = 1{,}009$, which makes it likely that the population standard deviation σ is about 1 gram.
Now for the sampling distribution. Its shape must be pretty close to normal, thanks to the central limit theorem (plus the fact that the population distribution already seems pretty normal). Its expected value $\mu_{\bar{X}}$ equals the population mean, so that should be about 8, and the standard error is $\sigma_{\bar{X}} = {}^{\sigma}/_{\sqrt{N}} \approx {}^{1}/_{\sqrt{230}} = 0{,}066$. We can see from this standard error how efficient the sample mean is: on average it will deviate from the population mean by a mere 0,066 grams.
c) This sample is large as well ($N = 145$); 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 therefore likely to be skewed to the left. The population mean μ will be roughly 11 grams and the standard deviation σ about 2 grams. 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: $\mu_{\bar{X}} \approx 11$, $\sigma_{\bar{X}} \approx {}^{2}/_{\sqrt{145}} = 0{,}166$.

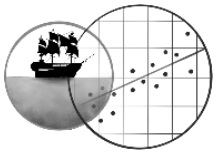
5. A

It looks like Nick did his best to draw a random sample. When that's the case, the expected value of the sample mean will always equal the population mean. And that defines it as an unbiased estimator of μ.
Answers B and C talk about the shapes of the various distributions (population versus sampling distribution, respectively). But these shapes are irrelevant. (The stroopwafel population might be normally distributed, but we can't be sure that this is the case. The central limit theorem does apply, which means the sampling distribution of $\bar{X}$ must be approximately normal; but again, this information has no impact on the expected value.)

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.[3] More than 50% of the scores lie to the right of Luca's z-score, so more than half of the stroopwafels (20) 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): $P(z < -0{,}20) = 0{,}4207$. So, Luca's waffle is roughly among the lightest 42%, but definitely not among the lightest 20%! Shame, now he has no excuse to take another one…

[3] Also holds for a skewed distribution, since z-scores express how many standard deviations a measure is above or below the mean.



CHAPTER 6 HYPOTHESIS TESTING (COMPLETE)

6.A 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, which leads to a particular expected value for your sample mean. You will need to compare your sample result against that fixed expected value. You can't compare it against a range of possible expected values. As for the alternative hypothesis, Santa Claus believed that red-nosed reindeers would lose their way less often. That makes a case for the left-sided alternative ($\mu < 10$).

2.

The distribution of sample scores contains an outlier – a gnome who got lost almost twenty-five times. This outlier may distort the sample mean: 10,14 may be unrepresentatively large. So what's the point? Always look at your data before doing a test! Santa should first decide what to do with this outlier. If it changes the mean a lot, it could be smarter to take it out (but this is not always allowed; see **Parrrt 2** of *Pirates, Peaches and P-values*).

3. C

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

$$9{,}75 \pm 1{,}96 * {}^{3}/_{\sqrt{36}}$$
$$9{,}75 \pm 1{,}96 * 0{,}5$$
$$9{,}75 \pm 0{,}98$$
$$[\ 8{,}77\ ;\ 10{,}73\ ]$$

This result should be interpreted like answer C says. The logic of the confidence interval pertains to the sampling distribution of $\bar{X}$; the components of the formula confirm this (look at the standard error for instance). The idea is when you draw many samples, 95% of the time the sample mean will be at most 1,96 standard errors away from μ. So now that we have drawn a single sample, we can be 95% confident that the real μ is no more than 1,96 standard errors away from $\bar{X}$. For more details, reread the theoretical discussion in section 6.4.

Thus, A and B are common misunderstandings. They refer to the population distribution (A) or to the distribution of sample scores (B). But confidence intervals do not describe the behaviour of individual scores; they're only an attempt to figure out the population mean. My student Marie also noticed that we have the distribution of sample scores: it's FIGURE 6.14… and you can clearly see in it that less than 95% of the scores fall between 8,77 and 10,73 misdirections. This is another way to debunk answer B.

4. B

If we perform the test, we obtain the following:

$$z = \frac{9{,}75 - 10}{{}^{3}/_{\sqrt{36}}} = \frac{-0{,}25}{0{,}5} = -0{,}50$$

Santa conducted a left-sided test, so you may look up a left-tailed 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.

$$p = P(z < -0{,}50) = 0{,}3085$$

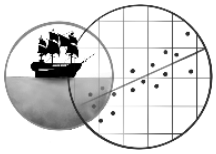
Thus the test is not significant (the p-value is way above the regular significance level of 5%), and we cannot reject the null hypothesis. But apparently, the challenge in the question is to interpret what the p-value says. B is the only correct interpretation. We cannot assign a probability to the null or alternative hypothesis itself; after all, the null hypothesis is either true or false. No probability is involved with that. We can only investigate how likely Santa's sample result would be, if the null hypothesis was true. A hypothetical probability for a hypothesis test… Since a sample like Santa's would be quite regular if the population mean equalled 10 misdirections indeed, we have no good reason to doubt the null hypothesis.

5. C

The only knowledge you can gain from the confidence interval is that the population mean probably lies within its borders. It could therefore be 10 in this case (H_0 might be true), but it could just as well be 13. All values within a confidence interval are equally likely. The logic of the interval doesn't let us say that certain values are more likely than others. You don't know how far your single sample mean is away from μ: it could be very close or *just* fall within the range of 1,96 standard errors (in case of a 95% interval).

6.

a) A type I error is a false rejection of a null hypothesis that is actually true. Since Santa did not reject the null hypothesis, such an error didn't take place here. But it was a close call: the p-value was almost low enough to get a significant test result.



b) A type II error is a failure to reject a null hypothesis that is actually false. The p-value is only just above 5%, so when you get a value like this, it's possible that the test should have been significant but wasn't. However, if we assume for this exercise that the null hypothesis was true, a type II error was not possible.

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, we may assume for this exercise that the null hypothesis was true, and then power as a concept is 'not applicable'. Short answer to the question: no.

6.B EXERCISES FOR PIRATES

1.

a) $z = \frac{33-40}{12/\sqrt{26}} = -2{,}97$

Since Gulliver should look for the right-tailed p-value, this yields that

$$p = P(z > -2{,}97) = 1 - P(z < -2{,}97) = 1 - 0{,}0015 = 0{,}9985$$

Which is totally insignificant…

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

c) $z = \frac{45-40}{12/\sqrt{26}} = 2{,}12$

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

$$p = 2 * P(z > 2{,}12) = 2 * (1 - 0{,}9830) = 2 * 0{,}0170 = 0{,}0340$$

Still significant though.

d) Not entirely: this sample isn't random anymore. Gulliver could (perhaps accidentally) have geared his ad campaign to these specific 26 supermarkets which he already knew he was going to target. As such, the mean of this sample is a **biased** estimator of the population mean!

2.

"This is metaphysically impossible": CORRECT. It is fundamentally – even philosophically – impossible to make 100% surefire claims about a population, when you only draw a sample from it.
"Here is your 100% confidence interval…": CORRECT. This sarcastic answer provides an interval that must contain the population mean, but is completely useless in practice.
"If you want certainty…": CORRECT. Nothing to add here.
"We can be 100% certain…": INCORRECT. An extremely low p-value would make the null hypothesis very, very unlikely, but not 100% certainly wrong. Note: I made a typo here. The statement is actually supposed to read: "… 100% certain that the population mean is not **40**…". Sorry! I hope you still understood what I was trying to say.

3.

a) 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 determines the width, and it consists of the critical z-value and the standard error; these are both constant in a z-distribution. As a result, the margin of error won't change.

b) C
It depends a bit on how you define 'similar', so I have counted both answers as correct and hope that the question made you think. ☺ Higher confidence levels make the intervals broader, so they may overlap more. But they can still vary to a small or large degree, depending on the sample size, so the similarities don't have to become much stronger when you increase the confidence level.
Intervals based on larger samples are narrower and tend to be close to the real population mean, so most of them will be located in a relatively small range. This causes them to be similar more quickly.

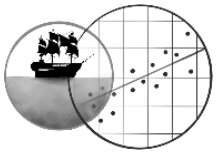
4.

a) $H_0: \mu = 32$
$H_A: \mu > 32$ if you wish to conduct a one-sided test (which seems defensible here). A two-sided test is always an option as well: $H_A: \mu \neq 32$.

b) A type I error (falsely rejecting the true null hypothesis). A larger sample cannot help prevent this, since the probability of a type I error exclusively depends on the significance level α. (It is thus usually 5% no matter what.)

c) A type II error (failing to reject the false null hypothesis).

d) That should not be too difficult if we use the fancy formula from section 6.5. Note: the answer is going to differ a bit depending on the alternative hypothesis you chose in exercise a. Why? Well, if you do a one-sided



test, the critical z-score is $z^* = 1{,}645$ but in a two-sided test, it equals $z^* = 1{,}960$. Remember that you can look this up in an Appendix table.
So here are both calculations. One-sided test:

$$z_\beta = z^* - \frac{|\mu_{true} - \mu_0|}{\sigma/\sqrt{N}} = 1{,}645 - \frac{|33 - 32|}{1{,}9/\sqrt{10}} = 1{,}645 - 1{,}664 = -0{,}02$$

According to the z-table, $\beta = 0{,}4920$. Note that the probability that you find directly in the table is always β. In other words: that's it! ☺
Two-sided test:

$$z_\beta = z^* - \frac{|\mu_{true} - \mu_0|}{\sigma/\sqrt{N}} = 1{,}960 - \frac{|33 - 32|}{1{,}9/\sqrt{10}} = 1{,}960 - 1{,}664 = 0{,}30$$

The z-table says that $\beta = 0{,}6179$. (Larger than in a one-sided test!)

e) The power is simply $1 - \beta$. For the one-sided test this becomes 0,5080, and it's 0,3821 for the two-sided test.

f) B
You don't really need to calculate this. If the true population mean was 34 grams instead of 33, the effect $|\mu_{true} - \mu_0|$ was bigger. The power increases in that case (see the end of section 6.5 for a further explanation). Still want the calculation to confirm it? Sure. We just need to do the same thing as in exercise d and e.
One-sided test:

$$z_\beta = z^* - \frac{|\mu_{true} - \mu_0|}{\sigma/\sqrt{N}} = 1{,}645 - \frac{|34 - 32|}{1{,}9/\sqrt{10}} = 1{,}645 - 3{,}329 = -1{,}68$$

According to the z-table, $\beta = 0{,}0465$ so the power equals $1 - \beta = 0{,}9535$. Larger than in exercise e!
Two-sided test:

$$z_\beta = z^* - \frac{|\mu_{true} - \mu_0|}{\sigma/\sqrt{N}} = 1{,}960 - \frac{|34 - 32|}{1{,}9/\sqrt{10}} = 1{,}960 - 3{,}329 = -1{,}37$$

The z-table says that $\beta = 0{,}0853$ so the power equals $1 - \beta = 0{,}9147$. Larger than in exercise e!

5.

a) This question requires a bit of recall from chapter 5. The hypothesis test can only be performed if the sampling distribution of the mean stroopwafel weight is normal. After all, the probabilities we look up in the z-table are based on a normal distribution. Are we sure that this requirement has been met? We don't know if the population distribution of the weights is normal, and the sample is only moderate in size so relying on the central limit theorem could be a bit tricky.

b) Tip: sketch the sampling distribution!
Cohen's d is estimated to be $\hat{d} = \frac{32{,}8 - 32}{1{,}9} = 0{,}42$. That's a small to medium effect, according to the guideline.
The z-score is

$$z = \frac{32{,}8 - 32}{1{,}9/\sqrt{20}} = 1{,}88$$

It follows from the z-table that

$$p = P(z > 1{,}88) = 1 - P(z < 1{,}88) = 1 - 0{,}9699 = 0{,}0301$$

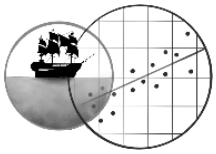
This test result is (barely) significant when you do a one-sided test. If you opted for a two-sided test, you should double this p-value and in that case it turns insignificant: $p = 0{,}0301 * 2 = 0{,}0602$. (For this reason, always pick your hypotheses before conducting the test!)

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

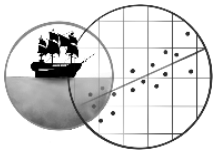
$$32{,}8 - 1{,}960 * 1{,}9/\sqrt{20} = 32{,}8 - 0{,}83$$
$$32{,}8 + 1{,}960 * 1{,}9/\sqrt{20} = 32{,}8 + 0{,}83$$
$$[\ 31{,}97\ ;\ 33{,}63\ ]$$

In other words, the population mean is probably (with 95% confidence) between the weights of 31,97 and 33,63 grams.
According to this interval, the null hypothesis $H_0: \mu = 32$ cannot be rejected because the value 32 falls between the boundaries. This outcome will contradict your answer to exercise b if you conducted a one-sided test; if you did a two-sided one, the outcome is consistent with exercise b. That's because confidence intervals can only be used as two-sided tests. (Another good reason to stick with two-sided testing, perhaps?)

6. 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 false – there's



no probability involved in that. The p-value only tells us how often you would draw a sample like the one you drew, if the null hypothesis was true.



CHAPTER 7-9 T-TESTS (COMPLETE)

9.A 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.

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

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

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

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

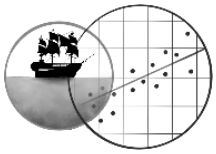
$$\bar{X}_d = \frac{0-1-2-4+2}{5} = -1$$

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: the same. This is -1, identical for both tests! I would not call that surprising, since you kept comparing the same two conditions.
- Standard error: smaller for the paired t-test. So 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	$d = X_{neutral} - X_{religious}$
$\bar{X}_1 - \bar{X}_2 = -1$ $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 1{,}923$		$\bar{X}_d = -1$ $s_d/\sqrt{N} = 1{,}000$

This typically happens when you match the participants well. The two groups can't differ because they have vastly different participants, so the difference scores usually have a smaller standard deviation than the separate scores.

- Degrees of freedom: smaller for the paired t-test. 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 participant groups who went to both vacation houses, not 10.)
- Power: greater for the paired t-test. Neither of the two tests are significant, but the paired t-test comes closest: it has the lowest p-value. This is due to the previous two points. Dependent samples have fewer degrees of freedom (by definition), which costs power, but they make up for this by strongly reducing the standard error (which *gains* power). The smaller standard error usually gives the power a decisive boost.[4]

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 measure 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.

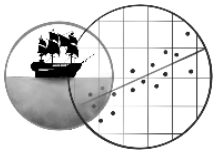
9.B EXERCISES FOR PIRATES

1.

This exercise is all about identifying the research design. Do we have one sample or two, and in the latter case, are the two samples independent or not? Here we go:

1. Repeated measures design (within-subjects): the same pilots were measured twice. So, a paired samples t-test.
2. Observation of one group. Thus, a one-sample t-test.
3. Two independent samples. Therefore, an independent samples t-test.
4. Be a bit careful. You might think this study compares two groups, but the behaviour of the old WIE Fighter was already known. R&D only drew a sample of the *new* model to test $H_0: \mu = 1100$, against $H_A: \mu < 1100$. Hence, a one sample t-test.

[4] 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.



5. Another tricky one. There were two independent samples… but each of them was also measured twice. So this is actually a so-called split-plot design (or mixed design): a combination of between- and within-subjects. That would normally be way too complicated for these chapters on t-tests. ☺ However, a simple solution was chosen here. The change between pre- and post-test was used as the dependent variable.[5] In this way, we 'erase' the within-subjects part of the design and can just compare the two independent groups on their change scores. Consequently, an independent samples t-test.

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 repeated measures design, so that looks good. But in option B, different fighter spacecraft were compared. Theoretically the research team might have matched individual WIE Fighters and SE-5 Seagulls, 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 WIE Fighters and the SE-5 Seagulls constituted independent groups. The study in option B then 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) We might expect that the training group will do better than the control group (which may not change on average). But the observer still preferred a two-sided test. So,

$$H_0: \mu_{change,training} = \mu_{change,control}$$
$$H_A: \mu_{change,training} \neq \mu_{change,control}$$

b) If both groups contained 40 pilots, the samples were large enough to make the t-test robust against non-normality. They were equally large as well, so the test would be just as robust against unequal variances. In that case we can certainly use the classic t-test that assumes equal variances.

c) Normality: the training group does not look normal at all (more like all over the place…), but it's still large enough that we don't have to care (29 participants). The control group looks reasonably symmetric – maybe a bit skewed to the right – but it still contains all 40 pilots, so that poses no problem either.
Equal variances: this assumption doesn't look good. The rule of thumb says that the training group probably had a larger standard deviation at population level $\left(\frac{50{,}111}{24{,}911} = 2{,}01 > 2\right)$, and the highly significant Levene's test confirms this ($p = 0{,}000$). The samples no longer had the same size due to the dropouts in the training group, so the independent t-test was <u>not</u> robust against this violation of equal variances. We should reconsider and use the Welch t-test that does not assume equal variances.

d) Here goes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-21{,}52 - 0{,}45}{\sqrt{\frac{50{,}111^2}{29} + \frac{24{,}911^2}{40}}} = \frac{-21{,}97}{10{,}105} = -2{,}174$$

If we use the simplified way to get the degrees of freedom, they become $df = smallest\ n_i - 1 = 29 - 1 = 28$. If you like to, draw the t-distribution with its mean of 0 and the t-value you just found. Now look up the p-value. It turns out that

$$0{,}01 < P(t < -2{,}174) < 0{,}02$$

However, we decided on a two-sided test. So double these bounds:

$$0{,}02 < p < 0{,}04$$

That's (just) significant: reject the null hypothesis.

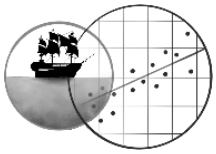
5. B

Namely, you could have resorted to… the confidence interval! The value 0 doesn't lie within this interval. 0 is thus <u>not</u> 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 <u>lower</u> for the trained pilots: they fiercely declined 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… 😐

[5] This is called a **gain score analysis**. For more information, see chapter 27 in **Parrrt 2** (but you may need to learn more about statistics first).



CHAPTER 10 ONE-WAY ANOVA (COMPLETE)

10.A EXERCISES FOR PEACHES

1.

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

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

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

$$\bar{Y} = \frac{2+7+6+1+10+8+8+10+4+6+5+5}{12} = \frac{4+9+5}{3} = 6$$

Now we can begin with the sums of squares. Are you not seeing what happens in these calculations? Have another look at section 10.4!

$$\boldsymbol{SS_{total}} = \sum_{ij} \left(\boldsymbol{Y_{ij}} - \bar{\boldsymbol{Y}}\right)^2 =$$

$$(2-6)^2 + (7-6)^2 + (6-6)^2 + (1-6)^2 + (10-6)^2 + (6-6)^2 +$$
$$(10-6)^2 + (10-6)^2 + (2-6)^2 + (2-6)^2 + (9-6)^2 + (7-6)^2 =$$
$$16 + 1 + 0 + 25 + 16 + 0 + 16 + 16 + 16 + 16 + 9 + 1 =$$
$$\mathbf{132}$$

$$\boldsymbol{SS_{between}} = \sum_{ij} (\bar{\boldsymbol{Y}}_{\boldsymbol{i}} - \bar{\boldsymbol{Y}})^2 =$$

$$(4-6)^2 + (4-6)^2 + (4-6)^2 + (4-6)^2 +$$
$$(9-6)^2 + (9-6)^2 + (9-6)^2 + (9-6)^2 +$$
$$(5-6)^2 + (5-6)^2 + (5-6)^2 + (5-6)^2$$

Remember that the group effect is calculated for every individual person. This is why the same group effect is taken four times in every condition. Actually, then, you can also formulate the calculation like this:

$$\boldsymbol{SS_{between}} = \sum_{i} \boldsymbol{n_i} * (\bar{\boldsymbol{Y}}_{\boldsymbol{i}} - \bar{\boldsymbol{Y}})^2 =$$

$$4 * (4-6)^2 + 4 * (9-6)^2 + 4 * (5-6)^2 =$$
$$4 * 4 + 4 * 9 + 4 * 1 =$$
$$\mathbf{56}$$

(Note that the j, for the participant, then vanishes from under the sum sign.)

$$\boldsymbol{SS_{within}} = \sum_{ij} \left(\boldsymbol{Y_{ij}} - \bar{\boldsymbol{Y}}_{\boldsymbol{i}}\right)^2 =$$

$$(2-4)^2 + (7-4)^2 + (6-4)^2 + (1-4)^2 + (10-9)^2 + (6-9)^2 +$$
$$(10-9)^2 + (10-9)^2 + (2-5)^2 + (2-5)^2 + (9-5)^2 + (7-5)^2 =$$
$$4 + 9 + 4 + 9 + 1 + 9 + 1 + 1 + 9 + 9 + 16 + 4 =$$
$$\mathbf{76}$$

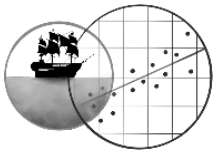
There, the worst is behind us. Now the ANOVA table. See the book chapter for the calculation rules. Look up the p-value in the Appendix; you'll find that it falls between 0,05 and 0,10.

	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) The F-test is not significant. There is no evidence that some mushrooms induce pleasant hallucinations more than others.

2. A

Fewer participants mean less power (see chapter 6) and that's a reasonable explanation for the non-significant ANOVA in exercise 1. Are any assumptions violated? Not terribly, it seems. 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 section 10.6 if you can't remember what the Bonferroni correction is for.



3. B

It's easy to apply the Bonferroni correction by hand, so that doesn't need to stop Marley. He 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

MS_{within} is always an unbiased estimator of the error variance; the same goes for $MS_{between}$ if the null hypothesis is true. This may be the case since the ANOVA is quite insignificant. It follows then that all alternatives must be correct (answer D). So why is the squared standard deviation of the Psycho Classic group also an unbiased estimator of the error variance? Well, it describes how test subjects within this group differ from each other – and this variance can only result from error effects. In fact, the squared standard deviation of the Fungus Fantasticus II group is an unbiased estimator of the error variance as well. (I just didn't list this one as an answer option.)

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 plus the treatment variance in the population; whereas MS_{within} is an unbiased estimator of the error variance only. If $MS_{between}$ is strikingly larger, that suggests that the treatment variance really exists. The question is: is the current result striking enough? That turns out not to be the case, as the ANOVA is not significant. The p-value is 0,214. This tells us that if the null hypothesis is true, then no less than 21,4% of the time will $MS_{between}$ be 1,594 times as large as MS_{within} (1,594 is the F-ratio).

10.B EXERCISES FOR PIRATES

1.

a) Most distributions of sample scores look rather ugly (tasty, sure, but their shapes are like hunchbacked grannies). But is that so odd? Actually not. A large number of customers simply buy no kringles, and if they do, they're of course not so bland that they buy just one tiny little kringle (save for one shabby student). Anyway, it is clear that the normality assumption has been violated. Fortunately, this does not hurt the ANOVA since all groups are large enough ($n_i = 30$).

b) The rule of thumb shows that the assumption has clearly been violated: $\frac{largest\ s}{smallest\ s} = \frac{4{,}693}{2{,}168} = 2{,}16 > 2$. The standard deviations look particularly larger in the two conditions where the advertisement contains a pun (*Krrringles*). This makes sense: some customers like puns more than others. Fortunately, all groups have the same size so the ANOVA is robust against this violation of equal variances.

2. A

MS_{within} describes the degree to which individuals within one and the same group 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.

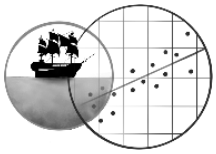
Answer B describes the square root of MS_{total}, a quantity which is commonly excluded from the ANOVA table since we don't need it. $MS_{total} = \frac{SS_{total}}{df_{total}} = \frac{2219{,}592}{119} = 18{,}65$. The square root of this is the standard deviation of all individuals relative to the general mean: 4,32. That agrees with the *Total Std. Deviation* in the *Descriptives* table! ☺

Answer C gives a rough impression of $MS_{between}$.

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 also by the size of the sample (N). One obtains more power when using larger samples, 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. $\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{593{,}558}{2219{,}592} \approx 0{,}267$. This indicates that the different advertisements seem responsible for 26,7% of the variation in the sales. Cohen's guidelines (see section 10.5) count this as a very large effect.



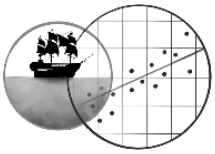
4.
The table below summarises all the answers.
In scenario 1, all group means are equal so there is no variation between the group means. We only have variation within each group. Thus there is no proof for a treatment effect at all. This is nicely expressed by the F-ratio, which equals 0. Note that F cannot be negative, so what is the p-value (the probability that F would be at least 0 if H_0 were true)? That's right: it equals 1, which is the least significant result we can get.
Scenario 2 is the opposite. In it, the variation in the sales exists only between the groups. This implies that the treatment is the only thing responsible for the variation in the sales. Error effects (sampling error) does not exist. So why should we even conduct a test in this hypothetical situation? The null hypothesis can clearly be rejected. In fact, a test is mathematically impossible: we cannot calculate the F-ratio since we would divide by 0. Mathematicians would say that as MS_{within} gets closer and closer to 0, F gets closer and closer to infinity. So in fact, you can see this result as the most significant one we could possibly have. Correspondingly, as F approaches infinity, the p-value approaches 0.
In reality, all ANOVAs will be somewhere in between these extremes.

	SCENARIO 1	SCENARIO 2
$MS_{between}$	$= 0$	> 0
MS_{within}	> 0	$= 0$
F-ratio	$= 0$	Cannot be computed; approaches infinity as MS_{within} approaches 0
P-value	$= 1$	Cannot be computed; approaches 0 as F approaches infinity
Null hypothesis	Cannot be rejected (seems true in fact)	Can be rejected (treatment is the only factor that affects sales)

5. B
Seeing that the ANOVA is significant and the null hypothesis can be rejected, there is demonstrably a group effect. This implies that $MS_{between}$ is an unbiased estimator of the error variance plus the treatment variance, whereas MS_{within} is an unbiased estimator of the error variance only. When drawing a sample, we can therefore expect $MS_{between}$ to be larger than MS_{within}, so that the F-ratio becomes larger than 1 as well. In other words, the expected value of F is larger than 1. It does not necessarily equal 14,115; that's just the F-ratio we observe in this single sample. The expected value, if you recall, is the average F-ratio we would obtain if we drew the sample again an infinite number of times.

5. B
Both A and C are actually saying: 'The probability that the null hypothesis is true equals 0,000.' However, there's no such thing as a probability that a null hypothesis is true[6]; the four groups simply have the same population mean or not. That's the 'truth'; there's no probability involved. We can only describe how often we would observe four samples with such divergent means as these, if the population means were actually the same – so if the null hypothesis were true. The current p-value says that we would virtually never see such a big difference between the sample means if they came from the same population.

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



6. A

Pay attention: we have 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[7] all the p-values, the following table results (I added the group numbers manually and highlighted the significant t-tests):

Multiple Comparisons

Dependent Variable: SALES

Bonferroni

(I) ADVERTISEMENT	(J) ADVERTISEMENT	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
(1) Kringles	(2) Krrringles	-2,567	,967	,054	-5,16	,03
	(3) Traditional butter kringles	-4,567*	,967	,000	-7,16	-1,97
	(4) Traditional butter krrringles	-5,900*	,967	,000	-8,49	-3,31
(2) Krrringles	(1) Kringles	2,567	,967	,054	-,03	5,16
	(3) Traditional butter kringles	-2,000	,967	,245	-4,59	,59
	(4) Traditional butter krrringles	-3,333*	,967	,005	-5,93	-,74
(3) Traditional butter kringles	(1) Kringles	4,567*	,967	,000	1,97	7,16
	(2) Krrringles	2,000	,967	,245	-,59	4,59
	(4) Traditional butter krrringles	-1,333	,967	1,000	-3,93	1,26
(4) Traditional butter krrringles	(1) Kringles	5,900*	,967	,000	3,31	8,49
	(2) Krrringles	3,333*	,967	,005	,74	5,93
	(3) Traditional butter kringles	1,333	,967	1,000	-1,26	3,93

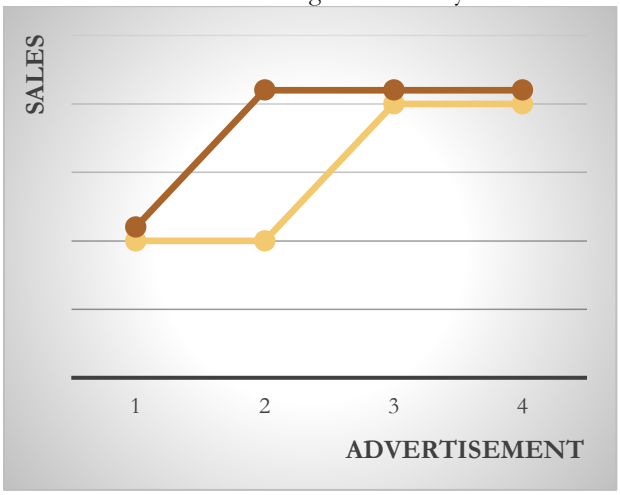
*. The mean difference is significant at the 0.05 level.

(Note that a p-value cannot exceed 1 or 100%, so some of them are capped at 1,000.)

After this rather heavy correction, only three unique comparisons remain significant:

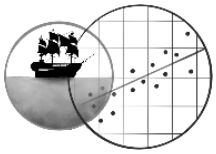
- The difference between advertisement 1 and 3;
- The difference between advertisement 1 and 4;
- The difference between advertisement 2 and 4.

This result contains something contradictory because it's impossible to plot the population means:

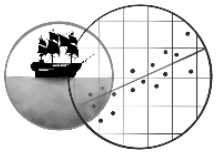


Advertisement 3 and 4 sell better than advertisement 1; so far, so good. Advertisement 4 also sells better than advertisement 2. But ad 3 is not better than 2, right? So does the dark pattern display the population means well?

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



But this pattern implies that ad 2 is better than 1, for which we have no evidence. So then we should draw the light pattern instead… but as we said, ad 3 is not better than 2. Contradiction! If advertisement 1 and 3 lead to a different number of mean sales, advertisement 2 cannot be equal to both of them.
How to explain this contradiction? The only explanation is that one of the pairwise comparisons contains a type I or type II error. Since advertisement 1 and 2 did differ significantly *before* the Bonferroni correction, the most plausible explanation is that we're making a type II error in this comparison now.



CHAPTER 11-14 CATEGORICAL TESTS (COMPLETE)

1.

a) We have two categorical variables: CONSIDERATION and PURCHASE. A χ^2-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**

$$H_0: \pi_{brief\ consideration} = \pi_{serious\ consideration}$$
$$H_A: \pi_{brief\ consideration} \neq \pi_{serious\ consideration}$$

Where π represents the population proportion of students who chose clothes. (Note: you can also let π stand for the proportion of students who picked Spencer's statistics book. This will lead to exactly the same z-score because the difference between the sample proportions will be identical. Feel free to try it if you have enough time to spare!)

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: the proportion of students who spend their money on clothes is demonstrably lower in the group that seriously considers their choice.

χ^2-test for contingency tables

$$H_0: in\ the\ population, there\ is\ no\ association\ CONSIDERATION \times PURCHASE$$
$$H_A: yes\ there\ is$$

The design assumptions are met (dependent variable categorical, independent groups). We're about to find out if the expected counts are large enough, but that seems likely considering the large samples.

So let's perform. First we make the contingency table:

observed		CONSIDERATION		
		at the beach	at the beach house	
PURCHASE	statistics book	66	26	**92**
	clothes	121	22	**143**
		187	**48**	**235**

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

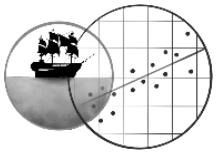
expected		CONSIDERATION		
		at the beach	at the beach house	
PURCHASE	statistics book	73,2	18,8	**92**
	clothes	113,8	29,2	**143**
		187	**48**	**235**

Now for the test statistic:

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

Check it: $z^2 = \chi^2$. Yes indeed: the z-test and the χ^2-test are in this case exactly the same – or data-equivalent, to use a difficult word. 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 chapter 18.

Anyhow, the χ^2-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 significant: the proportion of students who spend their money on clothes is demonstrably lower in the group that seriously considers their choice.

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 in the population, students who consider their choice seriously are between 3,2% and 34,6% more likely to choose Spencer's statistics book.

2.
We have a single variable with more than 2 categories. The table has $3 - 1 = 2$ degrees of freedom; a χ^2-test for goodness of fit is thus the only alternative.
There are 235 participants in total ($N = 235$). We calculate the expected counts using the formula $EC_i = N * \pi_i$:

$$EC_{can't\ wait} = 235 * 0{,}3 = 70{,}5$$
$$EC_{will\ see\ what\ my\ friends\ say} = 235 * 0{,}3 = 70{,}5$$
$$EC_{who\ is\ this\ guy} = 235 * 0{,}4 = 94$$

These are all amply large. Note that the observed counts show *fewer* students than expected who cannot wait, *fewer* than expected who will wait for their friends, and *more* students who have never heard of Devin Spencer. Do such deviations also exist at population level?
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. It's clear that Devin Spencer was less famous than Roan Rosch had hoped back in 2015.

3.

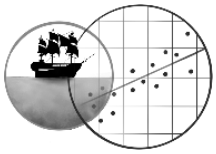
a) The test is significant. This happens when the observed counts differ a lot from the expected counts (and χ^2 becomes very large as a result).

b) Yes: it would be nice to conduct 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:

CONSIDERATION * EAGERNESS Crosstabulation

Count

		EAGERNESS		
		can't wait to read it	will see what my friends say	Total
CONSIDERATION	briefly	22	34	56
	seriously	23	32	55
Total		45	66	111



Chi-Square Tests

	Value	df	Asymptotic Significance (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 there's no indication that students who can't wait to read Spencer's book and students who will first see what their friends say differ in the way they consider their choice to buy the book or not. Both groups are equally serious about it (the sample percentages of serious considerations are 51% and 48%, respectively). Which is a good sign for Spencer (after all, a brief consideration leads more often to the purchase of clothes, as we saw in exercise 1).

CONSIDERATION * EAGERNESS Crosstabulation

Count

		EAGERNESS		
		can't wait to read it	who is this guy	Total
CONSIDERATION	briefly	22	87	109
	seriously	23	37	60
Total		45	124	169

Chi-Square Tests

	Value	df	Asymptotic Significance (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.

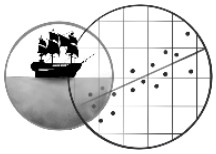
b. Computed only for a 2x2 table

After Bonferroni correction, it turns out that $p = 0{,}033$: significant. Students who can't wait to read the book are far more likely to consider their choice seriously (51%) than students who don't know Spencer (30%).

CONSIDERATION * EAGERNESS Crosstabulation

Count

		EAGERNESS		
		will see what my friends say	who is this guy	Total
CONSIDERATION	briefly	34	87	121
	seriously	32	37	69
Total		66	124	190



Chi-Square Tests

	Value	df	Asymptotic Significance (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

Bonferroni correction delivers $p = 0{,}033$: significant. Students who will see what their friends say are far more likely to consider their choice seriously (48%) than students who don't know Spencer (30%).

4. D

The student's age is a quantitative variable. χ^2-tests and z-tests for proportions are inadequate for this. Spencer needed to use an independent samples t-test or a one-way ANOVA.

5. C

Let's look at whether we have an association in the sample. This is easiest to express in terms of a z-test: by calculating proportions or percentages. Remember that relative frequencies are the same as percentages? So here's the table again:

		AUTHOR		
		multidisciplinary	specialised	
PRODUCT PLACEMENT	no	45 (75%)	35 (87,5%)	80
	yes	15 (25%)	5 (12,5%)	20
		60 (100%)	40 (100%)	100

This shows a clear association at sample level. The multidisciplinary authors used product placement twice as often, and that is the way Rosch expected it. So answer B is wrong… but we cannot pick A. After all, we don't yet know if this association is significant! It could be due to chance. With that, C must automatically be correct. If you want confirmation, you can proceed, though this is not strictly necessary for the exercise.
Let's say we conduct the χ^2-test anyway to be certain. Here are the expected counts:

expected count		AUTHOR		
		multidisciplinary	specialised	
PRODUCT PLACEMENT	no	48	32	80
	yes	12	8	20
		60	40	100

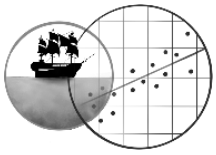
$$\chi^2 = \frac{(45-48)^2}{48} + \frac{(15-12)^2}{12} + \frac{(35-32)^2}{32} + \frac{(5-8)^2}{8} = \mathbf{2,34}$$

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.[8] In other words, Rosch could not reject the null hypothesis, but it was possible that he made a type II error. (This is always possible when you don't reject a null hypothesis. Of course, this error is less 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 Rosch could have 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!

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



CHAPTER 15 SIMPLE REGRESSION

1.

- The first statement is correct. The correlation indicates a weak relationship between mine transports and housing prices.
- The second statement is incorrect. This is a claim about the slope of the regression line – the appearance of the relationship. Remember that the correlation coefficient only describes the strength.
- The last statement is correct. The proportion of explained variation equals $R^2 = (-0{,}121)^2 = 0{,}015$, so mine transports explain only 1,5% of the variation in the housing prices. Other factors are responsible for the remainder of the variation.

2. C

(Of course, the last answer option is supposed to have the label 'C'. My bad!)
If no mine transports pass at all, $X = 0$. So we must be talking about the intercept housing price. What would be the correct prediction for the population? Just the intercept we find in the sample (answer A)? Nah: it's better to use the confidence interval for the population intercept. That's how you arrive at answer C.

3.

a) The slope b_1 of the regression equation equals -4,417. This means that one additional mine cart leads to a drop in the predicted price by 4,417 silver coins. So what do 20 mine carts do? They are associated with a price drop of $20 * 4{,}417 = 88{,}34$ silver coins.
b) This question should make you turn to the confidence interval. Just use the same logic as in question a and multiply the bounds by 20. The predicted change in the population should be

$$[\, 20 * -14{,}433 \,;\, 20 * 5{,}599 \,]$$
$$[\, -288{,}660 \,;\, 111{,}98 \,]$$

4.

The standard error is

$$s_b = \frac{s_{est}}{\sqrt{\sum_i (X_i - \bar{X})^2}} = \frac{389{,}099}{\sqrt{6071{,}345}} = 4{,}994$$

For the 95% confidence interval, we must first look up the critical t-value in the Appendix. We should round down the 53 degrees of freedom to 50 and then the critical t-value turns out to be 2,009. (As you can see, just using the value 2 as a rough number is usually okay for manual calculations.) With that the interval becomes

$$b \pm t^* s_b$$
$$-4{,}417 \pm 2{,}009 * 4{,}994$$
$$-4{,}417 \pm 10{,}033$$
$$[\, -14{,}450 \,;\, 5{,}616 \,]$$

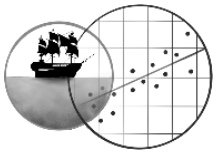
We understandably have some rounding error, but this nicely agrees with the interval in the SPSS output.

5. A

Declining prices are exactly as expected. It's true that I took out some values from the *Coefficients* table, but you have two other options to test the relationship between MINE TRANSPORTS and PRICE. First, there's the confidence interval for the slope that we've been talking about. It contains the value 0, so in the population, the effect of mine transports could be completely absent.
Second, the model ANOVA tests the same relationship and it yields the exact same result as the slope t-test when you have a simple regression model! And it turns out to be insignificant ($p = 0{,}380$).

6. A

The new *Coefficients* table shows that when the outlier was included, the relationship between MINE TRANSPORTS and PRICE appeared much more pronounced and it even was significant. So the outlier made Inthor overestimate the relationship between his variables: this relationship looked stronger. R^2 must therefore have been higher than in the main analysis.



CHAPTER 16 AGREEMENT (COMPLETE)

1.
For starters,

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

We can 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

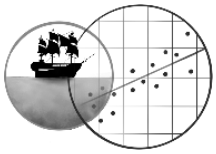
$$A_e = 1{,}25 + 1{,}25 + 1{,}875 + 1{,}875 + 0{,}625 = 6{,}875$$

In short, kappa becomes

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

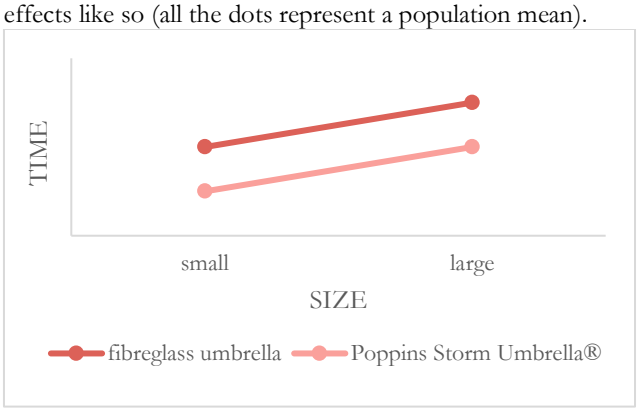
2. C
After all, kappa is downright pathetic. ☺

3. C
The colour is a nominal variable. You can also put its categories in a different order. Perhaps you will object to this by pointing at the spectrum of visible colours. Isn't orange more similar to red than green is? Well, kind of, but then I think that purple is again rather similar to red – even though red and purple are the farthest removed from each other. Assigning decent weight coefficients is going to be pretty difficult. We definitely can't use the standard method!

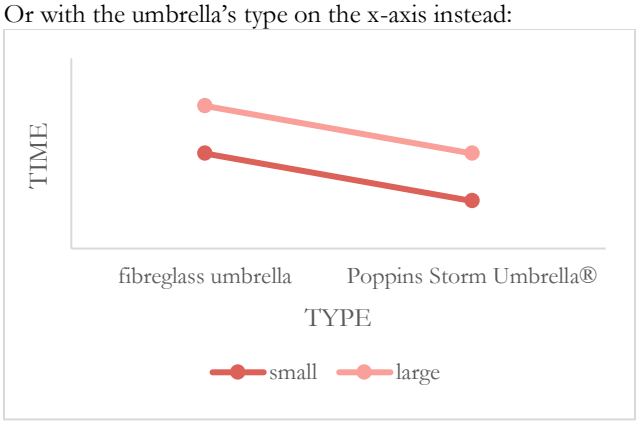


CHAPTER 21 BONFERRONI AND CONTRAST ANALYSES (COMPLETE)

1.
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:



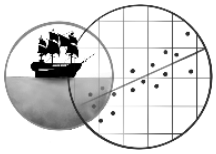
These are just schematic drawings, mind you. If you have other expectations, that's perfectly fine, as long as you can support them with solid theoretical arguments. ☺
In two-way ANOVA terms, I would expect a main effect of both the umbrellas' SIZE and TYPE, but no interaction.

2.
The best alternative, in my opinion, would be an adjusted Bonferroni correction. This is possible since the ANOVA is very significant ($p = 0{,}000$); it demonstrates that at least 1 group must differ from the other 3, and as such it protects 3 of the 6 pairwise comparisons against type I errors. If we apply adjusted Bonferroni, therefore, we should multiply the p-value of each comparison by 3. If you don't want that, a basic 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. B
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. Do you wish to understand why these



are the correct coefficients for interaction? I think that's a nice bonus, but check out the graphs I made for exercise 1 (you may need some knowledge from chapter 22). If there's no interaction, the lines run parallel so the umbrella size has the same effect regardless of the type. In other words, the difference in average TIME when using small versus large umbrellas should be the same for both types. Or, $H_0: \mu_1 - \mu_2 = \mu_3 - \mu_4$. Rewrite that to $H_0: \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$ 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 require no 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.

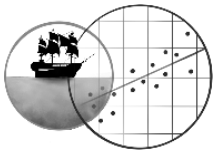
Answer C is therefore the worst answer you could've picked. The cubic contrast is statistically significant, but realistically meaningless. So is the linear term. The quadratic contrast isn't even statistically significant, so B still contains an erroneous statement. A is best, because it's fully correct.

6.

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

b) C

After all, the sample results (means) 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 22 TWO-WAY ANOVA (COMPLETE)

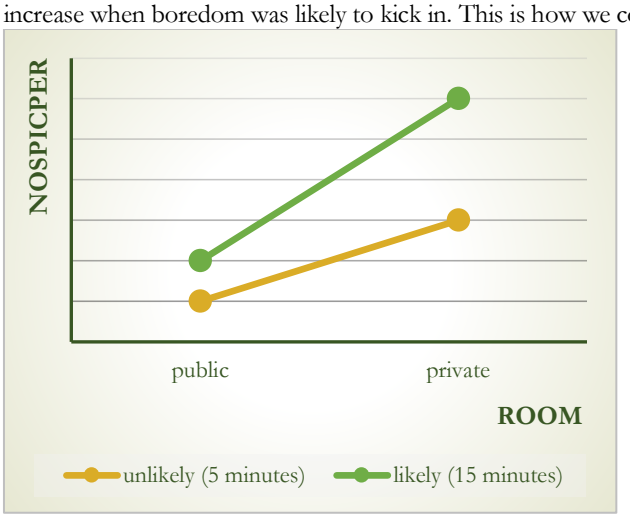
22.A EXERCISES FOR PEACHES

1. C

NEUROTICISM would be a constant if everyone in the study had the exact same level of neurotic traits. That doesn't sound likely. It is a variable that differs between individuals, which makes it belong to the bunch of factors that remain.

2.

Fingerheber expected that the mean NOSPICPER would be higher in the private room, but also that this difference would increase when boredom was likely to kick in. This is how we could sketch her expectations:



The graph makes it clear that she predicted an interaction effect.
(You could say that she also predicted a main effect of ROOM, but it would be more interesting to study the simple effects of ROOM when interaction is present.)

3.

a) Here are the means in a table:

		BOREDOM		
		unlikely (5 minutes)	likely (15 minutes)	
ROOM	public	3	7	**5**
	private	9	13	**11**
		6	**10**	**8**

b) SS_{total} compiles how everyone deviates from the grand mean:

$$SS_{total} = \sum_{ijk} (Y_{ijk} - \bar{Y})^2 =$$

$$(6-8)^2 + (0-8)^2 + (3-8)^2 + (11-8)^2 + (9-8)^2 + (9-8)^2 + (17-8)^2 + (9-8)^2 =$$
$$4 + 64 + 25 + 9 + 1 + 1 + 81 + 1 =$$
$$\mathbf{186}$$

Next up is SS_{ROOM}:

$$SS_{ROOM} = \sum_{ijk} (\bar{Y}_i - \bar{Y})^2 =$$
$$(5-8)^2 * 4 + (11-8)^2 * 4 =$$
$$\mathbf{72}$$

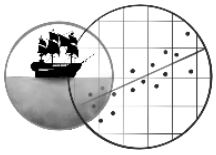
Then $SS_{BOREDOM}$:

$$SS_{BOREDOM} = \sum_{ijk} (\bar{Y}_j - \bar{Y})^2 =$$
$$(6-8)^2 * 4 + (10-8)^2 * 4 =$$
$$\mathbf{32}$$

Finally the interaction term:

$$SS_{ROOM*BOREDOM} = \sum_{ijk} (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2 =$$
$$(3-5-6+8)^2 * 2 + (7-5-10+8)^2 * 2 + (9-11-6+8)^2 * 2 + (13-10-11+8)^2 * 2 =$$
$$0 * 2 + 0 * 2 + 0 * 2 + 0 * 2 =$$
$$\mathbf{0}$$

The variation explained by the model as a whole is of course $SS_{(corrected)\ model}$:



$$SS_{model} = 72 + 32 + 0 = 104$$

We can simply add up the main effect and interaction sums of squares, because the design is orthogonal – each combination group had only $n_{ij} = 2$ participants.
Here's the SPSS output to confirm everything:

Tests of Between-Subjects Effects

Dependent Variable: NOSPICFREQ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	104,000[a]	3	34,667	1,691	,305
Intercept	512,000	1	512,000	24,976	,008
ROOM	72,000	1	72,000	3,512	,134
BOREDOM	32,000	1	32,000	1,561	,280
ROOM * BOREDOM	,000	1	,000	,000	1,000
Error	82,000	4	20,500		
Total	698,000	8			
Corrected Total	186,000	7			

a. R Squared = ,559 (Adjusted R Squared = ,228)

c) There is no interaction effect – not even at sample level! You can see this easily if you draw a plot of the sample means, but it also becomes apparent from the consequence: the interaction sum of squares equals zero. This term may therefore be taken out; the corresponding F-test won't be significant anyway.

d) SS_{error} must be the variation that remains unexplained:

$$SS_{error} = SS_{total} - SS_{model} = 186 - 104 = \mathbf{82}$$

You can see that it agrees with the SPSS output. ☺

e) The total degrees of freedom amount to 7. Each main effect we estimate requires 1 degree of freedom here, as well as the interaction effect:

$$df_{ROOM} = i - 1 = 2 - 1 = \mathbf{1}$$
$$df_{BOREDOM} = j - 1 = 2 - 1 = \mathbf{1}$$
$$df_{ROOM*BOREDOM} = (i - 1)(j - 1) = 1 * 1 = \mathbf{1}$$

Hence, for the error term, 4 degrees of freedom remain:

$$df_{error} = df_{total} - df_{model} = 7 - 3 = \mathbf{4}$$

f) If each combination group had only 1 participant, N would be 4, and then $df_{total} = 3$. This would leave us with 0 degrees of freedom for the error term:

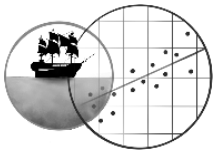
$$df_{error} = df_{total} - df_{model} = 3 - 3 = 0$$

In such a case, SS_{error} cannot be calculated, let alone MS_{error} (you can't divide the sum of squares by 0). And that makes sense: the effect of remaining factors is estimated by seeing how individuals in the same treatment group still differ from each other. But if we have only 1 person in each treatment group, we cannot compare any persons who had the same treatment…

g) This sample shows no interaction effect, so we could take it out of the model. The degree of freedom from the interaction term will then return to the error, and so will the sum of squares. Granted, this solution would only work if the interaction sum of squares was not 0, and in such a case, removing it would be controversial. In short, all effects need degrees of freedom in order to be estimated.

4.
The first and the last statement are true; the others are false.
Fingerheber had complete control over her independent variables: she could simply assign an equal number of participants to each combination group of ROOM and BOREDOM. This gave her the opportunity to set up an orthogonal design – which is recommended in all cases. ☺
What is important, then, is whether the two independent variables show a relationship. The second and third statement describe them individually, but their individual behaviour doesn't matter. We need to make a contingency table of the sample sizes[9]:

[9] Note that this is a different table than the one for exercise 3, which describes the means of the groups on the dependent variable NOSPICPER.



		BOREDOM		
		unlikely (5 minutes)	likely (15 minutes)	
ROOM	public	34 (50%)	34 (50%)	68 (50%)
	private	34 (50%)	34 (50%)	68 (50%)
		68 (100%)	68 (100%)	136 (100%)

Clearly, the independent variables were perfectly balanced: the relative ROOM frequencies were the same in each BOREDOM group.

5. A

Answers B and C describe effectively the same thing: that the design is orthogonal. But that's not what this is about. Interaction and confounding are completely unrelated! The fact that we have no interaction does not mean that the group factors don't overlap (i.e. that they cannot confound each other); instead, it means that they do not change each other's *real* effects on the dependent variable NOSPICPER.

6.

First of all, the interaction effect was not significant. Thanks to the orthogonal design, Fingerheber could immediately look at the main effects; she did not have to remove the interaction term first. The main effects were both significant.
Next, would she need pairwise comparisons? Not per se: both group factors have only two levels. The pairwise comparisons would simply provide the same p-values for the main effects of ROOM and BOREDOM as the ANOVA does. But the pairwise comparisons could still give her confidence intervals, which would be a nice extra.
If we look back at the sample means (check the *Descriptive Statistics* or the *Profile Plot*), nose-picking is clearly a more popular activity in the private room (the significant ROOM effect in the ANOVA tells us that this is also the case at population level). In addition, long waiting times (i.e. boredom) lead to more frequent nose-picking as well. Interestingly, we have a main effect of BOREDOM here, so apparently this effect also manifests in public spaces. I wonder what kind of population Fingerheber has been investigating… Might this be typical behaviour of people who often eat take-away food?

22.B EXERCISES FOR PIRATES

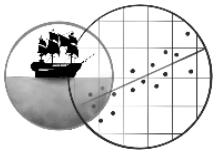
1.

a) This was a somewhat non-orthogonal design. That's because the participants were no longer distributed across the conditions at an entirely constant ratio:

		STOMACH	
		weak	strong
FAST FOOD	no	19	16
	yes	16	20

This isn't odd; it's not an experiment after all. The participants were observed in their natural environments (ahem), so King & McDonald had no control over the combination group sizes. Now the factors FAST FOOD and STOMACH display an association, and the risk of confounding exists. We can see these frequencies in the *Descriptive Statistics* table.
To correct for potential confounding, SPSS calculates 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: $991{,}650 + 306{,}344 + 210{,}712 = 1508{,}706 \neq 1417{,}286$.
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.

b) 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. If we work with my favourite method to check normality, we can see that multiple skewness and kurtosis values fall outside the [-1, 1] bound. The kurtosis of the first combination group is especially bad (3,837). 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.



c) Equal variances: this assumption is obviously violated. The largest and the smallest standard deviation differ strongly $\left(\frac{9,667}{2,701} = 3,58 > 2\right)$. Is the ANOVA robust against this violation? Well, the groups differ in size somewhat, but not too drastically. Again, I'm inclined to say that I would perform the ANOVA, but only trust extreme p-values (below 0,01 or above 0,10) If your lecturer is stricter, dear reader, your conclusion is allowed to sound: 'No, the ANOVA is not robust and we actually shouldn't perform it.'[10]

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 ($p = 0,027$). C, finally, speaks of an association between the two independent variables, STOMACH and FAST FOOD (so not of an interaction effect). This is a rather silly claim which is tested nowhere in the output.

Incidentally, C is even false at sample level: the people with a weak stomach ate fastfood slightly less often $\left(\frac{16}{35}\right)$ than the people with a strong stomach $\left(\frac{20}{36}\right)$ (in other words, the design is non-orthogonal).

3. B

First look at the interaction term in the full model. 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, as it is now, main effects may be rendered meaningless. We should study simple effects in such a scenario.

4. A

$SS_{FAST\ FOOD}$ is equal to 300,538. This is the variation that FAST FOOD uniquely explains, when overlap with STOMACH is removed. Consider FIGURE 22.9 in the book chapter: different variables, same kind of model. If we take FAST FOOD out of the model, this unique bit of explained variation (which does not overlap) will go to the error term. So we can simply add $SS_{FAST\ FOOD}$ to SS_{error}.

5. B

You will probably conclude – correctly – that FAST FOOD has an effect at population level: the p-value of this main effect is 0,011. But did it also have an effect in the sample? Certainly! How can you prove that an effect exists in the population if it doesn't even show up in the sample? There are a few ways in which you can see this main effect:

- The sum of squares for FAST FOOD is larger than 0, meaning that there is at least some variation between the group means of the fast food conditions. (If you wanted, you could calculate an eta squared using this and the total sum of squares: $\eta^2_{FAST\ FOOD} = \frac{SS_{FAST\ FOOD}}{SS_{corrected\ total}} = \frac{300,538}{4172,958} = 0,072$. That's a medium-sized effect at sample level.)
- One page earlier, we find the estimated marginal means of the two fast food groups. They differ by 4,1 bathroom visits. The two-way ANOVA tests if this difference is significantly large.

Note: in a sample, you will virtually always see all effects to some extent, just because it's a sample and its groups are likely to differ due to chance. Wouldn't it be interesting if these 35 patients who ate no fast food had the *exact* same TOILET average as these 36 fastfood eaters? The statistical test is meant to assess whether the differences were caused by more than mere chance.

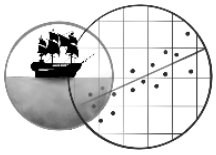
6. B

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 FAST FOOD, for the two STOMACH groups separately. Among people with a weak stomach, a clear effect of FAST FOOD shows itself when we compare the sample means (8,74 versus 16,38 bathroom visits). This simple effect is significant ($p = 0,009$ if we assume equal variances, $p = 0,013$ if we don't[11]). Among people with a strong stomach, the FAST FOOD groups differ less strongly (the means are 4,69 and 5,40) and this difference is not significant ($p = 0,605$ if we assume equal variances, $p = 0,583$ if we don't). These results are consistent with the interaction effect we found in exercise 2. The required Bonferroni correction (both p-values times 2) does not change the conclusions.

[10] 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 far for a regular statistics course.

[11] "Should we assume equal variances?"
If you check the assumption using the rule of thumb from chapter 20, it holds (also in the strong stomach group). On the other hand, my pro-tip from chapter 9 in Parrrt 1 recommends not to assume equal variances when you do an independent t-test – unless the two groups are equally large.



CHAPTER 23 ANCOVA (COMPLETE)

23.A EXERCISES FOR PEACHES

1. A

Mind the research design! The PARENT TYPE groups already existed before the (quasi-)experiment began, so the average FITNESS might differ considerably between the samples. We can expect FITNESS to also affect the DISTANCE a child could traverse, so this would distort Khan's measure of the PARENT TYPE effect.
It's true that, in quasi-experiments, the covariate often explains some error variance in the dependent variable as well. As such, you would think that including the covariate provides more power. Unfortunately, correcting for confounding also *costs* power (we lose degrees of freedom and the sum of squares of the PARENT TYPE effect may decrease). The net result is usually a power loss.

2.

All statements are correct, but the second one cannot be concluded from the *Coefficients* table; when we make this one, we already assume that we can draw straight lines through the scatterplot. Here's how you can verify everything:

- Fitter children are usually quicker at travelling through the jungle: within each group, FITNESS has a positive slope b_1. In addition, these (simple) slopes are all significant.
- The linearity assumption of ANCOVA is met: it seems that the straight lines in the scatterplot capture the relationships pretty well. The *Coefficients* table contains nothing to tell you this, as the values it contains are already based on the straight lines that have been drawn through the point clouds.
- The non-interaction assumption of ANCOVA might be violated: the lines in the scatterplot don't run quite parallel… in other words, they have different slopes. The *Coefficients* table indicates that the three sample slopes are 0,139, 0,162 and 0,112. What we don't know yet is if these slopes are different in the PARENT TYPE populations as well. We should still test this using a SCHMANCOVA. (I did that in fact, and no worries: the interaction effect was not significant.)
- FITNESS would be a useful covariate: as we said before, it has a significant relationship with DISTANCE in all groups. So it could be a confounder if the groups don't have the same mean FITNESS score – and it's wise to play it safe in a quasi-experiment. Even if FITNESS does not confound the PARENT TYPE effect, its significant relationship with DISTANCE indicates that it explains a significant slice of error variance.

3. B

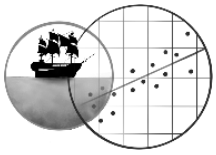
Tough question, eh? Let's start by comparing the results of the ANOVA versus the ANCOVA. PARENT TYPE has a (visible) p-value of 0,000 in both models, so its effect is quite significant. But does it have the same magnitude? Let's compare. The sum of squares between groups in the ANOVA is 224,958 (which, incidentally, amounts to a huge η^2 of 0,331). But the sum of squares for the PARENT TYPE in the ANCOVA is only 78,449 (or 0,115 in terms of an η^2). In other words, it reduces to roughly one third(!) of its original size when we correct for differences in FITNESS. The difference between the sums of squares, 146,509, is the part that overlaps with the FITNESS effect.[12] So we can argue that FITNESS was definitely a confounder: after correcting for it, we still conclude that PARENT TYPE has an effect, but we become much more modest about its impact.
Two further points:

- You can also use the sample means to come to a similar conclusion about the reduced effect size. Check the raw (uncorrected) sample means in the output of the one-way ANOVA, and compare them with the estimated marginal means that the ANCOVA produces. Notice the change? The estimated marginal means (which are corrected for differences in FITNESS) are more similar. ☺
- The conclusions of the study actually *do* change somewhat when we use the corrected pairwise comparisons after the ANCOVA! More on that later as well.

4. A

Your drawing should look like FIGURE 23.7 in the book chapter. Put FITNESS (the covariate) on the x-axis, DISTANCE (the dependent variable) on the y-axis, and make three lines this time (one for each PARENT TYPE group). Now indicate the point of reference as a vertical line: it was 9,69, according to the footnote below the estimated marginal means table. Continue by plugging in the means (schematically). The orangutans group had the highest estimated marginal mean, so the top line belongs to this one; the humans group is in the middle, and the sloth bears group at the bottom. Lastly, it's time to draw the uncorrected sample means from the one-way ANOVA in your drawing – and then you will find the answer. The sloth bears group had an uncorrected mean of 4,451, which is lower than the estimated marginal mean of 5,048… and this forces you to draw it further to the left on the regression line. Likewise, the orangutans group had an uncorrected mean of 8,011, but an estimated marginal mean of 7,242 – so the uncorrected mean lies higher, or further to the right on the regression line.

[12] You can add $SS_{\text{PARENT TYPE}}$, SS_{FITNESS} and the overlap to obtain $SS_{\text{corrected model}}$: $78{,}449 + 123{,}670 + 146.509 = 348{,}628$.



5. A

You can basically use your drawing from exercise 4 again. The intercepts are the points where the lines pass the y-axis. We could as well compare the (corrected) mean DISTANCE of the groups here, rather than at the average FITNESS level of 9,69. In short, the intercept is just another point of reference that we could use! If each group has a different intercept, this indicates a PARENT TYPE effect at sample level. And the *Coefficients* table reveals that the groups had different intercepts indeed.

Different slopes don't imply a main effect of PARENT TYPE; rather, they indicate interaction (the FITNESS effect on the DISTANCE depends on the PARENT TYPE), as we already discussed in exercise 2. Lastly, answer C is rather meaningless since a single participant doesn't tell the whole story. We are more interested in means.

6. B

It has already become clear that PARENT TYPE has an effect, whether or not we correct for FITNESS. But which groups really had different population means? Let's study the pairwise comparisons. It's wiser to use those that come with the ANCOVA (these pairwise comparisons are based on the estimated marginal means). The orangutans group traversed a somewhat longer mean DISTANCE than the humans group (7,242 versus 6,773 kilometres), but this difference is not significant ($p = 0{,}927$). The mean of the sloth bears group, however, is significantly lower than the other two sample means. Conclusion: don't let sloth bears adopt your human child.

Answer C can be picked when all pairwise comparisons are significant (but this is not the case here).

Bonferroni correction has already been applied to the table, by the way (check the footnote in the output).

23.B EXERCISES FOR PIRATES

1. C

You may 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 PRE-EMOTIONALITY 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.

If PRE-EMOTIONALITY *was* likely to be a confounder (quasi-experiment), answer B is the wrong way to think. A variable can confound your model if you don't include it. You include it in order to take its effect into account and correct for that.

2.

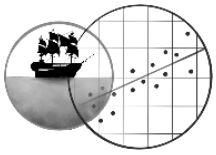
a) It's an ANOVA, which can only compare groups (categories). The independent variable must therefore be MUSIC, as the *Descriptives* table makes clear as well. The model heading indicates that the dependent variable is the pre-test (PRE-EMOTIONALITY). Thus, this ANOVA tests whether the three music groups had the same population mean on this pre-test. It turns out that they probably did, since the test is not significant at all ($p = 0{,}919$).
b) They did not, it would seem.
c) No: MUSIC and PRE-EMOTIONALITY are (almost) orthogonal. 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 level-headed 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 as a trait before the experiment began, this can explain to a great extent why they their emotional states differed directly after. The pre-test thus explains a good deal of variation in the post-test scores. It's a very typical covariate for experiments.

3. A

Here we go again, just like in the exercises for peaches. Put the dependent variable (EMOTIONALITY) on the y-axis and the covariate (PRE-EMOTIONALITY) 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 a pre-emotionality equal to 51,54 (this is reported below the estimated marginal means tables). This is the average pre-test score in the entire sample. If everyone had been 10 points more emotional before the computer task, the ANCOVA model also predicts higher emotionalities after the task (look at the regression lines). All groups then obtain a higher estimated marginal mean. Since the ANCOVA assumes non-interaction, however, the lines run parallel: hence the three estimated marginal means all go up to the same degree. Their differences will thus stay the same.

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



4. B

The scatterplot in Model 3 shows the sample result: the three regression lines don't run parallel in the sample, and so the slope coefficients b_1 are not exactly equal. In short, light interaction occurs. This interaction turns out to be barely significant with $p = 0{,}046$ (see Model 4).

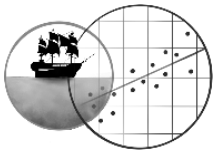
5.

a) B
Since the interaction effect in Model 4 is significant, we should not remove it in order to study the main effect of MUSIC. Model 5 does the latter.

b) C
Model 5 finds no significant main effect of MUSIC ($p = 0{,}147$), but that doesn't say much, since we found a significant interaction in Model 4.

6.

Yes, we should change our answer. Use the graph in Model 3 for a good impression. The lines disperse when PRE-EMOTIONALITY increases, so the estimated marginal means would not only increase if we shifted up the point of reference, but their differences would grow as well. In short, you should now pick answer B.



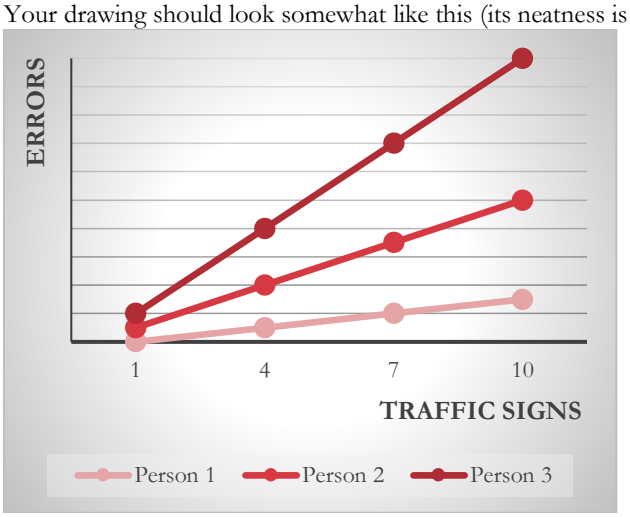
CHAPTER 24 WITHIN-SUBJECTS ANOVA (COMPLETE)

1. A

Leaving out the PERSON variable causes SS_{error} to remain far too large. We also assume way too many degrees of freedom, by the way: there are way fewer than 56 independent observations within the groups, since each set of 4 observations belongs to the same person.
Between-subjects ANOVA does not assume sphericity, only equal variances in the individual conditions.

2. C

Your drawing should look somewhat like this (its neatness is obviously exaggerated):



First consider how the scores change from 1 traffic sign to 4 traffic signs. Not much; they all go up only a little. In other words, these difference scores are pretty similar so they have a small variance.
Now check how the scores change from 1 to 10 traffic signs. The change (difference score) is pretty small for person 1, but it's really big for person 3. This means that these difference scores between 1 and 10 signs have a much larger variance!
In other words: interaction between PERSON and the within-subjects factor almost always leads to a violation of sphericity. (Exceptions are mathematically possible, but they would be theoretically weird.) Since such interaction is commonplace, sphericity gets violated more often than it gets met.

3. C

Mauchly's test is meh. I would not use it. Just study the epsilon estimates in the same table. They are all pretty low; in the book chapter, I indicated that you can use 0,7 as a guideline. Epsilon values below this threshold indicate a severe violation of sphericity.
Answer A makes an interesting allusion. It is true that the standard deviations in the *Descriptive Statistics* table are very different… but what standard deviations are these? They belong to the separate conditions… but sphericity is about the standard deviations of the <u>difference</u> scores.
(To be fair, when the separate conditions have different variances, the same usually goes for the difference scores.)

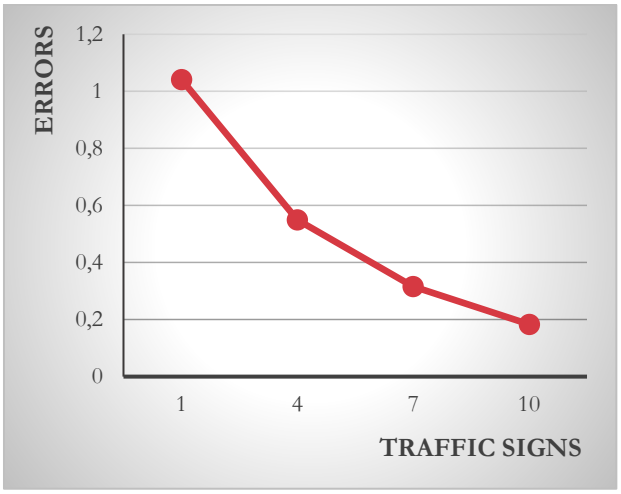
4. C

The F-test with lower-bound correction is significant. This is a heavier correction than Greenhouse-Geisser, so Greenhouse-Geisser is bound to yield a p-value lower than lower-bound. It must therefore be significant as well.

6.

a) C

Researcher Clarkson was expecting a <u>trend</u>: a curvilinear trend, to be precise. The number of errors should increase ever faster when more traffic signs were present. Since TRAFFIC SIGNS is a scale in disguise, we can use the polynomial contrast analyses provided in the output. The linear and quadratic contrast test are both significant ($p = 0{,}000$ and $p = 0{,}002$), which implies that the trend follows a <u>partial U-curve</u> that goes up or down overall.
So how do we find out the exact shape of the curve? By looking back at the sample results. You can use the means from the *Descriptive Statistics* table or the estimated marginal means (they are the same here). If you plot them, they look like this:

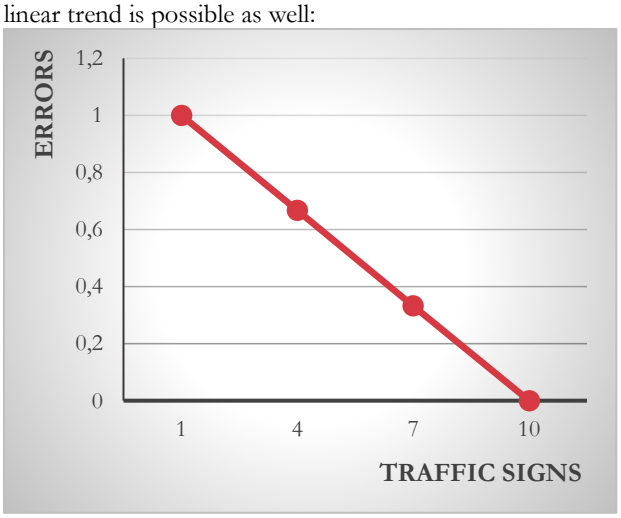


… and that is actually a gradual *decrease* of the errors. Hm! That's the opposite of what Clarkson expected! True story: as it turns out, when a traffic situation *appears* really dangerous, road users become much more careful and *fewer* accidents happen. That's traffic psychology for you. ☺

b) C

The univariate ANOVA 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 four conditions have the population mean, against the alternative that at least one mean is different. If the number of errors increases faster and faster with an increasing abundance of road signs, all population means must be different. We cannot demonstrate the latter using the ANOVA alone.

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 (7 signs) differ. Then something like in my previous graph may be going on, but the following linear trend is possible as well:



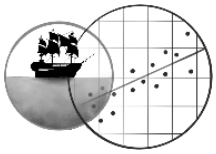
That's why the pairwise comparisons also fail to be specific enough. On top of that, they're less powerful due to a necessary Bonferroni correction.

The quadratic trend Clarkson expected *can* be studied directly in the *Tests of Within-Subjects Contrasts* table.

5.

Only the last statement is correct:

- The mean differences are not averaged, only the standard errors.
- It's true that assuming sphericity gives more power; all pairwise comparisons in *GLM Univariate* assume more degrees of freedom. But if we average the standard errors, that means some of them will become smaller… while others will become larger. This causes the corresponding pairwise comparisons to have a higher p-value. So not all individual p-values will necessarily be smaller when we assume sphericity. They will only tend to be smaller on the whole.



CHAPTER 25-26 TWO-WAY WITHIN-SUBJECTS AND SPLIT-PLOT ANOVA

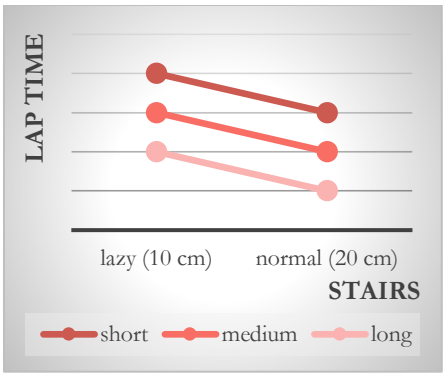
26.A EXERCISES FOR PEACHES

1.

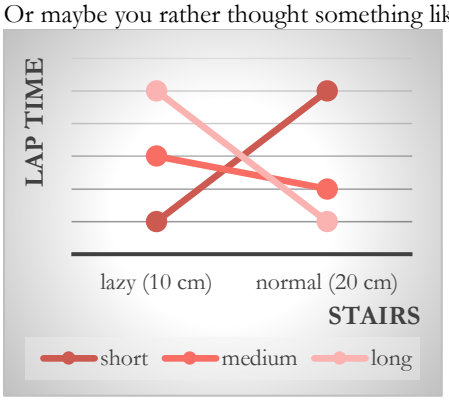
a) LAP TIME.

b) LEGS: between-subjects factor (length differs between persons). STAIRS: within-subjects factor (all persons try both stairs). This means a split-plot (or mixed) 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. C

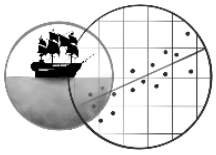
Box's test evaluates the assumption of equal covariance matrices. Its null hypothesis is that all of these matrices are equal at population level, and this is not rejected. But what does it imply to have equal covariance matrices? It implies that that the three leg groups (so the different groups of the between-subjects factor):

- Have the same variance in lap times on the lazy staircase;
- Have the same variance in lap times on the normal staircase;
- Have the same covariance (correlation) between lap times on the lazy and normal stairs. This is exactly what answer C says.

Remember: the variances within one matrix don't have to be equal. The ANOVA does not require the variances on the lazy and normal staircase to be the same. We only need equality between matrices. This is why A is wrong.
Finally, what is the correlation between the lap times of the different leg groups? It's 0. The leg groups are independent!

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 here.



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. 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.

6.

Because we find a significant interaction effect, we shouldn't look at main effects… but the pairwise comparisons test the main effect of LEGS! 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 stairs.

26.B 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. For instance, have a look back at the superhero study in book chapter 24: 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

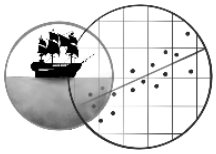
'If the MOUTH increases, the ORDERS also increase?' Um, what? MOUTH is a categorical variable! Trend analyses don't make any sense. You can only use them if the independent variable represents a scale. See chapter 24 for details.

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 $p = 0{,}000$).

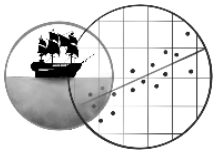
BOTTOM has 2 levels, so the ANOVA result tells us enough. It's still useful to check out the estimated marginal means though, to see which condition completes the most ORDERS on average. That's the chili peppers condition, as we established already in exercise 1.

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. Then only one significant difference remains, between empty mouths (1) and chili (2). The other differences are not significant (anymore)



In a scientific report, these results are often supplemented with the mean differences and their confidence intervals, plus effect size estimates such as Glass' Δs.

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



CHAPTER 29 CONTINGENCY TABLE ANALYSIS (COMPLETE)

29.A EXERCISES FOR PEACHES

1.

a) If we take spitting on the picture as a sign of intolerance or disgust, she will have expected that the llamas which lived in a mixed environment had lower odds of spitting.

b) Given the coding scheme, it makes sense to construct the contingency table like this:

		ENVIRONMENT		
		white (0)	mixed (1)	
SPIT	didn't spit (0)			
	spat (1)			

So the category of interest always gets the 1, while other category gets the 0. Domínguez hypothesised that the odds of spitting would be lower in the mixed environment group, which implies a negative relationship. The odds ratio would then have to be smaller than 1.

I've noticed that this is a complicated question. But in a way, that's precisely the point: the behaviour of the odds ratio, and therefore the answer, depends entirely on the coding scheme. So let's always use the same coding principle of 0 = 'no' and 1 = 'yes'! That way we can communicate odds ratios without confusing each other with their meaning.

c) It could be that the llamas from poor farmers in Dominguez' sample lived in a mixed environment more often than the llamas from rich farmers, or less often. This can happen due to sampling error, or because the population is unbalanced (which is often the case). Income can also have an impact on your tolerance toward refugees, since you have less food to share. This may confound the environment effect.

d) Perhaps the environment would especially make llamas from poor farmers more tolerant towards refugees, since llamas from rich farmers already lived a wealthy life and felt less threatened by unexpected newcomers to begin with. In that case, the environment effect would be stronger or weaker depending on the farmer's income.

2.

a) A separate contingency table of the two independent variables will help us with this. So make it real quick, and ignore the dependent variable SPIT for this part of the exercise.

		INCOME	
		low	high
ENVIRONMENT	white	487	379
	mixed	120	38

The image looks clear to me already: the llamas from rich farmers almost never lived in a mixed herd (38 out of 417), while roughly 20% of the low income group did (120 out of 607). An odds ratio indicates as well that there's a strong negative relationship between the two Xs:

$$OR_{ENVIRONMENT*INCOME} = \frac{A * D}{B * C} = \frac{487 * 38}{120 * 379} = 0{,}407$$

Confounding is therefore possible!

b) Best to keep them separate. If you merge them, INCOME can confound the main effect of ENVIRONMENT.

c) Back to the contingency table to calculate the simple effects. Below I express the simple effects of ENVIRONMENT in two odds ratios:

- Low level of education: $OR_{SPIT*ENVIRONMENT} = \frac{398*8}{112*89} = 0{,}319$
- High level of education: $OR_{SPIT*ENVIRONMENT} = \frac{360*1}{37*19} = 0{,}512$

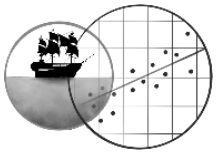
These two are not equal, so we've got ourselves an interaction effect. Llamas from mixed herds were less likely to spit on the picture than those from all-white herds (a negative relationship), but this effect was stronger (i.e. more negative) in the low income group.

d) Yes. Its effect was negative in both income groups. This means that if the interaction effect turns out to be non-significant later, we may average the simple effects of ENVIRONMENT into one (negative) main effect.

3.

a) No. INCOME is the control variable here; that's because it's found on the left side of several tables. These tables split the data on the control variable. Typical of contingency table analysis is the fact that the effect of the control variable can't be tested.

b) We must redo the analysis and swap INCOME and ENVIRONMENT.



4. C

Of course, these tests *could* have given a different result. Only the *Tests of Conditional Independence* correct for the potentially confounding effect of the control variable. Thus, I hereby advise against the use of the *Total* block from the *Chi-Square Tests* table.

5. A

The odds were half as large, not the probability. That's similar but not quite the same.

6. B

This analysis tests the effect of ENVIROMENT (see exercise 3), so answer A can be dropped. First check if the interaction effect we found in the sample is significant. The Tests of the Homogeneity of the Odds Ratio say it's not ($p = 0{,}668$). An analysis of the main effect of ENVIROMENT is therefore appropriate. We see in the *Mantel-Haenszel Common Odds Ratio Estimate* table that llamas who lived in a mixed group were racist less often (negative relationship, $OR = 0{,}337$), and this effect turns out to be quite significant ($p = 0{,}003$).
Answer C describes an interaction effect, which we disqualified.

29.B EXERCISES FOR PIRATES

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: $OR_{HAPPY*SES} = \frac{32*43}{26*51} = 1{,}04$

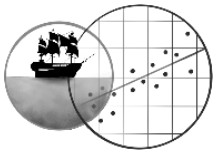
   All in all, we see an effect in both groups: the odds ratio is slightly above 1, so there's a positive association. People with a high SES are happy slightly more often. Mind: this may have resulted from sampling error (or from background variables altogether, since this is just an observational study).

d) The fact that the simple effects (the odds ratios) differ each time means that we have some light 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. 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. C

The dependent variable and the primary independent one need to be dichotomous: odds ratios can only be calculated for 2*2 tables. However, when we wish to look at the effect of PAPER TOWELS, it is possible to control for more than two SES groups. If we find interaction, we can just study the simple odds ratio of the PAPER TOWELS effect within each SES group. No interaction? Then we can average out all the simple effects and study the main effect of PAPER TOWELS (in the form of the Mantel-Haenszel common odds ratio).

5. B

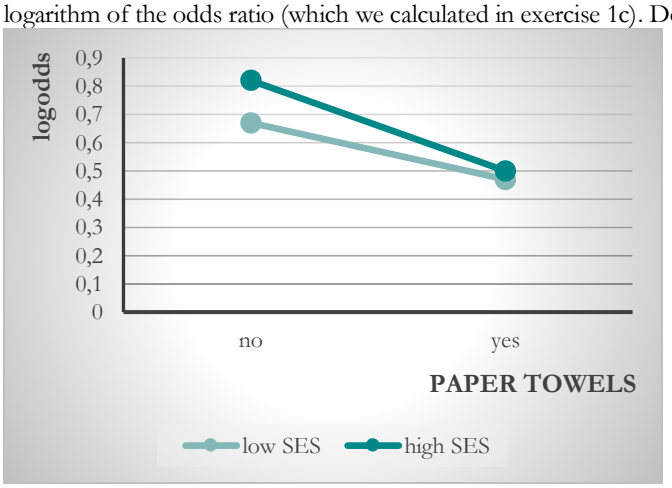
As we said earlier in exercise 3, the effect of SES is not tested in this output. A can therefore be dropped. In the sample we do see the interaction effect C mentions, but this effect is not significant (*Tests of Homogeneity of the Odds Ratio*: $p = 0{,}791$). With that, B remains and indeed: the main effect of PAPER TOWELS on HAPPY is not significant. You can see this in the *Mantel-Haenszel Common Odds Ratio Estimate* table ($p = 0{,}222$), but also in the *Tests of Conditional Independence* table (Cochran's: $p = 0{,}222$, Mantel-Haenzel: $p = 0{,}266$).

6.

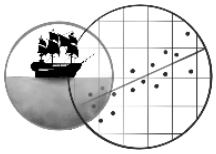
Here's the contingency table again, now with all the odds and logodds included:

	low SES		high SES	
	no paper towels	paper towels	no paper towels	paper towels
unhappy	49	32	40	26
happy	96	51	91	43
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*	**0,67**	**0,47**	**0,82**	**0,50**

We can make a graph of these logodds as displayed below. The slope of the line is consistently equal to the natural logarithm of the odds ratio (which we calculated in exercise 1c). Don't mind the minor rounding error.



- Low SES: $\log OR = \log 0{,}81 = -0{,}21 \approx 0{,}47 - 0{,}67$
- High SES: $\log OR = \log 0{,}73 = -0{,}31 \approx 0{,}50 - 0{,}82$



CHAPTER 30 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 31-A); this would render the analysis impossible. (SPSS will automatically throw a superfluous dummy out.)

4. C

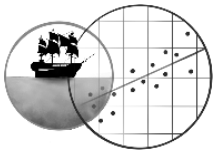
What is asked for is a comparison between group 2 and 3. However, group 4 is the reference group in the current coding scheme. The dummy variables thus compare group 1, 2 and 3 against group 4; they cannot make any other comparisons.

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 39, section 39.1, and bridge the gap between ANOVA and regression! ☺



CHAPTER 31-A MULTIPLE REGRESSION: MAIN EFFECTS

1.

a) $\hat{Y} = 20{,}668 + 0{,}205 * NEUROTICISM - 0{,}429 * PRESSURE - 0{,}140 * AAAH!$

b) This is the predicted cortisol concentration (20,668 micrograms per decilitre) for a woman who is not neurotic whatsoever (0), has a shower without any water pressure (0) and with an extremely sensitive cold water knob which basically changes the temperature if you turn it 0 degrees. Would someone like that exist? If so, she had better buy a new shower.

c) If a woman scores 1 point higher on the NEUROTICISM scale, her CORTISOL concentration is expected to rise by 0,205 micrograms per decilitre. This is the expected rise when we keep PRESSURE and AAAH! constant, so when we assume that this person's shower stays the same.

2.

a) Outliers in the y-direction: yes, because at least one studentised residual is greater than 3.
Outliers in the x-direction: no, since the centred leverage values should all be below $\frac{3(\#X+1)}{N} = \frac{3*4}{30} = 0{,}40$ and the biggest leverage value in L'Irréel's data equals 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 her neurocitism, water pressure and AAAH! score. Judging from the Cook's distances, this woman has an overly strong impact on some regression coefficients. (Note: I'm assuming that the influential person(s) is/are also the one(s) with the high residual(s).

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 has not been modelled quite adequately in the regression analysis. Some distortion is the consequence: the persons in the upper right corner pull the regression line toward themselves.

3. B

The PRESSURE predictor has a tolerance value which almost equals 1: it thus hardly overlaps with AGGRESSION and AAAH!. AGGRESSION does correlate substantially with the other two predictors, just like AAAH!: both have tolerance values around 0,6. Since we know that they don't overlap with PRESSURE, they must overlap with each other.

4. 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.
Alternatively, you can study the part correlations (after all, correlations are standardised quantities as well). They lead to the same conclusion.

5. C

This simple regression model would not have corrected the AGGRESSION effect for overlap with PRESSURE and AAAH! (the latter in particular, as we learnt from exercise 3). In other words, it would be based not on the part correlation of AGRESSION with CORTISOL (0,058), but on the zero-order correlation (0,444). This zero-order correlation is more than seven times as large as its corrected part sister, so it's likely that the t-test on the effect of AGGRESSION would have been more significant.
(Note that this statement is not completely certain: the multiple regression model with all predictors also contains less unexplained (residual) variance, which means that the standard error of the AGGRESSION effect may be smaller in this model. That would cause the AGGRESSION effect to be somewhat more significant in the multiple model again. It's all a matter of how things balance out.)

6.

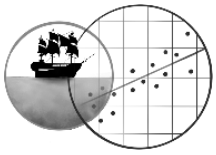
The model ANOVA tells us that the first two statements are true. Let's calculate the proportion of explained variation:

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{515{,}768}{1114{,}141} = 0{,}463$$

This is almost 50% indeed.
The multiple correlation is simply the square root of the above:

$$R = \sqrt{0{,}463} = 0{,}680$$

The last statement is not necessarily supported. Although the ANOVA is highly significant, this only tells us that the model contains at least one useful predictor. In fact, the *Coefficients* table makes it clear that only AAAH! is significantly related to stress levels.



CHAPTER 31-B MULTIPLE REGRESSION: INTERACTION AND SIMPLE SLOPES

1. C
D

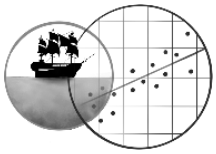
2. B
Rub-a-dub-dub

3. A
Uitrekenuh

4. B
Niet schreeuwen aub

5.
   a) Schaal niet normaal
   b) Spiegeldenken

6. C
Yups



CHAPTER 32 LOGISTIC REGRESSION

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 slope (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 at population level.
In addition, the conclusion which follows from model 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. To be fair, this is not too likely a scenario, 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. Then again, some birds might 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 slope b_1 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 windows. 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 as well. 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 Hosmer and Lemeshow test for model 3 is not useful anymore. It effectively tests whether you should add an interaction term or a curvilinear term to the model. Well, no, you don't need to do that – because model 3 *already has* an interaction term. The Hosmer and Lemeshow test is only interesting for models that have not been saturated yet.
Lastly, 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 C a no-go area.

3. B

The regression equation is as follows:

$$\log odds = b_0 + b_1 W + b_2 S + b_3 WS$$

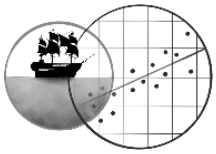
I have used W for the WINDOW predictor and S for the SPECIES predictor.
When we fill in the bs (using the *Variables in the Equation* table), we obtain:

$$\log odds = -0{,}547 - 0{,}840W + 0{,}898S + 0{,}488WS$$

Now we can calculate the logodds that a headless lark will crash into a normal window. WINDOW = 0 and SPECIES = 1 for this situation:

$$\log odds = -0{,}547 - 0{,}840 * 0 + 0{,}898 * 1 + 0{,}488 * 0 * 1 = 0{,}351$$

The higher the logodds, the greater the probability of a crash. Remember the tipping point? When the logodds is greater than 0, the probability is… greater than 50%.



4.

a) Let's follow the book chapter, and look at the regression equation. We take out the components that make up the WINDOW effect:

$$\log odds = b_0 + b_1 W + b_2 S + b_3 WS$$

For pigeons, the SPECIES variable equals 0. This turns the WINDOW effect into

$$b_1 W + b_3 W * 0 = \boldsymbol{b_1 W}$$

So for pigeons, tinting the window changes the logodds by b_1. That's equal to -0,840, according to the SPSS output.

b) This odds ratio is simply e raised to the power of the simple WINDOW effect: $OR = e^{b_1} = e^{-0{,}840} = 0{,}432$.

c) In the case of headless larks, SPECIES = 1 so the WINDOW effect becomes:

$$b_1 W + b_3 W * 1 = \boldsymbol{b_1 W + b_3 W}$$

So for larks, tinting the window changes the logodds by $b_1 + b_3$. That equals $-0{,}840 + 0{,}488 = -0{,}352\ldots$ a smaller WINDOW effect than for pigeons.

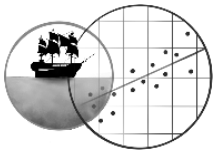
d) You'd have to raise e to the power of this specific WINDOW effect: $OR = e^{b_1 + b_3} = e^{-0{,}840+0{,}488} = 0{,}701$.

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. Answer C discusses a main effect of the SPECIES and basically says that it does not exist, but the main effect of SPECIES in model 2 is highly significant ($p = 0{,}000$). 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. Birds therefore crash into a tinted window less often than into a normal window.

6. B

The Mantel-Haenszel common odds ratio 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 SPECIES in a model without an interaction term. This is model 2 of course. We find the odds ratio for the relationship between CRASH and SPECIES under *Exp(B)*: 3,083.



CHAPTER 34 POWER ANALYSIS

1.

a) This is a comparison of two independent groups on a quantitative dependent variable. In other words, we should conduct a power analysis for an independent t-test. We may assume equal variances (the standard deviation equals 3 for both animals), so 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. Let's go for equal sample sizes.

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

So yeah, the outcome is that you'd need less than one cat and dog to prove this with 99,9% power. We can still add 2 to this result, since we're using the rule of thumb formula.

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) This is a comparison of two independent groups as well, but now on a dichotomous dependent variable. That implies a power analysis for a z-test for 2 proportions.
First we need Cohen's d again:

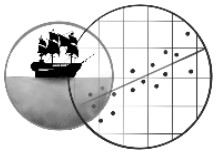
$$d = \frac{|\pi_1 - \pi_2|}{\sqrt{\frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{2}}} = \frac{|0{,}05 - 0{,}35|}{\sqrt{\frac{0{,}05*0{,}95+0{,}35*0{,}65}{2}}} = \frac{0{,}30}{0{,}371} = 0{,}81$$

Now the sample size calculation:

$$n_1 = n_2 = 2 * \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{d}\right)^2 = 2 * \left(\frac{1{,}645 + 1{,}282}{0{,}81}\right)^2 = 26{,}17 \rightarrow 27$$

This makes 54 cat-dog pairs in total (27 per group).

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.



CHAPTER 35 FACTOR ANALYSIS

NOTE
Due to a baffling mistake on my part, the items in the SPSS output have Dutch labels. 😐 My apologies! Here's a translation overview:

blijdschap	joy
aarzeling	hesitation
enthousiasme	enthusiasm
stress	stress
wanhoop	despair
sarcasme	sarcasm
woede	anger
verdriet	sorrow
medeleven	compassion

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 have had a sufficiently high eigenvalue, namely 1,279.

3. 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 uniqueness, a minor 10%. This is very little! ☺

4. A
The communality describes how well the variance in the separate items is explained (see exercise 3); 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.

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

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

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

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.

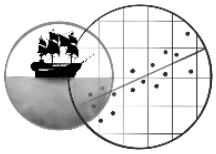
5.
- The structure has become simple, which it wasn't before rotation

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 ($communality = a^2$). 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 ($\hat{r} = a_1 a_2 \cos\alpha$). This statement is false.

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

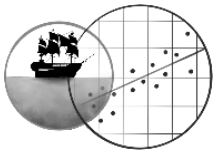
Oblique rotation allows the factors to correlate (orthogonal rotation does not). That makes this statement correct.

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 these items 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 36 RELIABILITY

1.

a) I think not: rather than measuring pyromania and sadism, item IV rather measures how much the chemists 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) Item IV is not parallel to the others; it clearly measures a different trait.
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.

- The participants have scored too high on it: correct. 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.
- The participants' scores on this item are too similar: correct. The standard deviation of item IV is very limited as well. Nearly all scientists must have indicated 4 or 5 points – they (fully) agreed with the statement that molecules are mega-cool. Of course they did – they're chemists. Thus, we encounter a ceiling effect. This item fails to discriminate the participants properly; it cannot tell us how they differ in terms of their passion for the field. Which turns the item into a pretty useless psychometric instrument.
- Without item IV, Cronbach's alpha will rise substantially: correct. We saw this already in the previous exercise.
- Without item IV, the questionnaire will become more reliable: incorrect. Thought this was the same statement as the third one? ☺ But what did we conclude again in exercise 1d? We said that Cronbach's alpha was going to underestimate the reliability of the questionnaire. You should be aware that Cronbach's alpha does not *equal* the reliability; it's just an *estimate* of the reliability. So by removing item IV, we don't actually make the survey more reliable. Instead, the value of Cronbach's alpha just gets closer to the *true* reliability. We remove (most of) the bias from the estimation.
  Also consider that *adding* items can make a test more reliable; removing them will do the opposite at worst (and do nothing at best, such as in this case).

3. B

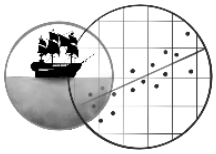
Cronbach's alpha *(Reliability Statistics)* is too low; it should be 0,70 at least, and preferably 0,80. The researchers might achieve this by extending the questionnaire with more (parallel) items. The current items – save for number IV – appear pretty parallel and each one correlates nicely with the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

4. C

Of course the items are parallel, but this does not guarantee that they correlate perfectly. No, *if* they are parallel, then the correlation tells us how reliable they are. After all, that correlation can still weaken due to random measurement errors. So the fact that the correlation is perfect tells us that no measurement errors were made. And this makes sense: there are a few things that we can measure without error, such as a person's age. (All chemists knew their exact date of birth for sure. As always, there are exceptions: elderly people may sometimes misremember, and orphans did not always have their date of birth registered.)

5.

a) The average grade that a chemist's students score on the CSE may depend on the cohort. Every year, the chemist may teach a different pool of students who are a bit smarter or dumber (to put it bluntly). In addition, the difficulty of the CSE is likely to vary a bit from year to year. In short, the average grade will fluctuate over the years.
b) Hiroshima had only one grade for every chemist (participant). This means that he could not calculate a Cronbach's alpha. He could, however, study how the grades fluctuated over time. He had the option to request the CSE grades from two subsequent student cohorts (years), and to see how the chemists' grades from the first cohort correlated with the grades from the second cohort. In short, he could look at the retest reliability.
c) He could request the CSE grades from multiple cohorts, and take the averages of those multiple cohorts. How well do a chemist's students score on average across three years rather than one? This provides a more reliable image of their educational skills.
d) Both the questionnaire and the average CSE grade are not perfectly reliable. Hence, when we correlate them, that correlation will come out lower than the true correlation; uncorrelated measurement errors are in the mix after all. This is the attenuation effect.



e) For this we use the attenuation formula. The reliability of the questionnaire is still estimated at 0,569 (Cronbach's alpha in the Supplement). Let's call the sum scores on the questionnaire X and the average CSE grades Y.

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

f) A positive trend ensues: chemists with stronger sadistic and pyromaniacal tendencies are better teachers sometimes as well. The relationship is not perfect, however.

g) In principle, we want to know whether the true correlation is larger than 0 at population level. I would therefore prefer to test the true correlation. Only, 0,45 has to really be the true correlation from the sample in that case; in other words, the attenuation formula must be correct for this situation. It is only correct if the assumptions of the true score model have been met, and here they are not (item IV from the questionnaire is not parallel to the others, which makes the Cronbach's alpha of 0,569 that we plugged in an underestimate). In that case, you had best be careful and just test the observed correlation.

6. C

A is the wrong answer anyway: Fremmels' 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)

$$\sigma_{e_s}^2 = (1 - \rho_{SS'}) * \sigma_s^2 = (1 - 0{,}569) * 10{,}576 = 0{,}431 * 10{,}576 = 4{,}558$$

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

$$\sigma_{e_{s\,difference}}^2 = 2 * 4{,}558 = 9{,}116$$

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

$$SEM = \sqrt{\sigma_{e_{s\,difference}}^2} = \sqrt{9{,}116} = 3{,}02$$

Assuming that the measurement errors are random (their mean is 0) and normally distributed, we can state that 95% of the time we compare two individuals, the difference between their true scores falls within the range

$$measured\ difference \pm 2 * SEM$$
$$23 - 19 \pm 2 * 3{,}02$$
$$4 \pm 6{,}04$$
$$[\mathbf{-2{,}04}\,;\mathbf{10{,}04}]$$

Hence, the true difference may also amount to 0 points. Hiroshima and his peers could not prove that Fremmels was truly more sadistic than Interlinn. This might have been possible with a more reliable questionnaire.