University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Midterm Exam
March 4, 2019

Aids allowed (printed or handwritten): My lecture overheads (slides); Any notes that you have taken in this course; Your marked assignments; My assignment solution; Non-programmable, non-communicating calculator.

This exam has 23 numbered pages of questions. Check to see that you have all the pages. There is an additional empty page that you can use if you need more space for any answers.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question).

The maximum marks available for each part of each question are shown next to the question part.

Your code should use `tidyverse` and `ggplot` ideas, as used in this course.

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Question 1 (14 marks)

Drivers in the Vancouver area suspect that the price of gas depends on the community they are in. One driver collected data on the price of a litre of regular gas at three randomly-chosen Chevron, Esso and Shell stations in each of three communities. The data are shown in Figure 2 in the booklet of Figures, as laid out in the file, and the file is stored as `vancouver_gas.txt` in your current folder in R Studio.

(a) (3 marks) Give R code to read in the file to a data frame `gas` and to display (some of) that data frame.

My answer: Data aligned in columns, so `read_table`:

```
gas <- read_table("vancouver_gas.txt")
```

```
## Parsed with column specification:
## cols(
##   community = col_character(),
##   station = col_character(),
##   price = col_double()
## )
```

```
gas
```

```
## # A tibble: 27 x 3
##    community station price
##    <chr>     <chr>   <dbl>
##  1 Langley   Chevron  93.9
##  2 Langley   Chevron 102.
##  3 Langley   Chevron 102.
##  4 Langley   Esso    102.
##  5 Langley   Esso     93.9
##  6 Langley   Esso    102.
##  7 Langley   Shell   104.
##  8 Langley   Shell   105.
##  9 Langley   Shell   102.
## 10 Surrey    Chevron 102.
## # ... with 17 more rows
```

Two marks for getting `read_table` right, and one for remembering to display your data frame in some fashion. It's not `read_tsv` because the gas prices (look at the ones less than 100) are *right*-justified, and if it were tab-separated, (i) the columns would be left-justified, or (ii) might not even be lined up, depending on the width of the text:

```
d <- read_tsv("vancouver_gas.txt")
```

```
## Parsed with column specification:
## cols(
##   `community  station  price` = col_character()
## )
```

This might *look* all right, but it's *one* column of text, called `commmunity station price`:

```
summary(d)
```

```
##  community  station   price
##  Length:27
##  Class :character
```

Question 1 continues...

```
##  Mode  :character
```

Other things:

- trying `read_tsv` is not unreasonable but doesn't work (as discussed above). One out of two.
- `read.table` is not `tidyverse`, but it does work here (perhaps a bit luckily). 1.5 out of 2, if you remembered the `header=TRUE`; expect a further deduction if you missed that.
- `read_delim` does not work because the values are separated (at least sometimes) by more than one space. Zero out of two.

Displaying the data frame ought to be a gimme point! Make sure you don't miss out.

(b) (2 marks) We want to make a plot showing how gas prices compare among the different communities, ignoring which station each price comes from. Explain briefly why a boxplot is suitable for this.

My answer: We have two variables, one of which is quantitative (gas price) and the other of which is categorical (community).

One point for saying one quantatitive and one categorical without naming the variables of each type. If you don't name the variables of each type, you are not showing that you understand completely what's going on. Maybe *you* know what's going on, but the grader cannot see what's in your head, only what is on the paper.

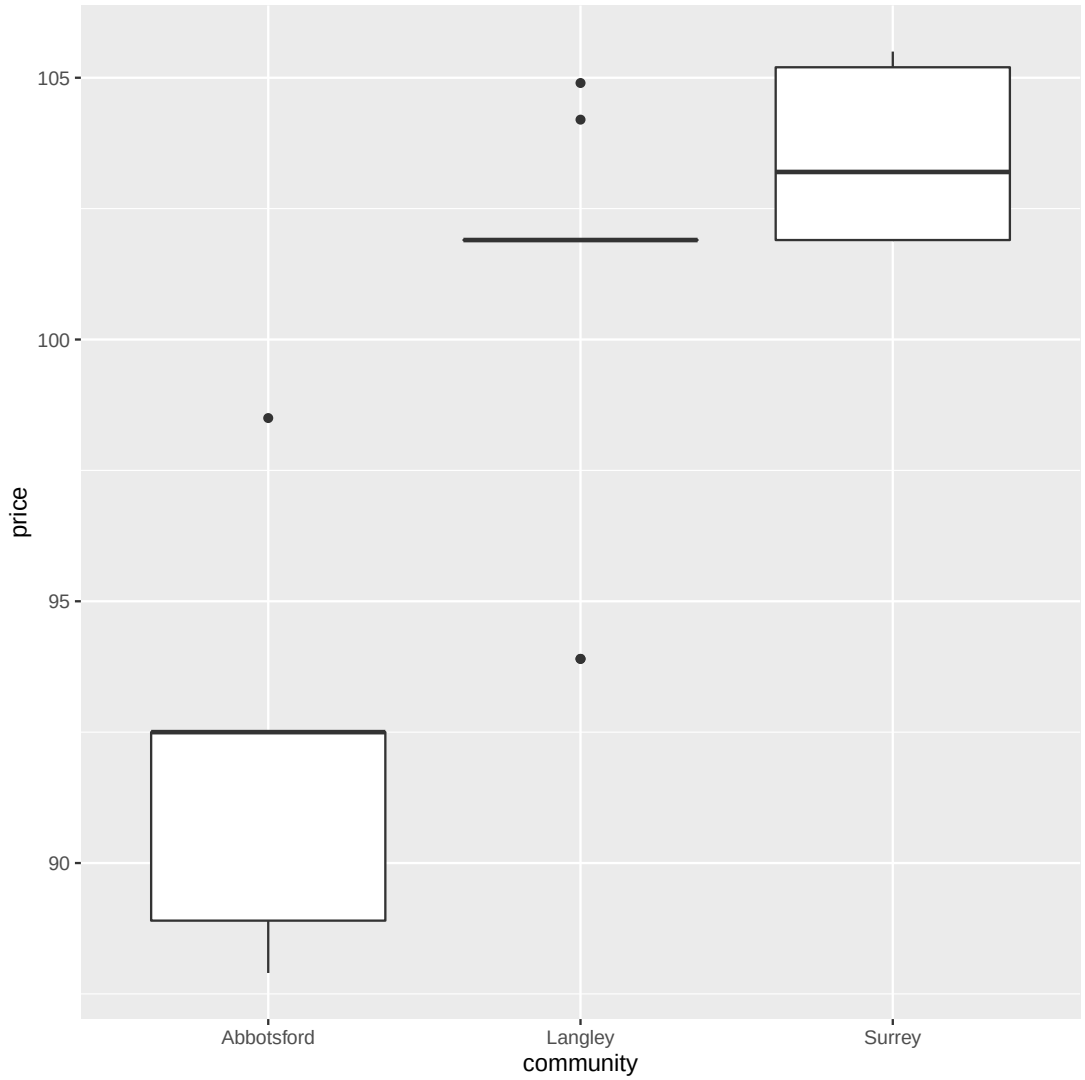
One point for saying something less specific but reasonable, such as "a boxplot enables us to compare medians and spreads for different groups". This is true, but it doesn't really get at why a boxplot rather than something else. If you talk about *means* rather than medians, expect to lose a half point, because this is not what boxplots show. Likewise, be sure to distinguish between a *categorical variable* and its *levels*: `community` is one categorical variable with three levels (not three categorical variables). Also, you'll lose a further half point (at least) if you talk this without saying that you're comparing different communities, or at least comparing something.

Think about what you've learned *in this course*.

(c) (2 marks) Give R code to make a boxplot as described in the previous part.

My answer:

```
ggplot(gas, aes(x=community, y=price)) + geom_boxplot()
```



Making this pretty easy for you.

If you wrote something like `geom_box` or something else that wouldn't actually work but suggests the right thing, minus 0.5. If you drew some other plot entirely, or got one of the variable names wrong, minus 1.

(d) (2 marks) The boxplot drawn by your code is shown in Figure 3. Is there a community where gas is noticeably cheaper or more expensive on average than the other communities? Explain briefly.

My answer: The median gas price is noticeably lower in Abbotsford than everywhere else. The middles of the boxes are the medians, which will serve as an average. Expect to lose

something if you call them means, which they are not.

(e) (2 marks) Why is it that the Langley boxplot doesn't have a box? Explain briefly.

My answer: It looks as if the median and both quartiles are all the same. Go back and look at Figure 2 to confirm this: in Langley, six of the nine gas prices were 101.9. This is enough to make the median and the first and third quartiles all the same, and the other three values appear to be outliers. Hence, a non-existent box because its top and bottom and middle are all the same.

There is something, probably 1, for noting that a lot of the gas prices in Langley are the same, but to get the second point you have to get as far as seeing that the *quartiles* are the same as well, so that the box part of the boxplot is non-existent. (If you say that a lot of the values are the same and the rest are outliers, you might get more than 1, depending on how well you state your case.)

A boxplot uses the *median* and quartiles, *not* the mean. This is because the mean can be greatly affected by outliers (as it will be here), but the median will not be. Talking about the mean in connection with a boxplot is an error (probably costing you 0.5 here).

(f) (3 marks) Give R code to compute the mean and standard deviation of gas prices, along with the number of stations, for each community.

My answer: This is `group_by` and `summarize`:

```
gas %>% group_by(community) %>%
    summarize(n=n(), xbar=mean(price), s=sd(price))
```

```
## # A tibble: 3 x 4
##   community      n  xbar     s
##   <chr>      <int> <dbl> <dbl>
## 1 Abbotsford     9  91.5  3.27
## 2 Langley        9 101.   4.02
## 3 Surrey         9 104.   1.66
```

Give the summaries any names you like.

Minus one per error, down to a minimum of one if you have something substantial correct. I might deduct only 0.5 if you make what I consider to be a "small error". This kind of work requires attention to detail: it's easy to make two errors and finish up with 1, if you are not careful enough. I stopped counting after 2 (1-point) errors; if you had something remaining that was a decent part of the way to a working answer (like the code for the mean and SD), one point; if not, zero.

If you don't remember `n()`, do something like this in two stages:

```
gas %>% group_by(community) %>%
    summarize(xbar=mean(price), s=sd(price))
```

```
## # A tibble: 3 x 3
##   community   xbar     s
##   <chr>      <dbl> <dbl>
## 1 Abbotsford  91.5  3.27
## 2 Langley    101.   4.02
## 3 Surrey     104.   1.66
```

```
gas %>% count(community)
```

```
## # A tibble: 3 x 2
##   community      n
##   <chr>      <int>
## 1 Abbotsford     9
## 2 Langley        9
## 3 Surrey         9
```

Not very elegant, but it gets the job done, which is better than not doing it. I could have given this 2.5 (for the lack of elegance), but it gets 3 because it works.

This doesn't work:

```
gas %>% group_by(community) %>%
    summarize(xbar=mean(price), s=sd(price), n=count(station))
```

```
## Error in UseMethod("summarise_"):  no applicable method for 'summarise_' applied
to an object of class "character"
```

since you can't use `count` like that. But I thought trying to count in this way was better than not trying to count at all, so I deducted only 0.5 for this.

Also, `n()` has to be empty. If you try to put something in it, this happens:

```
gas %>% group_by(community) %>%
    summarize(xbar=mean(price), s=sd(price), n=n(station))
```

```
## Error in n(station):  unused argument (station)
```

Minus 0.5, on the same basis as before (it's better to try to count rows than to not count them at all).

Question 2 (9 marks)

Is it really true that adult males in North America have a mean weight of more than 160 pounds? To assess this, a random sample of 16 males was taken, with the weights shown in Figure 4. The data frame is called `weights`, and the column of weights in it is called `weight`.

(a) (2 marks) A normal quantile plot is shown in Figure 5. What do you conclude from this? Explain briefly.

My answer: I would say that the points are acceptably close to the line, and therefore the weights have approximately a normal distribution.

If you want to argue for a curve and therefore (left-) skewness, I think you would need the highest value to be only about 180 rather than 200, and you would need the two lowest values to be lower than 120 rather than above it. So I don't think left-skewness is the best answer, but I would give it 1. Saying "skewed" without a direction is only 0.5, and even that only if you provide a reason.

Talking about skewness, you can (if you must) remember which direction of skewness corresponds to which kind of curve (upward- or downward-opening), but then there is software that switches the sample and the theoretical values from what you see here and *you will get the answer wrong* unless you are paying attention. It's much easier to look at the "sample" axis: it's a short tail if the values are bunched up on that axis, and a long tail if they're spread out. One short tail and one long tail means skewed in the direction of the long tail. (This also enables you to see short-tailed distributions, compared to the normal, and long-tailed ones.)

Question 2 continues...

In your answer, make sure you are clear that this is a normal quantile plot and not some kind of scatter plot (which you cannot have here because there is only one quantitative variable). Talking about "upward trends" on a plot like this is dangerous because you are now in an uphill battle to convince the grader that you really know what this plot even is.

(b) (3 marks) Give R code to run a suitable t-test.

My answer: Before coding, think: we are trying to prove that the mean weight is *greater* than 160, so we need that as our one-sided alternative hypothesis. This is a *one*-sample t, which *does not* take a `data=`, so you need to specify the data frame somehow. Hence, this:

```
with(weights, t.test(weight, mu=160, alternative="greater"))
```

```
##
##  One Sample t-test
##
## data:  weight
## t = 0.13498, df = 15, p-value = 0.4472
## alternative hypothesis: true mean is greater than 160
## 95 percent confidence interval:
##  152.5077      Inf
## sample estimates:
## mean of x
##   160.625
```

or this:

```
t.test(weights$weight, mu=160, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  weights$weight
## t = 0.13498, df = 15, p-value = 0.4472
## alternative hypothesis: true mean is greater than 160
## 95 percent confidence interval:
##  152.5077      Inf
## sample estimates:
## mean of x
##   160.625
```

You need the alternative, the null mean, and something saying what data frame the column `weight` is coming from. Expect to lose one per error.

The question said "t-test", so don't expect anything for giving code for a sign test, even if you thought that's what should be run.

(c) (2 marks) The output from your t-test is shown in Figure 6. What do you conclude, in the context of the data?

My answer: The P-value 0.45 is not smaller than 0.05, so we do not reject the null hypothesis (or "we retain the null hypothesis"). Therefore there is no evidence that the mean weight of (all) adult males in North America is larger than 160 pounds (best), or we conclude that the mean weight is (or, better, does not exceed) 160 pounds.

"Accept the null hypothesis" is *wrong*. Expect to lose points for saying that.

Grading: one point for making an appropriate conclusion about the null hypothesis, and one point for saying something appropriate about weights of adult males.

Extra: there is a clue in the output about the coding, since it says what the alternative hypothesis is, so you need to make sure you include something in your coding to make that come out.

Extra 2: If I ask you to do a test, you need to give code that will obtain a P-value appropriate for what you are testing. Don't try to do it with a confidence interval. For example, in Figure 6, the output includes a (one-sided) 95% confidence interval that includes 160, but all that tells you is that the P-value is greater than 0.05. If somebody was reading that and wanted to do their test at $\alpha = 0.10$, knowing the confidence interval wouldn't help them, but giving a P-value of 0.45 would. Worse still is to obtain the default two-sided confidence interval:

```
t.test(weights$weight)
```

```
##
##  One Sample t-test
##
## data:  weights$weight
## t = 34.689, df = 15, p-value = 9.693e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  150.7555 170.4945
## sample estimates:
## mean of x
##   160.625
```

This tells you only that a *two-sided* test would have a P-value greater than 0.05. To make this into a one-sided P-value, you would have to check that you were on the correct side (you are: why?) and then halve it: the one-sided P-value is greater than 0.025, and that tells you nothing about what to do at $\alpha = 0.05$.

(d) (2 marks) Figure 7 shows a summary of the data frame. Someone says to you "but the mean is greater than 160: you should definitely reject the null hypothesis." How do you respond to them? Explain briefly.

My answer: The null hypothesis is a statement about the *population*; the `summary` here is a statement about the data in the *sample*. The sample mean is indeed greater than 160, but it could have been greater than 160 by chance. Indeed, the P-value here says that this kind of sample mean is *exactly* the kind of thing that could have happened by chance, even if the population mean were exactly 160. If you wanted to reject the null hypothesis in favour of an

alternative of "greater", you would need a sample mean that was considerably bigger than this, given the amount of variability present and the smallish sample size.

One point for distinguishing sample and population, one point for proposing chance as an explanation of what happened, with population mean actually being 160.

Question 3 (8 marks)

In each of the scenarios below, state whether we have one sample or two independent samples (or some other kind of sampling), and whether we should use a one-sided or two-sided procedure. In each case, justify your choices briefly. (One point for an answer about samples or sidedness *with a good reason*. No credit if you have no reason.)

(a) (2 marks) Many students have complained that the soft-drink vending machine in the student recreation room dispenses a smaller amount of drink than the vending machine in the faculty lounge. A student randomly samples servings from each machine and records the size of each serving in millilitres.

My answer: Two independent samples (one from each vending machine). There is no way to associate a measurement from one vending machine with one from the other.

If you think this is one sample of serving sizes, think about how the data would be laid out: one column of serving sizes, and a second column showing which machine the serving size came from. The aim here is to compare the serving sizes from the two machines with *each other*, not with some external standard, which is exactly what we are doing with a two-sample test. (Such a data format would be what you would get from `pivot_longer`, if you had started out with the serving sizes for each machine in its own column.)

This is not matched pairs. For that, you'd need something like 10 students to collect the samples, and each student would obtain a sample drink from each machine. Then you'd need a convincing reason why a *student* would have an impact on the serving size (or on the measurement of the serving size). I think this is too hard to justify.

The student wants to prove that the serving size is *smaller* from the machine in the recreation room, so a one-sided test is called for.

On all of these, if you do not say *why* you made your choice, you get zero points. I am not interested in guesses; I am interested in decisions made for a good reason that you can articulate.

(b) (2 marks) An automotive company wants to compare the wearing quality of two brands A and B of tire. To do this, six test cars are used and one tire of each brand is placed on one randomly chosen wheel. (A standard brand of tires is placed on the other wheels.) After the test, each tire of brands A and B is assessed for wear (in thousandths of an inch).

My answer: This is two samples (two brands of tire), but not independent samples because they are tested on the same car: that is to say, each car produces two measurements and we have matched pairs.

You can call it one-sample if you want, but you need to explain where the one sample is coming from, for example the difference in tire wear A minus B for each car (or B minus A, as long as you go this way for each car).

There is no mention of a particular brand that is expected to wear better, so a two-sided test is appropriate. The contrast between this one and the previous one is that we knew without looking at any data which soft drink machine might produce smaller servings (the one in the student lounge), but here we have no idea which tire brand might be better ahead of time, and we want our test to find a difference either way, if it exists. Hence, two-sided. (Another way to say this is that a one-sided test is good if we want to show that a *particular* or *pre-specified* group is better, like the faculty lounge drink sizes in the previous part; otherwise a two-sided test is needed.)

(c) (2 marks) The Robertson square-drive screw has several advantages over a slotted or Phillips-head screw. A catalog reported that Robertson #8 wood screws fail only after an average of 48 inch-pounds of torque is applied (a much larger torque than for other types of screw). An independent testing laboratory randomly samples 22 Robertson #8 wood screws and records the inch-pounds of torque at which each of them fails, to see whether the catalog is correct or whether the average torque is less than 48 inch-pounds.

My answer: There is only one sample (of Robertson screws). (There is an implied comparison with other types of screw, but the testing laboratory did not sample those.)

The lab wants to see whether the average torque is less than 48, so a one-sided test is called for.

(d) (2 marks) The pulse rates of 13 randomly-chosen women were recorded, and a 95% confidence interval for the mean pulse rate of all women was calculated using `t.test`.

My answer: One sample (pulse rates of women). Two-sided because our confidence intervals in this course are two-sided. You might think that the sidedness *of a test* doesn't matter since we are not doing one, but in `t.test`, which we are using here, doing a one-sided test would also get you a one-sided confidence interval, going all the way up or down to infinity, which is not what we want here. So, to get the right *interval*, you also have to ask `t.test` to do the right kind of *test*.

Question 4 (7 marks)

Encyclopedia Britannica defines latent heat as "energy absorbed or released by a substance during a change in its physical state (phase) that occurs without changing its temperature". For example, melting ice turns it from a solid from a liquid, and requires more heat than would simply heating ice without melting it.

In an experiment, two different methods were used to study the latent heat of ice fusion (melting). Water was cooled to -0.72 degrees Celsius (so that it froze). The water specimens were then heated back up to 0 degrees Celsius, and the heat required to do so was measured by one of two methods, an electrical method or a method of mixtures. The method used for a particular specimen was chosen at random. The required heat is called `heat_change` in the data. The data, in data frame `fusion`, is shown in Figure 8.

(a) (4 marks) A boxplot of the data is shown in Figure 9. Three different analyses are shown in Figures 10, 11, and 12. Based on this information, carry out what you think is the most reasonable analysis, justifying your decision, and obtain a conclusion in the context of the data.

My answer: Look at the boxplot first, and decide whether you think the measurements for both methods are normally distributed. If you think they are (which I have to say is unlikely), then go ahead and assess the spreads for equality. If you think they are not, use analysis 1; if you think they are, use analysis 2. If you don't think both groups are normal, you don't need to assess spread; just go ahead and use Mood's median test, analysis 3.

One point for a sensible assessment of the boxplot; one point for a choice of analysis based on your assessment. I don't really think you can justify both samples being normal enough (so

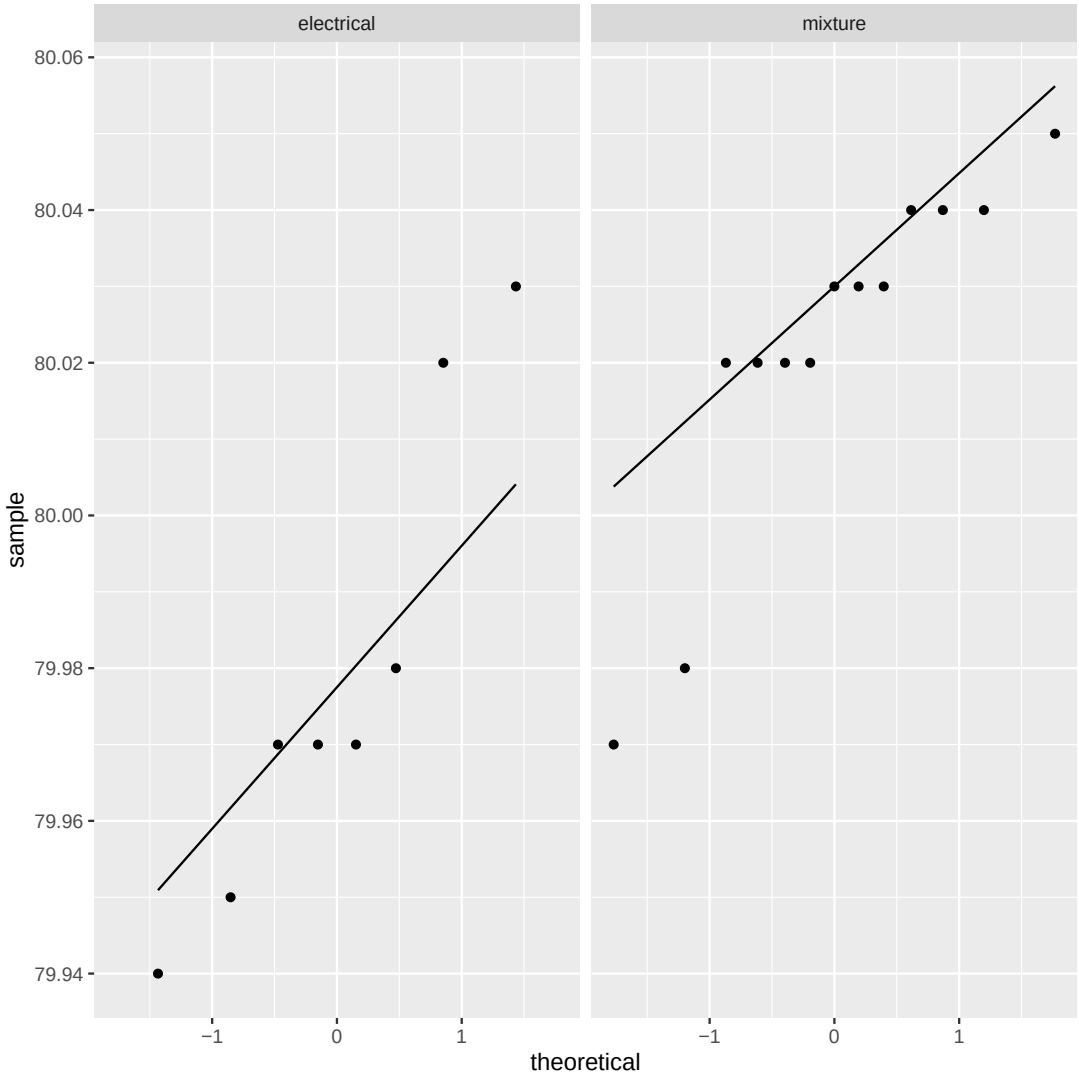
I think "not normal" is the only reasonable conclusion for the first point), but if you follow through properly, you can get the second point by choosing the analysis consistent with your conclusion from the boxplot.

The P-value for whichever test you choose is less than 0.05, so you can reject the null hypothesis that the two medians (means) are the same, in favour of the alternative that they are different. One more point for getting this far. (All three tests are two-sided, since that is the default.) Thus the two methods have *different* median (mean) heat change. One more point.

(b) (3 marks) The people who gave you the data say to you "the boxplot is nice, but what we would really like to see is a suitable normal quantile plot". Give R code to produce a normal quantile plot that is suitable for this analysis.

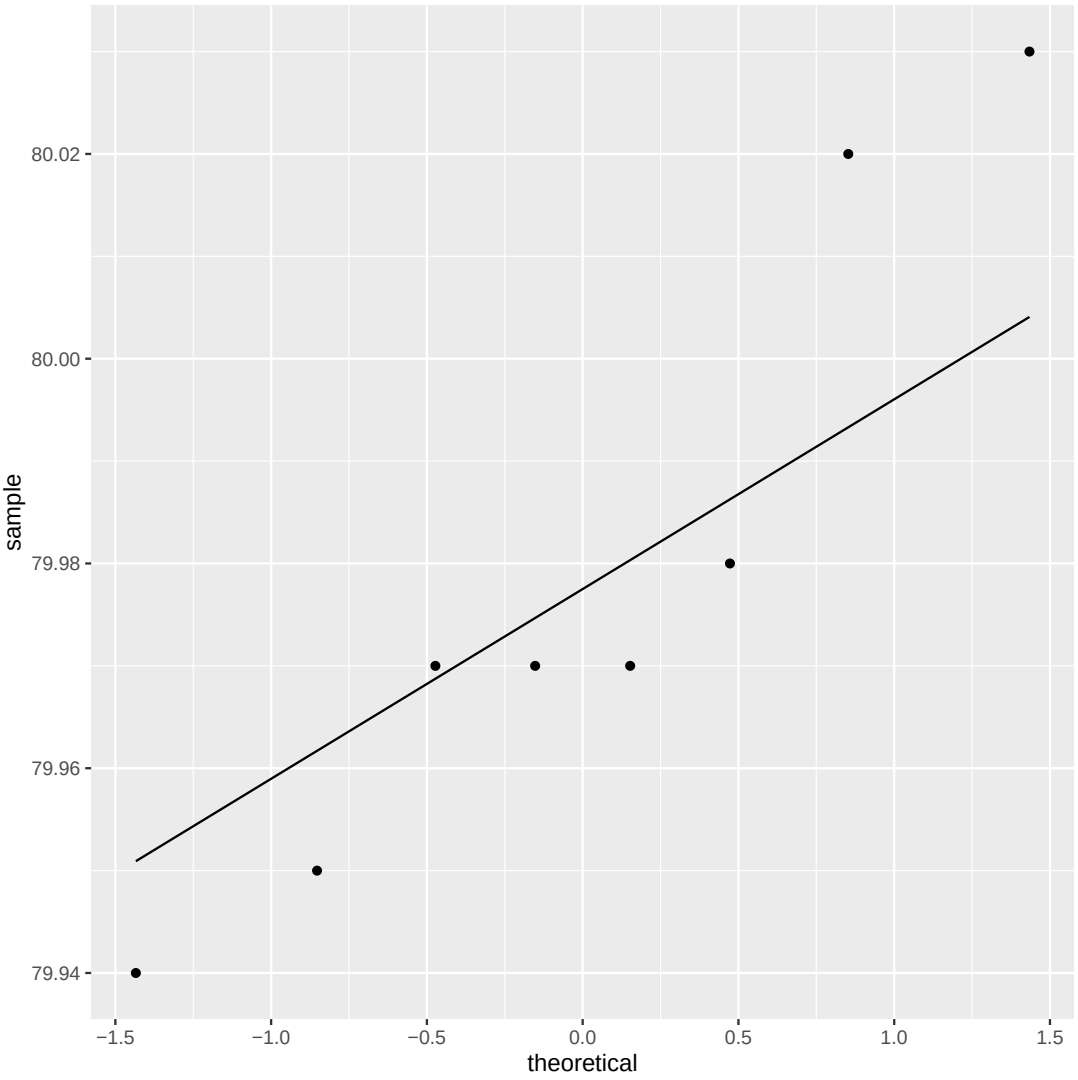
My answer: What needs to be normal is each group separately, so we need a separate normal quantile plot for the two methods. The easiest way to do this is to facet on `method`, thus:

```
ggplot(fusion, aes(sample=heat_change)) + stat_qq() + stat_qq_line() +
  facet_wrap(~method)
```



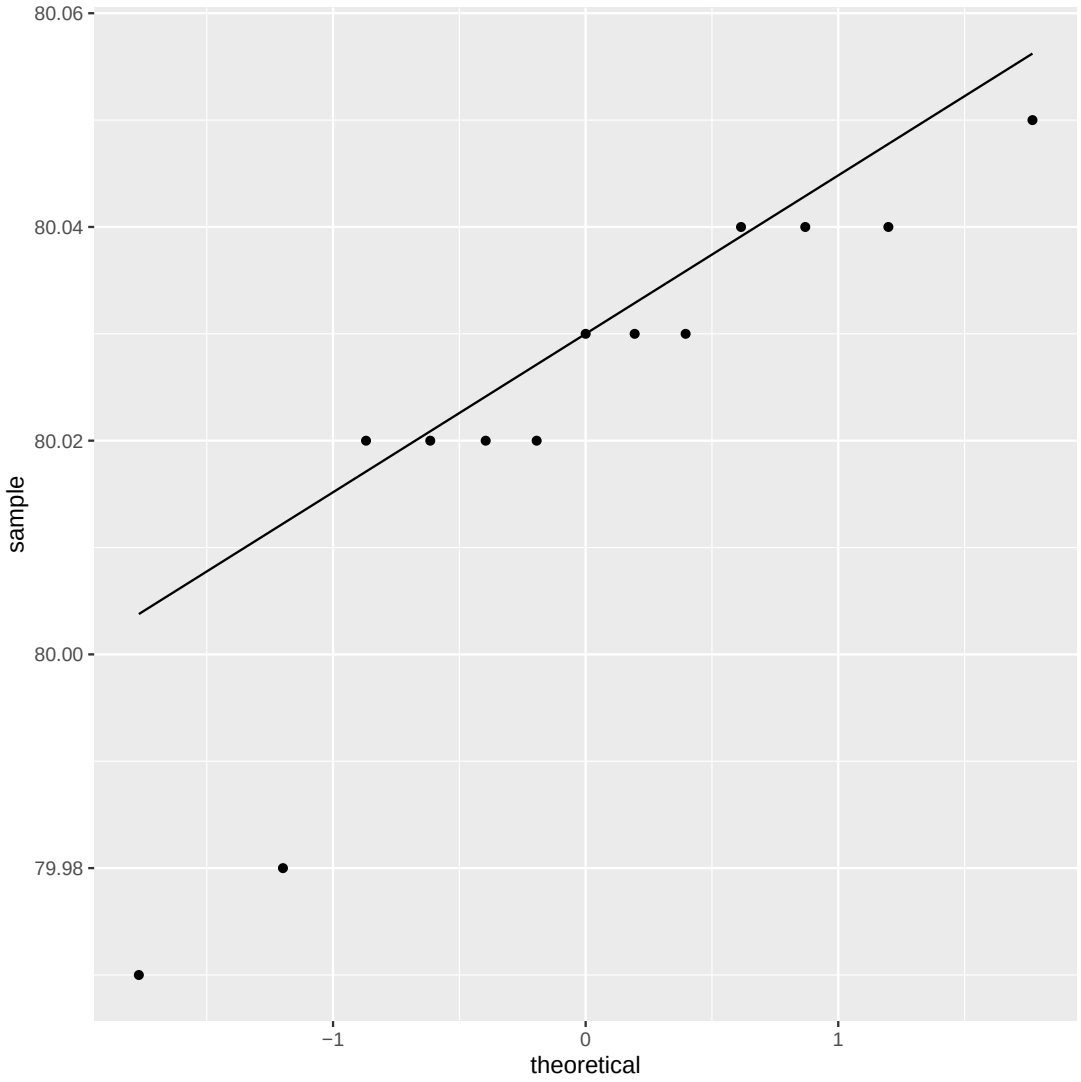
The other way is to make the plots one by one by pulling out the two methods separately:

```
fusion %>% filter(method=="electrical") %>%
    ggplot(aes(sample=heat_change)) + stat_qq() + stat_qq_line()
```



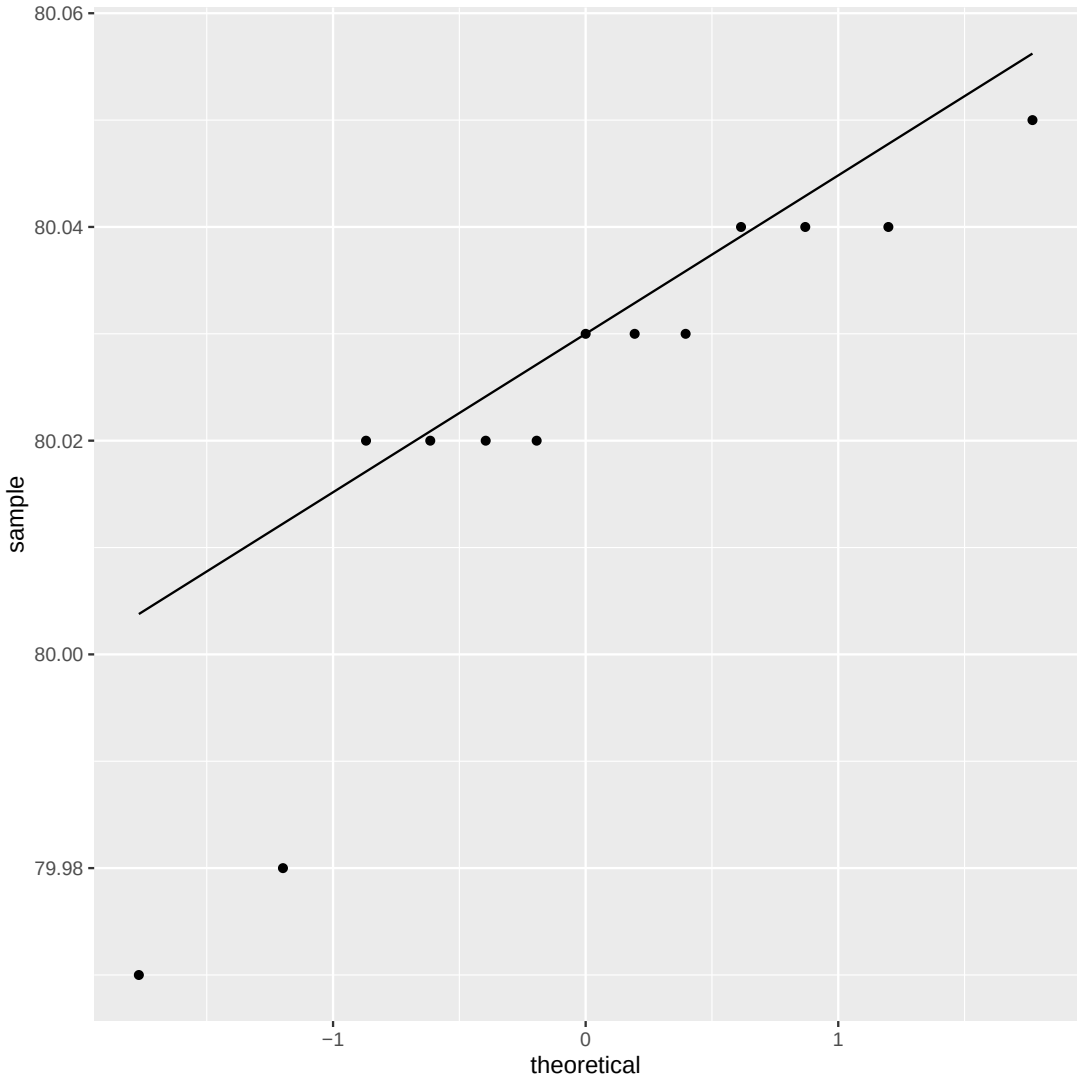
and

```
fusion %>% filter(method=="mixture") %>%
    ggplot(aes(sample=heat_change)) + stat_qq() + stat_qq_line()
```



or by doing something equivalent in base R (though to my mind you really ought to get out of the habit of this), eg:

```
fusion1 <- fusion[fusion$method=="mixture",]
ggplot(fusion1, aes(sample=heat_change)) + stat_qq() + stat_qq_line()
```



and then the other one.

Full marks if you get this. The first way, if you miss out the `facet_wrap` you are kind of missing the point, so only 1 out of 3 for that. Otherwise one off per error, down to 1 if you have something substantial correct in the grader's judgement.

Extra: as I see it, neither of these plots looks convincingly normal. On the `mixtures` plot, the two low outliers really show up, but also on the `electrical` plot, the points are not really very close to the line. The two apparent outliers at the top end are the outlier that shows up on the boxplot, plus the point at the top end of the upper whisker.

Question 5 (12 marks)

The exponential distribution has density function $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (0 otherwise). For example, the exponential density function for $\lambda = 0.5$ is shown in Figure 13. It has the same shape for any value of λ. The distribution has mean $1/\lambda$ and median $\ln(2)/\lambda$, with $\ln(2) = 0.6931$ approximately. The R function `rexp` will draw a random sample from an exponential distribution; it has two inputs, the sample size and the value of λ, in that order.

(a) (2 marks) If we believe that our data comes from an exponential distribution, explain briefly why we would prefer to do a sign test for the median rather than a t-test for the mean.

My answer: The exponential distribution is skewed to the right, and not even close to being normal in shape. Therefore we should favour a sign test for the median over a t-test for the mean.

(b) (2 marks) How might you have guessed at the shape of the exponential distribution (in general) even if I had not shown you Figure 13? Explain briefly.

My answer: No matter what λ is, the median is less than the mean, since $\ln(2) < 1$. This would (correctly) suggest right-skewness.

Or, note that the density function, as a function of x, is exponential decay, and so decreases monotonically for all x. Or show that its derivative is negative for all x. Statistically, this means that the distribution has no left tail at all and a long right tail, so it must be skewed to the right. This is rather a lot of work, compared to the first paragraph of my solution, but if this is what you see and you can explain it properly, it's good.

(c) (2 marks) What value of λ would produce an exponential distribution with a median of 10? Show your calculation.

My answer: Let's call the median m; then $m = \ln(2)/\lambda$, so $\lambda = \ln(2)/m$. Hence, for a median of 10, we must have $\lambda = \ln(2)/10 = 0.0693$.

(d) (4 marks) The function `pval_sign0` in `smmr` takes as input a null median and a vector, and returns the two-sided P-value for the sign test of that null median. Figure 14 shows how it works. Give code that uses `pval_sign0` to estimate the power of the sign test to reject a median of 5 when the median is actually 10, against a two-sided alternative, for data from an exponential distribution and a sample size of 50. You may assume that `smmr` has been loaded with `library(smmr)`. For full marks, do this without a loop.

My answer: First, save the value of λ that goes with a median of 10. `log` does natural (base-e) logs by default. There actually is no `ln` in R, though I won't penalize you if that's what you wrote:

```
lambda=log(2)/10
lambda
```

```
## [1] 0.06931472
```

Then: generate a whole bunch of samples of size 50 from an exponential distribution with this λ, for each one run the sign test and obtain a two-sided P-value, and then tabulate the results. This is my preferred way of doing it, the idea you have seen in lecture:

```
rerun(1000, rexp(50, lambda)) %>%
  map_dbl(~pval_sign0(5, .)) -> pvals
tibble(pvals) %>% count(pvals<=0.05)
```

```
## # A tibble: 2 x 2
##   `pvals <= 0.05`     n
##   <lgl>           <int>
## 1 FALSE             200
## 2 TRUE              800
```

`map` also works instead of `map_dbl`. I thought the `map_dbl` was needed because `pval_sign0` returns a number rather than some bigger data structure (like the result of a t-test from which we might have to extract the P-value). But you can do it either way. (I think the `tibble(pvals)` will create a data frame out of a vector of values, as from `map_dbl`, or from a list of values, as from `map`.)

It can also be done with a for loop, if you must (but you haven't learned that in this class, so expect a maximum of 3 out of 4 for this:

```
pvals <- numeric(1000)
for (i in 1:1000) {
  sample <- rexp(50, lambda)
  pvals[i] <- pval_sign0(5, sample)
}
table(pvals<=0.05)
```

```
##
## FALSE  TRUE
##   164   836
```

There is more to go wrong here. You have to initialize a place to hold the P-values, and you have to remember to store the P-value you get in the right place each time.

Either way, a point off per error, with a minimum of 1 if something substantial is correct.

All of the problems in this course can be solved with methods learned in this course, so that's what I want to see.

(e) (2 marks) I got an answer of about 0.8 for my estimated power. Describe, to someone who doesn't know about power, what your result means.

My answer: If my population has an exponential shape and I draw a sample of 50 values from it with a median of 10, I estimate that I will be able to correctly reject a median of 5 about 80% of the time.

Or, if I have a sample of 50 values from an exponential distribution, I would expect to correctly reject a (population) median of 5 about 80% of the time if the median is actually 10.

Extra: if the someone comes back to you and says that this isn't very good, since the median is nowhere near 5 and the sample is large, you tell them that the exponential distribution is very variable, and if they want better results they will have to take a larger sample.

Question 6 (12 marks)

Are there more arrests made for violations of the narcotics drug laws in larger cities than in smaller communities? A study was made of 24 communities that were classified by size: "large cities" (greater than 250,000 people), "small cities" (under 250,000 people), "suburbs", "rural". For each community, the rates of arrest (for these violations) were recorded per 10,000 inhabitants. The data are shown in Figure 15, in data frame `narc`.

(a) (2 marks) What makes analysis of variance an appropriate method to consider for these data? Explain briefly.

My answer: We have a quantitative variable (arrest rate) that we want to compare over four (in general, more than two) groups (city sizes).

You should name the variables that are of each type, to show that you know what you are looking at.

This is not saying that ANOVA will be the best choice, but merely that it is something we could consider using.

(b) (2 marks) On studying the boxplot in Figure 16, the statistician involved with the study decided that the arrest rate values were sufficiently close to normally distributed, given the small sample sizes (six observations per group). The statistician therefore proceeded with the analysis in Figure 17. What should the statistician conclude from the analysis, in the context of the data?

My answer: The P-value of 0.00038 is very small, so the four sizes of city do not all have the same mean arrest rate (or something equivalent).

You might be tempted to look back at the boxplot and start some discussion like "this is because...". *Don't.* Going any further here is an error and will cost you a point. This is what Tukey is for, in the next part.

(c) (3 marks) Is it useful to study Figure 18? Explain briefly why or why not. If it is useful, summarize as concisely as possible what you conclude, in the context of the data. For this, you might like to think about how the city sizes rank in terms of arrest rates.

My answer: Figure 18 is Tukey's Honestly Significant Differences. This is useful here because the ANOVA said that not all the mean arrest rates are the same, and we want to find out which ones are different; this is exactly what Tukey does. One point.

There are $\binom{4}{2} = 6$ comparisons between the four means, which is a lot to interpret, so you should try to summarize (as hinted in the question). You could say that all differences are significant except for suburb vs. small city. Or, for greater insight, look for which city sizes have a significantly greater or lesser arrest rate than all the others: large cities have a significantly greater arrest rate than all other sizes of city, and rural communities have a significantly *smaller* arrest rate than all other sizes of city. (Suburbs and small cities are kind of stuck in the middle and not distinguishable.)

Two points if you can say something concise about what is happening that gets at which kinds of cities have the highest and lowest arrest rates. One point for less insight, such as simply listing the city size comparisons that are significantly different in terms of arrest rates.

(d) (2 marks) An alternative analysis is shown in Figure 19. Under what kind of circumstances might such an analysis be appropriate? Explain briefly.

My answer: This is a Welch ANOVA followed by Games-Howell. This would be done if we believe the observations within groups are sufficiently close to normal, but we don't believe that they have equal spreads.

Extra: you might argue that this would be *more* appropriate here than the analysis that was actually done, since in the boxplot the suburbs' arrest rates look more variable than the others. See the next part for more on this.

(e) (3 marks) There are two important differences between the results of the analysis of Figures 17–18, and those of the analysis of Figure 19. What are they? What feature of the data do you think might have caused this to happen? Explain briefly.

My answer: "Important differences between the results" suggests to look at the P-values and find ones where the conclusions are opposite. The F-values are both strongly significant, so

it's not that. Compare the P-values in the Tukey and the Games-Howell; in the latter, suburbs are not significantly different (in terms of mean arrest rate) from *anything* else, since the two P-values for comparing suburbs with large cities and with rural communities are now just over 0.05, whereas in the Tukey they were just under. Two points, one for each thing you find.

The last point is an intelligent suggestion about why that is. The Games-Howell comparisons use the spreads (variances) of the groups being compared, and suburbs (from the boxplot) have the largest spread. (Tukey uses the "pooled variance", which is an average spread of all the groups.) In comparing suburbs with something else, Games-Howell is using the large spread of the suburbs' arrest rates, and that will make it more difficult to prove that the mean arrest rate in the suburbs is different from the mean arrest rate somewhere else. This, I think, is the main reason why the suburb differences are no longer significant. The last point.

Another angle you might think of taking is to say that the Welch analysis is actually more appropriate, because the suburbs have a different (larger) spread than the other city sizes. In that case, the reason for the difference is that the regular ANOVA is not the appropriate analysis, and by doing Welch we have gotten P-values we can trust, which we didn't have before. If you get all the way through that, I think that's worth the last point too.

Extra: it can be insightful to compare two competing analyses, as these two here. If the conclusions come out the same, then it doesn't matter which one you do (as we often found with the pooled and Welch t-tests). But if they come out different, then you have some thinking to do. Here, it seems that it depends on whether you think there is really greater variability in arrest rates in the suburbs compared to elsewhere, or whether you think that was just a statistical fluke. With only six observations per group, this is a hard one to settle. Something like Levene's test will not shed any light at all, because the samples are so small:

```
library(car)
```

```
## Loading required package:  carData
```

```
##
## Attaching package:  'car'
```

```
## The following object is masked from 'package:dplyr':
##
##    recode
```

```
## The following object is masked from 'package:purrr':
##
##    some
```

```
leveneTest(arrest_rate~factor(city_size), data=narc)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.3527 0.7877
##       20
```

You'll remember that I don't care for testing variances to decide which test for means to do. This is half of why: when the samples are small, it's almost impossible to prove the spreads different even if they are. (The other half of the why is that when samples are large, even a very tiny difference in the spreads, one that has no impact at all on the comparison of the means, can be significant.)

Perhaps the best argument for equal vs. unequal variances would come from the criminology: perhaps it is true that suburbs are more heterogeneous than the other city types, in that some suburbs have a big drug problem and some do not at all, and so the arrest rate could depend crucially on the kind of suburb you are looking at. I have no idea whether this is true or not, but it would be something to discuss with a criminologist, and, if true, *that* would be a sound reason to favour the Welch analysis over the standard ANOVA.

Use the area below if you need more space to write your answers. Be sure to label any answers here with the question and part that they belong to.