STAC32 Assignment 2

Due Thursday September 19 at 11:59pm

Hand in your answers to Question 2. Question 1 contains suggested problems from PASIAS to work through. They may contain hints for the question to hand in.

The assignment is due on the date shown above. An assignment handed in after the deadline is late, and may or may not be accepted (see course outline). My solutions to the assignment questions will be available when everyone has handed in their assignment.

You are reminded that work handed in with your name on it must be *entirely your own work*.

Assignments are to be handed in on Quercus. See `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html` for instructions on handing in assignments in Quercus. Markers' comments and grades will be available there as well.

You will probably want to begin with this:

```
library(tidyverse)
```

(If you have opened a new project on `rstudio.cloud` for this assignment, you will most likely need `install.packages("tidyverse")` first.)

1. In PASIAS , work through problems 6.1 through 6.4.

   Hand the next question in:

2. How accurate are radon detectors of a type sold to homeowners? In a study, researchers placed 12 detectors of this type in a chamber that exposed them to a certain dose of radon. The data are in `http://www.utsc.utoronto.ca/~butler/assgt_data/radon.txt`. There is only one column of values. These values are in units of picocuries per litre.

   (a) (2 marks) Read in and display the data.

   Solution:

   There is only one column, so you can pretend the file is a CSV or a text file delimited by anything you like. I think treating it as a CSV is easiest:

```
my_url="http://www.utsc.utoronto.ca/~butler/assgt_data/radon.txt"
radon=read_csv(my_url)
```

```
## Parsed with column specification:
## cols(
##  reading = col_double()
## )
```

```
radon
```

```
## # A tibble: 12 x 1
##    reading
##      <dbl>
##  1    91.9
##  2    97.8
##  3   111.
##  4   122.
##  5   105.
##  6    95
##  7   104.
##  8    99.6
##  9    96.6
## 10   119.
## 11   105.
## 12   102.
```

Extra: this also works, and is therefore full marks:

```
radon2=read_delim(my_url, " ")
```

```
## Parsed with column specification:
## cols(
##  reading = col_double()
## )
```

```
radon2
```

```
## # A tibble: 12 x 1
##    reading
##      <dbl>
##  1    91.9
##  2    97.8
##  3   111.
##  4   122.
##  5   105.
##  6    95
##  7   104.
##  8    99.6
##  9    96.6
## 10   119.
## 11   105.
## 12   102.
```

I called the one column `reading` so that you could use `radon` to name the data frame.

(b) (3 marks) Obtain numerical summaries for the radon readings, one to summarize the centre, and the other to summarize spread.

Solution:

This means either mean and standard deviation, or median and IQR. *I don't mind which you choose.* Thus either this, where you get to choose the names for the summaries:

```
radon %>% summarize(xbar=mean(reading), s=sd(reading))
```

```
## # A tibble: 1 x 2
##    xbar     s
##   <dbl> <dbl>
## 1  104.  9.40
```

or this:

```
radon %>% summarize(med=median(reading), iqr=IQR(reading))
```

```
## # A tibble: 1 x 2
##     med   iqr
##   <dbl> <dbl>
## 1  103.   9.4
```

Note that *getting* the interquartile range requires Capital Letters, but you can call the result whatever you like. There is no `group_by` because there is no grouping variable.

The point of this part was to get some numerical summaries, and I don't mind which pair you chose (or even, I guess, something funky like a trimmed mean and the range, if you can figure out how to get those). I would like your measure of spread to be the SD if you chose the mean, or the IQR if you chose the median, since these usually go together.

Extra 1: note that the mean and median are very close (given the amount of variability there is in the data) so that it really doesn't matter which one you use. It turns out (by coincidence) that the SD and IQR are to this accuracy the same; that is not actually meaningful because for normally-distributed data the IQR is about 1.35 times the SD (why? see later in the course).

Extra 2: you can make a choice between the two pairs of summary statistics by looking at your graph from later. The consideration is actually the same as whether you would trust a t-procedure: if you think the data are normal enough, a t procedure is OK, and the mean and SD are OK. If not, a t-procedure is shaky, and you would use median and IQR to summarize the data with. (This is actually not quite accurate, because you have the Central Limit Theorem which means that t-procedures can tolerate a certain amount of non-normality, but as a rough guideline it's good enough.)

(c) (4 marks) Obtain a 90% confidence interval for the mean radon level (of all detectors of this type). Make a brief statement containing your conclusion in the context of the data.

Solution:

This is a non-default confidence level, so you will need to go back to your notes and find out how to change the confidence level. It goes like this:

```
with(radon, t.test(reading, conf.level=0.90))
```

```
##
##  One Sample t-test
##
## data:  reading
## t = 38.386, df = 11, p-value = 4.537e-13
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##   99.26145 109.00521
## sample estimates:
## mean of x
##  104.1333
```

I think this is the cleanest way, but if you are a fan of the dollar sign, that works too (and is also full credit):

```
t.test(radon$reading, conf.level=0.90)
```

```
##
##  One Sample t-test
##
## data:  radon$reading
## t = 38.386, df = 11, p-value = 4.537e-13
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##   99.26145 109.00521
## sample estimates:
## mean of x
##  104.1333
```

Getting the output (in one of these ways) is two marks. For the last two, you need to pick out the confidence interval from the output and say something about what it tells you, for example

> With 90% confidence we say that the mean radon level that would be measured by all detectors of this type is between 99.3 and 109.0 picocuries per litre.

Your answer needs to contain: (i) the confidence level, (ii) what the population mean actually *is* in this case (I can live with something like "the population mean radon level" for this), (iii) the confidence interval itself, (iv) the units, in some order. If you are missing any of this, expect to miss some marks. You should also round off the confidence limits in the output to something that a human can consume: maybe one or two decimal places, but definitely not five. (A useful rule of thumb is to look at how accurately the data values were measured, one decimal place here, and take about one more decimal place in things like sample statistics and confidence intervals. That would justify two decimal places here, but not more. If you have a really large sample, like in the hundreds, you know things a bit more accurately than we do here.)

If you just talk about "the population mean", you are not showing that you know what the population mean *is* in this case, and if you just talk about "the mean", this is a more serious problem in that you don't know what a confidence interval is actually *for*. There is a sample mean and a population mean, and they are probably different. You know the first, but you care about the second.

As a general point, you as a statistician have two jobs to do: you have to run the analysis (correctly), and you also have to *interpret* the analysis in terms that the person who gave you

the data can understand and act upon. You need to get into the habit of doing all of this when you do an analysis. In this case, the researchers who collected the data filled the chamber with some known concentration of radon, which you don't know (yet), and if the confidence interval you calculate does not include that value, the company that manufactures the detectors may have to go back and re-calibrate them. (They would probably organize another experiment first, since this one was rather small.)

(d) (3 marks) The researchers actually filled the chamber with radon at a concentration of 105 picocuries per litre. Test whether the mean reading (of all detectors of this type, if they had been placed in this chamber) differs from 105. What do you conclude, in the context of the data?

Solution: This is asking for a hypothesis test: that is, you run `t.test` again, but this time specifying a null mean (105) and not a confidence level. We should probably pick an α here before looking at the results; $\alpha = 0.05$ is fine. Or you could reason that since we used a 90% confidence interval before, we should use the corresponding α of 0.10. That is in fact a good idea.

On the output, you ignore the confidence interval that comes out (which will actually be a 95% one) and look at the P-value:

```
with(radon, t.test(reading, mu=105))
```

```
##
##  One Sample t-test
##
## data:  reading
## t = -0.31947, df = 11, p-value = 0.7554
## alternative hypothesis: true mean is not equal to 105
## 95 percent confidence interval:
##   98.1625 110.1042
## sample estimates:
## mean of x
##  104.1333
```

The P-value is 0.7754. Two marks for getting this far. (If you quote the P-value in your explanation without pulling it out here, that's fine too.)

The P-value is clearly *not* less than 0.05 (my α) or 0.10 (maybe yours), so we *do not* reject the null hypothesis, and there is no evidence that the mean reading of all radon detectors of this type (when placed in this chamber) differs from 105 picocuries per litre. The plain-language takeaway from this experiment is that these radon detectors are properly calibrated, and are safe to sell to the public. (When the concentration is actually 105, the mean reading of all of the detectors is inferred to be not different from 105.)

In your conclusion, you need to: (i) quote the P-value, (ii) compare it with your α (or with 0.05), (iii) make a decision about rejecting the null hypothesis, (iv) explain what that decision means for this data set. (I don't need my plain-language takeaway, but it's nice if you include something like it.)

In this course, and in Statistics generally:

- asserting that you reject the null hypothesis (or don't) is only about halfway to a conclusion.

- a null hypothesis is never "accepted". All we are ever doing is trying to prove it *wrong*, and so if we cannot prove it wrong (as happened here) that's what we have to say. For example, the population mean radon reading might be something like 105.2; unless we had a very large sample, we would never be able to prove that the mean was different from 105 even though it was, and in any case, this kind of inaccuracy in the detector might not be important. A null hypothesis is either "rejected" or "not rejected". Or, if you like, "retained" instead of "not rejected", because this has the proper implication that we don't have any evidence to the contrary, rather than believing with all our hearts that the null hypothesis is true.

(e) (2 marks) Why might you have guessed, looking at your confidence interval, that your hypothesis test would come out the way it did? Explain briefly.

Solution: See where 105 is in relation to your confidence interval. It's inside the interval, in fact, almost in the middle of it. From that point of view, 105 is a "plausible" value for the population mean reading, and we would not expect to reject it. In our test, we were nowhere near rejecting a null mean of 105, so this is all consistent.

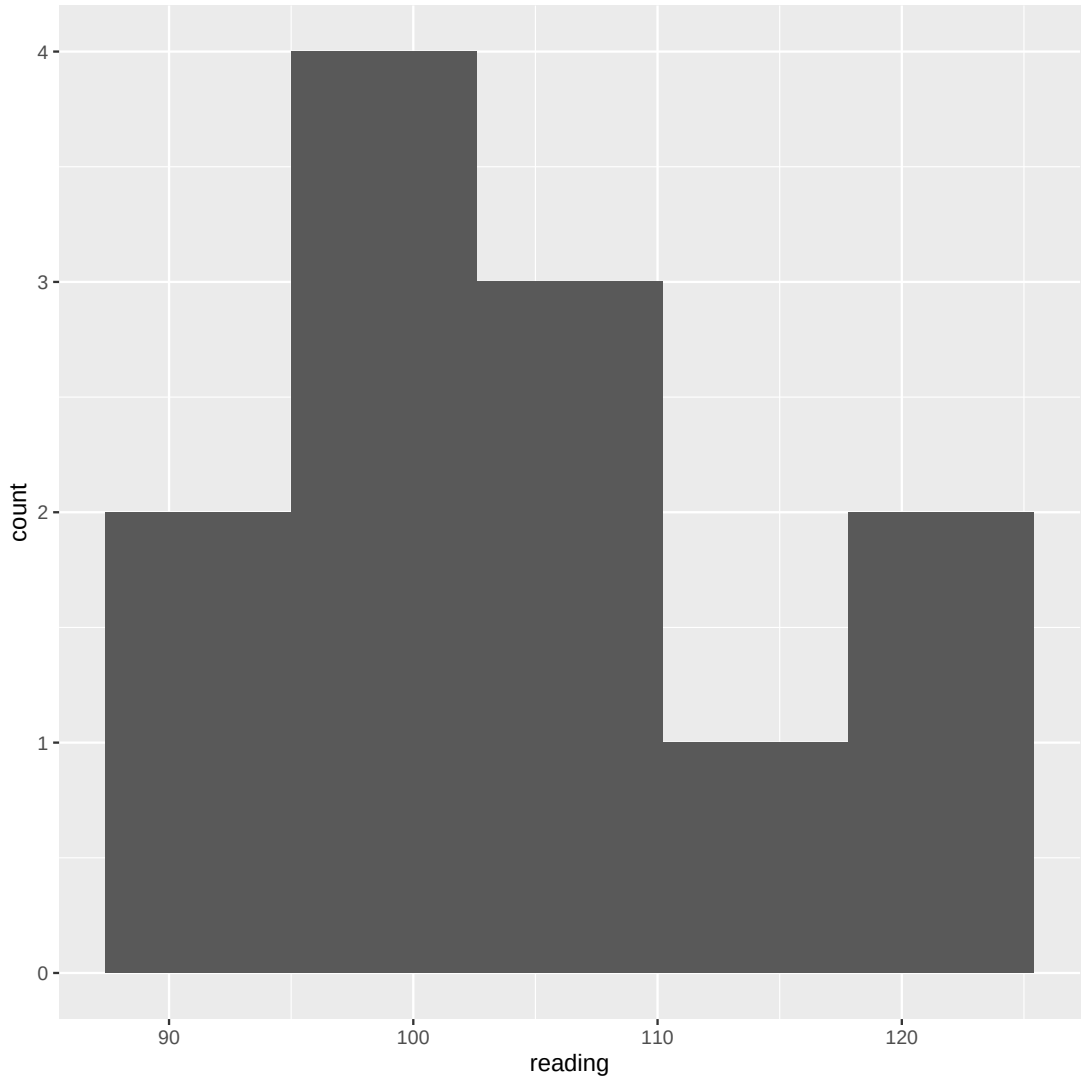
I like the word "plausible" to describe a population mean that is inside a confidence interval or that is not rejected by a test. This helps guide my thinking as to what I would expect to see.

We come back to this issue later (when we look at the sign test, or to be precise just before that). I thought it would be a good idea to have you think about it now, so that it makes sense when we come back to it.

(f) (3 marks) Make a suitable graph to display your data. Does your graph suggest that it is reasonable to use t-procedures here, or not? Explain briefly.

Solution: One quantitative variable suggests a histogram. Don't use too many bins because there are only 12 observations:

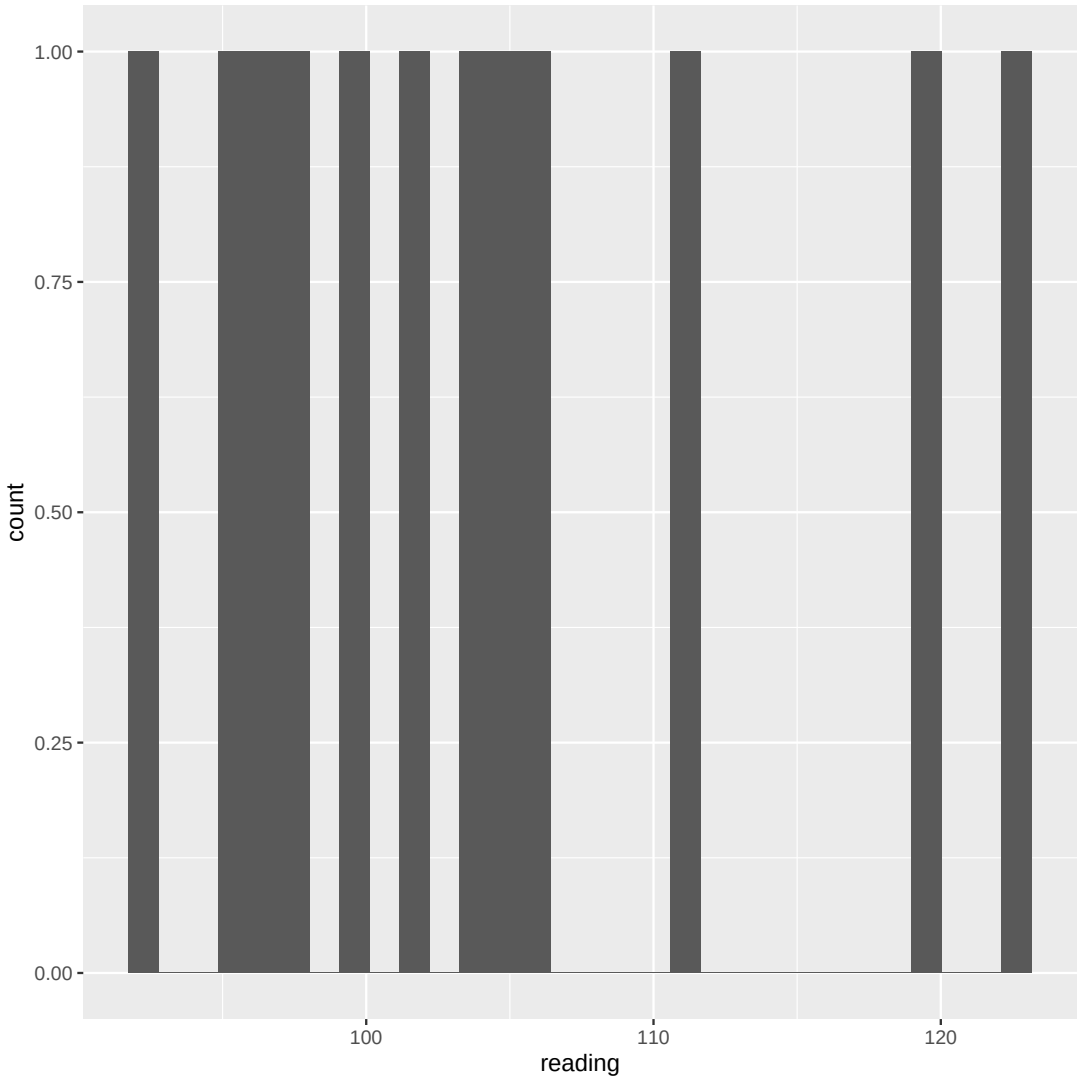
```
ggplot(radon, aes(x=reading))+geom_histogram(bins=5)
```



I used only 5 bins because I wanted to get a sense of the shape of the distribution. I think fewer bins would be too few to show the shape, and more would make it difficult to see what the shape is. For example, the default 30 bins definitely does not show the shape:

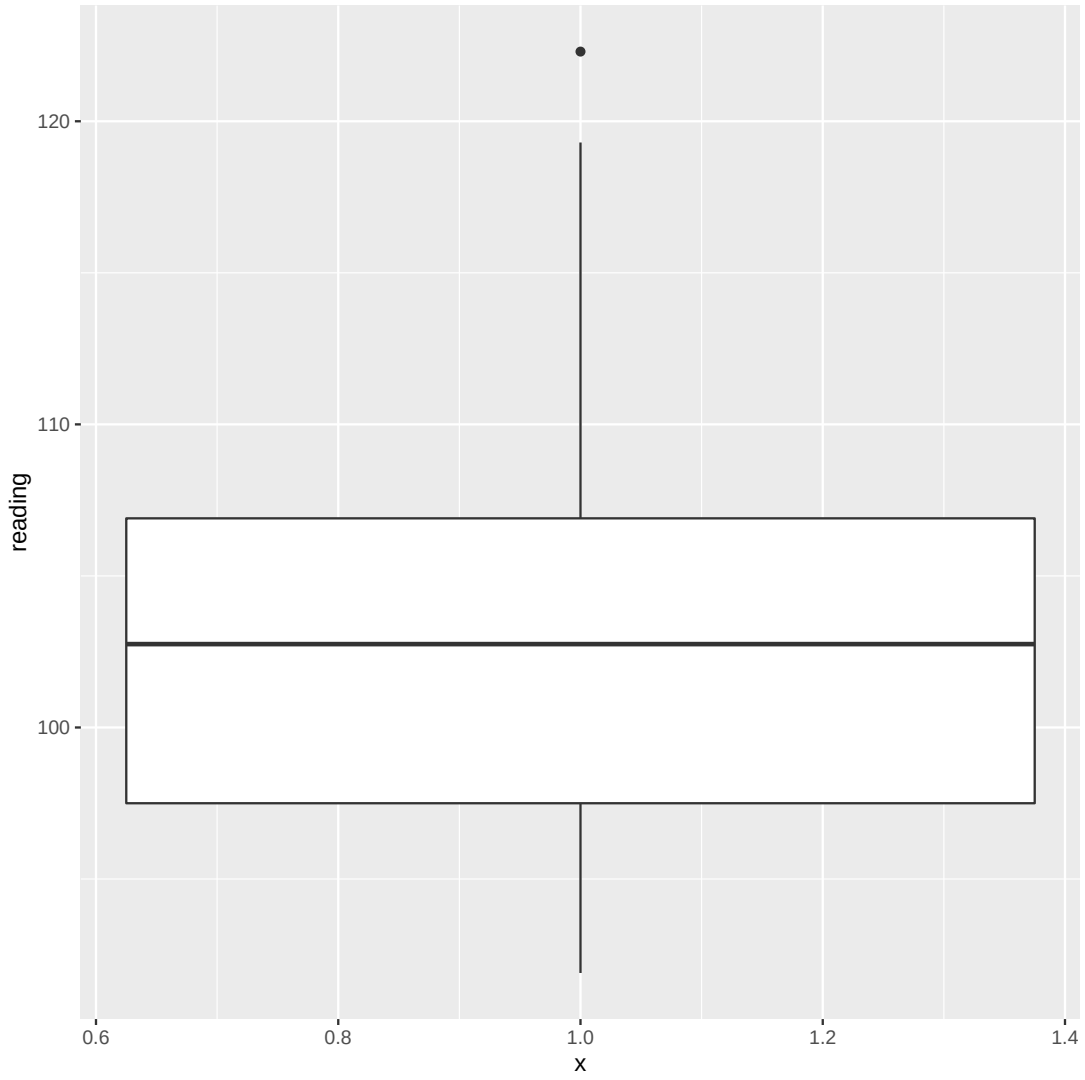
```
ggplot(radon, aes(x=reading))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



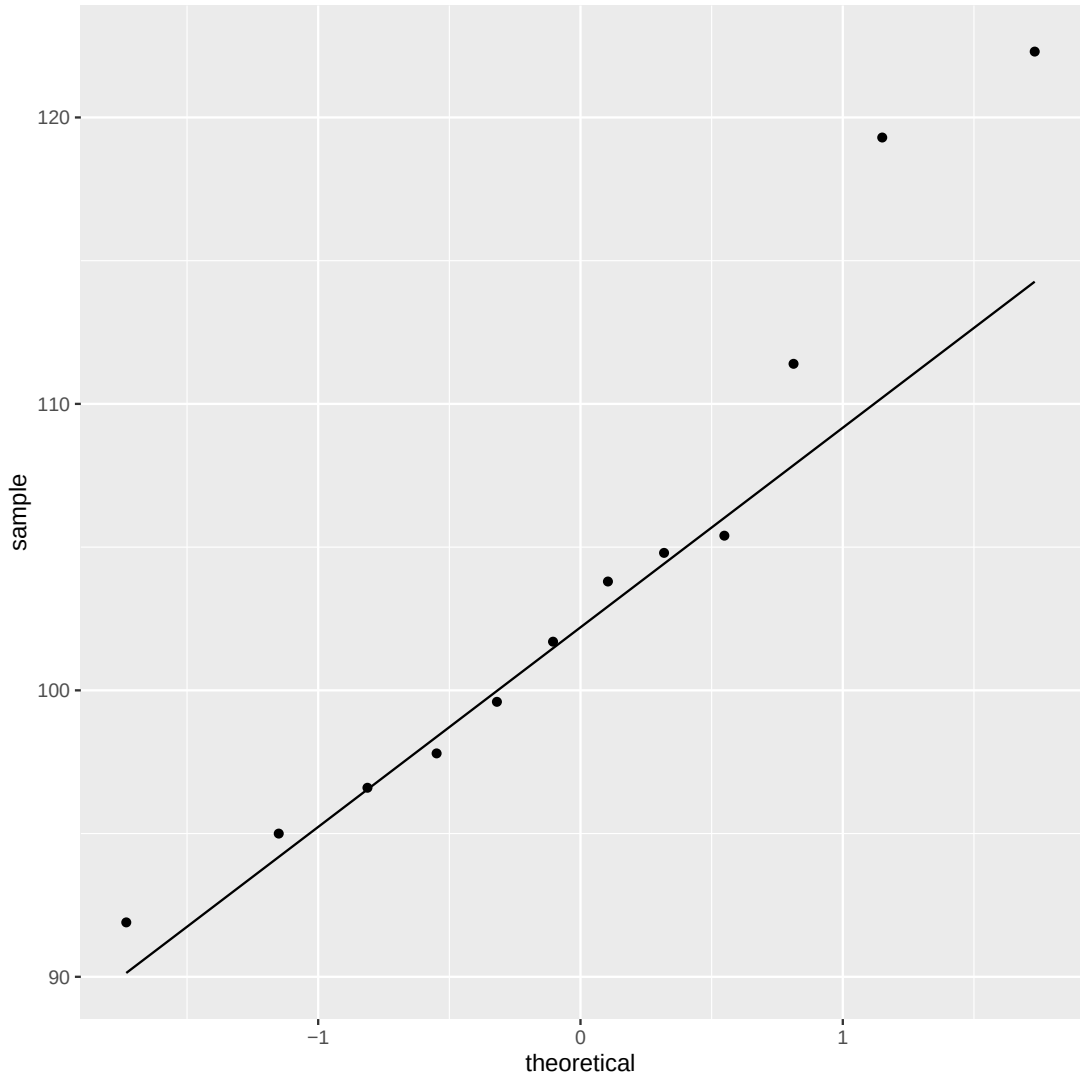
A one-group boxplot is also acceptable:

```
ggplot(radon, aes(x=1, y=reading))+geom_boxplot()
```



Some people knew about a normal quantile plot (which I haven't gotten to yet in class), and that is also a good choice:

```
ggplot(radon, aes(sample=reading)) + stat_qq() + stat_qq_line()
```



I think only one mark for choosing and making a graph, since you've done that before.

The issue here is whether you think this is sufficiently close to a normal distribution, bearing in mind that with $n = 12$ observations, we have a little help from the Central Limit Theorem (not much, but some). So I think you need one of these two approaches:

- The distribution is clearly skewed to the right (or has one or two outliers at the upper end, depending on the graph you chose). This is not close enough to being normal for the mean to be an appropriate measure of centre, and therefore I don't trust the t-procedures that we used.
- The distribution is somewhat skewed to the right (or has at least one mild outlier at the upper end). However, the skewness is not severe, and so I think with 12 observations that the t-test will be acceptable.

I don't mind which way you go; what I care about is the reasoning by which you get there. (This is a common theme in this class; applied statistics is all about having opinions and supporting them with good reasons.) I'm guessing that most people will go the first way, although my

preference is actually for the second alternative (I explain why below), but either approach, properly justified, is full marks.

Something to carry forward is that actual data will never look completely normal (see the lecture notes later on the normal quantile plot) even if the mechanism that generated it really was normal, so that you cannot hope for perfection. The kind of question to ask yourself, therefore, is whether the data is *non-normal enough to cause a problem*. See the blog post at `https://thestatsgeek.com/2013/09/28/the-t-test-and-robustness-to-non-normality/`. This is for a two-sample t (I think) but the issues are the same as here. With small samples, we are therefore in a difficult position: the normality matters as far as the t-test is concerned, so we need to make a call about normality, but even if the data really does come from a normal distribution, its histogram may not look very normal and it is easy to make a "type II error" (if you think of "the population distribution is normal" as a kind of null hypothesis): that is to say, to reject normality when the data really did come from a normal distribution. I don't know whether this is what happened for the radon readings, but it might have done.

Extra: later we learn about the sign test and its associated confidence interval for the median. Here, that goes like this (just read this bit for now, since there are extra issues with installing `smmr` that we get to later):

```
library(smmr)
ci_median(radon, reading, conf.level=0.90)
```

```
## [1]  96.60513 111.39946
```

```
sign_test(radon, reading, med0=105)
```

```
## $above_below
## below above
##     8     4
##
## $p_values
##   alternative   p_value
## 1       lower 0.1938477
## 2       upper 0.9270020
## 3   two-sided 0.3876953
```

The confidence interval is similar to (a bit longer than) the t one, and the P-value for the test is a bit smaller (the one to look at is the two-sided 0.388) but also nowhere close to rejection. So my take is that it doesn't really matter what you do, and thus the t procedures are acceptable enough. (If you can justify using t, then go ahead and use it.) There is no right and wrong here, though; if you see a problem with the t-test and CI, go ahead and use the sign test when we've learned about it. I have no problem with that.