Assignment 4

Due Thursday October 3 at 11:59pm on Quercus

As before, the questions without solutions are an assignment: you need to do these questions yourself and hand them in (instructions below). 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.

As ever, begin with this:

```
library(tidyverse)
```

1. Work through problems 9.1 and 9.2 in PASIAS. If you like, also work through problem 9.3. (This last problem is not immediately relevant to this assignment, but you have all the background to make sense of it, and doing so will give you some context for what is going on.)

   Hand the next one in.

2. Federal and provincial governments often specify a minimum wage level in government-tendered contracts. To ensure the wage levels are reasonable, government officials often survey wage rates for similar contracts elsewhere. An official believes that the median wage rate for part-time construction millwrights on contract is $15 per hour. A sample of workers in this category was selected, and their wage rates recorded. The data are in `http://www.utsc.utoronto.ca/~butler/assgt_data/millwrights.txt`.

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

   Solution:

   You can pretend this is a `.csv` or a delimited value (delimited by whatever you like). I'm going to make it easy and pretend it's a `.csv`:

```
my_url <- "http://www.utsc.utoronto.ca/~butler/assgt_data/millwrights.txt"
millwrights <- read_csv(my_url)
```

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

```
millwrights
```

```
## # A tibble: 12 x 1
##      wage
##     <dbl>
##  1  15.5
##  2  14.5
##  3  17
##  4  13.4
##  5  17.2
##  6  18.5
##  7  13.5
##  8  14.4
##  9  16.8
## 10  17.5
## 11  16.2
## 12  18.4
```

Aside: according to `https://www.careersinconstruction.ca/en/career/construction-millwrightindustri` a "millwright" is an industrial mechanic, whose job is to install, maintain and repair industrial machinery and mechanical equipment, for example on a construction site, in a factory or even at an amusement park or ski hill.[1] The original meaning of "millwright" is "person who builds mills"; it is one of a number of -wright words that survive today mostly as surnames, like "cartwright" or "wainwright", people who built carts or wagons respectively.[2]

(b) (3 marks) Use R to count the number of values above and below \$15. Would you expect a sign test to reject a null median of \$15, against a two-sided alternative, or not? Explain briefly (without doing any more calculation.)

Solution:

This can be as simple as this:

```
millwrights %>% count(wage>15)
```

```
## # A tibble: 2 x 2
##   `wage > 15`     n
##   <lgl>       <int>
## 1 FALSE           4
## 2 TRUE            8
```

Or you can make a new column with whether or not each wage is above 15 (there are no wages exactly equal):

```
millwrights %>%
  mutate(is_above=(wage>15)) %>%
  count(is_above)
```

```
## # A tibble: 2 x 2
##   is_above     n
##   <lgl>    <int>
## 1 FALSE        4
## 2 TRUE         8
```

or something a bit more aesthetically pleasing:

```
millwrights %>%
  mutate(above_below=ifelse(wage>15, "above", "below")) %>%
  count(above_below)
```

```
## # A tibble: 2 x 2
##   above_below     n
##   <chr>       <int>
## 1 above           8
## 2 below           4
```

Or count the number of values below rather than above. You'll get the same answer either way: 8 above and 4 below.

As to whether you'd expect to reject a median of 15 in a sign test: I would say this is pretty close to 50-50 above and below 15, so I would definitely *not* expect to reject a median of 15. Think about tossing a fair coin 12 times; 8 heads and 4 tails would not be terribly surprising (or 4 heads and 8 tails) if the coin is fair. Maybe a split of 10–2 or 11–1 would be unbalanced enough to lead to rejection for this sample size, but I'm guessing a split like this is about what it would take. (With a larger sample size, like $n = 100$, having two-thirds of the data values above the hypothesized median *would* be enough to reject the null, but not with a sample size this small.[3])

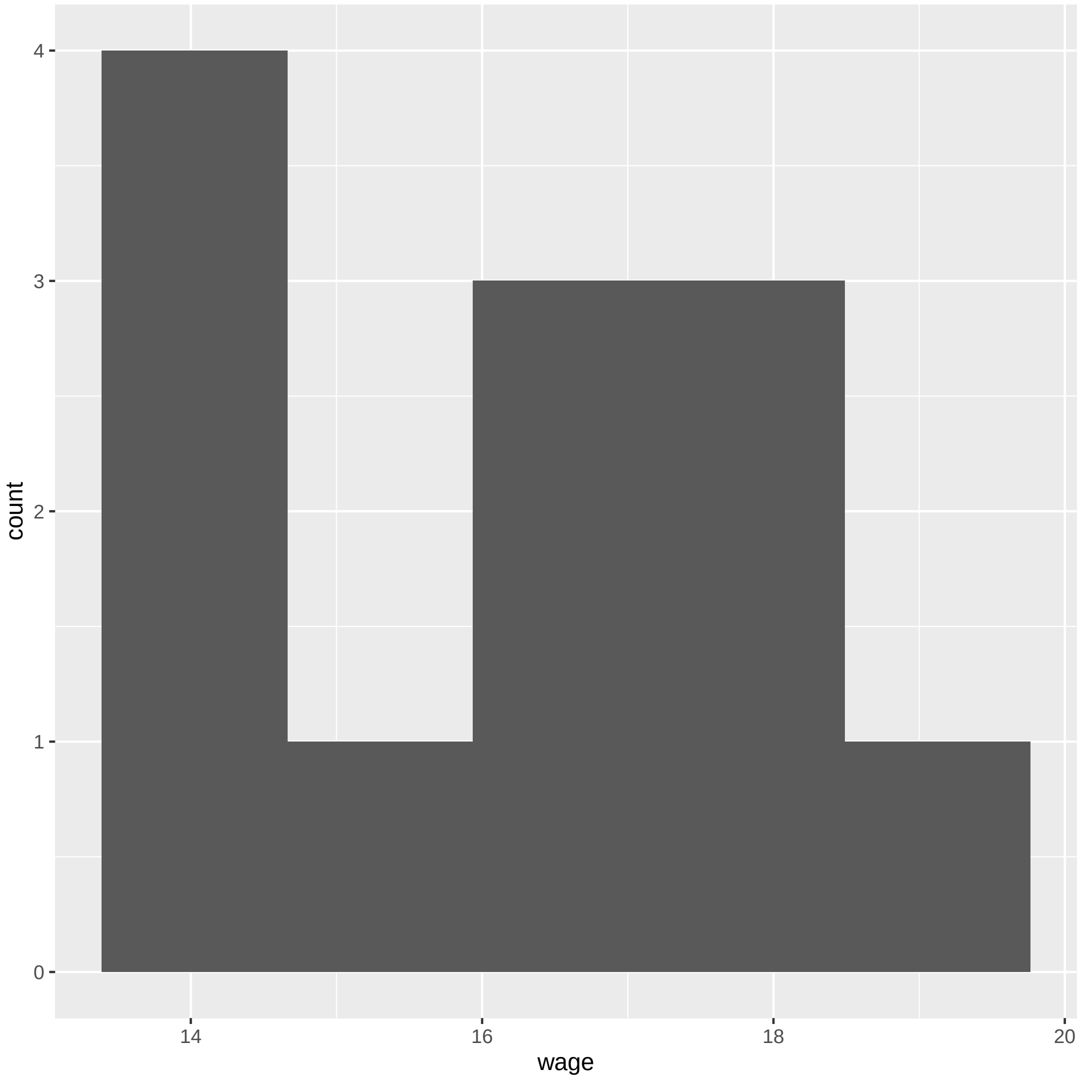
The key thing here is a discussion of whether you think the observed counts are unbalanced enough to indicate that we should be rejecting. If you want to say that an 8–4 split is unbalanced and we should expect to reject the null, you might want to revise your answer later, but the logic here is sound.

(c) (2 marks) Make a suitable graph of the wage values. Comment briefly on its shape.

Solution:

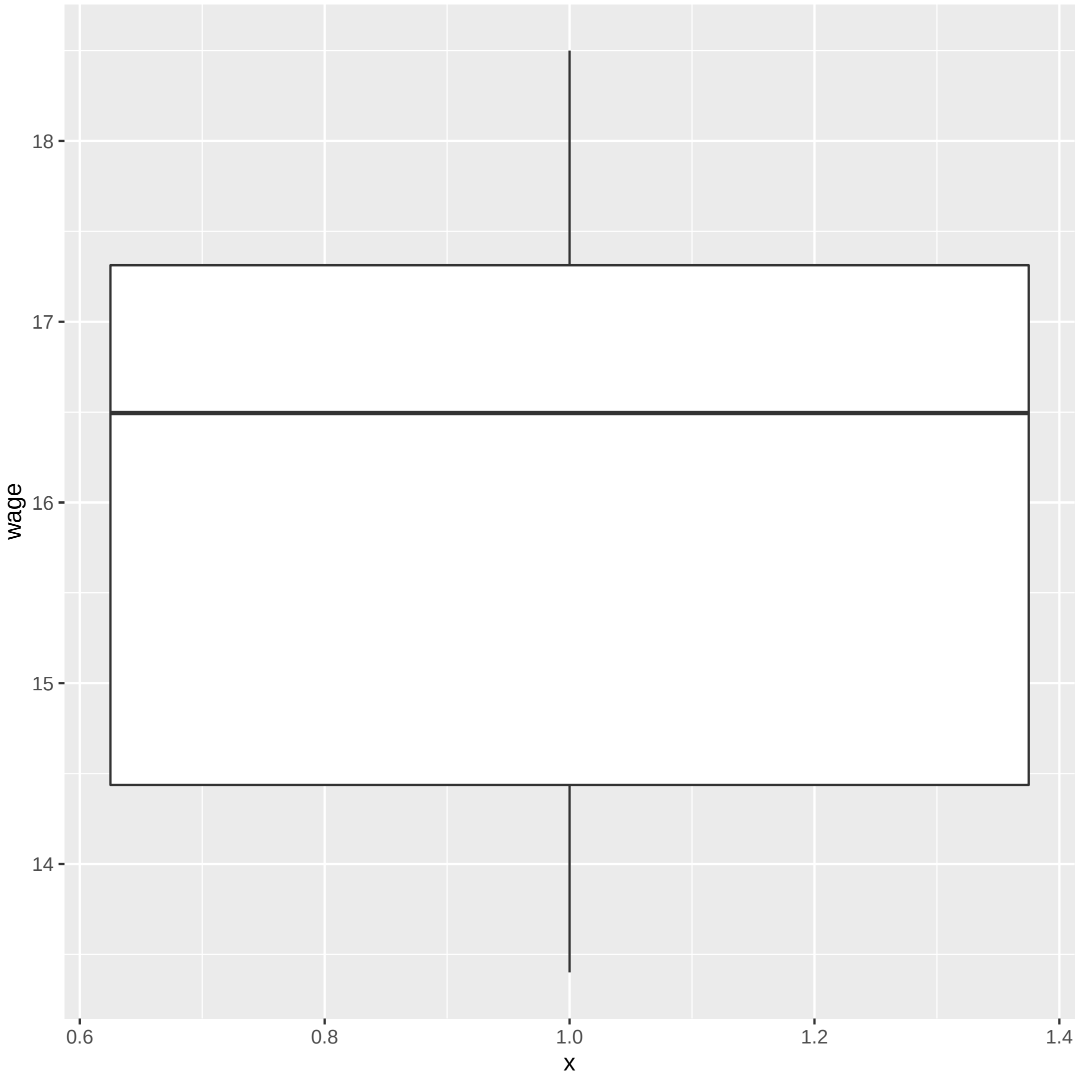
The obvious one is a histogram, with a small number of bins (since there isn't much data):

```
ggplot(millwrights, aes(x=wage)) + geom_histogram(bins=5)
```



or you could justify a one-sample boxplot:

```
ggplot(millwrights, aes(x=1, y=wage)) + geom_boxplot()
```

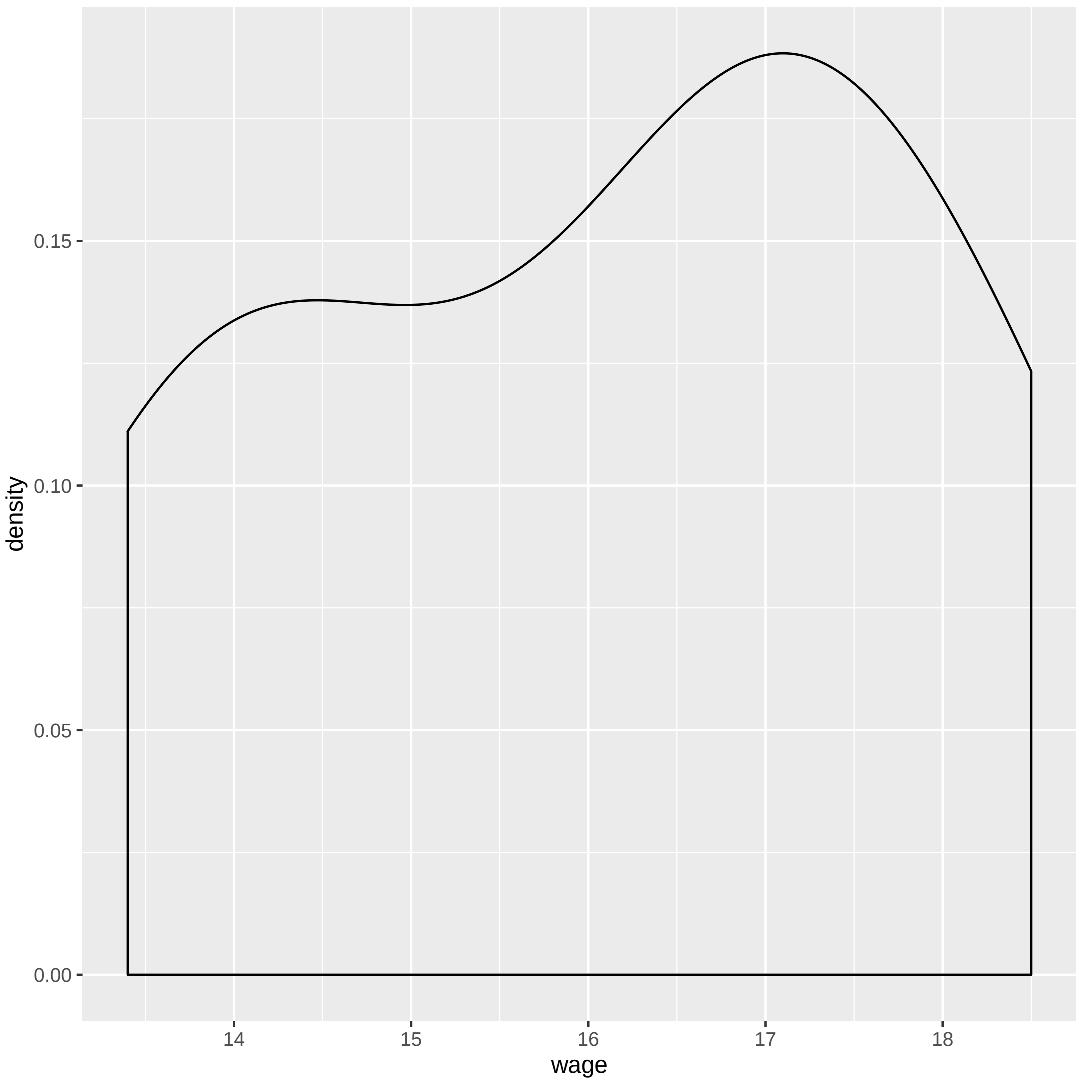


You ought to (very briefly) justify your choice of graph.

The boxplot looks symmetric (I wouldn't worry, with the small sample size, about the asymmetric median in the boxplot). The shape you get from the histogram seems to be very dependent on the number of bins you use. Here, with five bins, you could reasonably describe it as skewed to the right, but when I used six bins, it looked more or less "flat-topped": symmetric but maybe uniformly-distributed rather than normal. Make a call. As long as your graph supports it, I'm happy.

Extra: some other graphs of interest here: a "density plot" is a smoothed-out version of the histogram, which saves you choosing a number of bins (though there is a "smoothing parameter" lurking in the background for which a default choice been made for you):

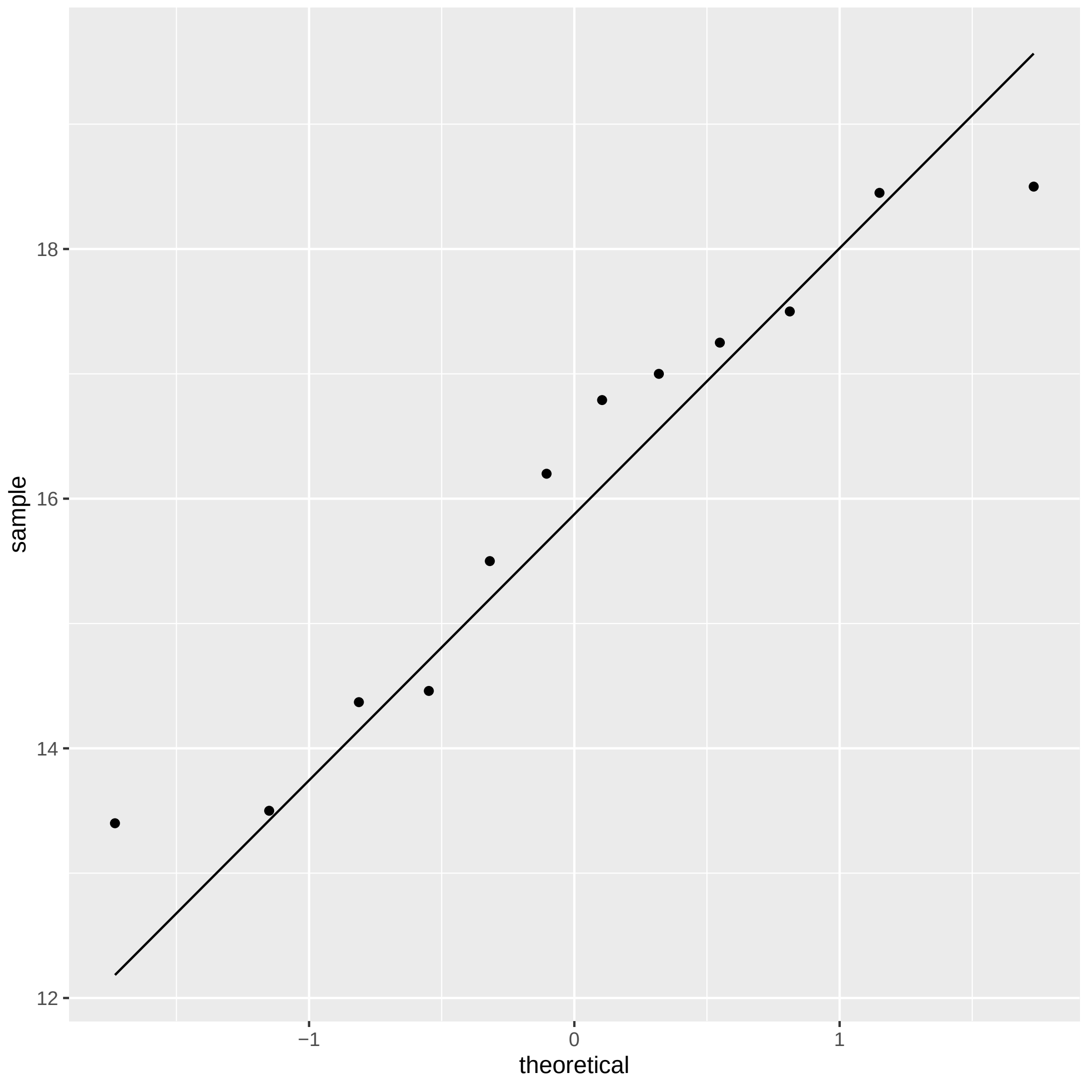
```
ggplot(millwrights, aes(x=wage)) + geom_density()
```



This one is actually slightly skewed to the *left*, though the density is high all the way across, so it is not far from being “flat-topped” or uniformly-distributed.

Or you might have thought of a normal quantile plot:

```
ggplot(millwrights, aes(sample=wage)) + stat_qq() + stat_qq_line()
```



I wasn't asking about normality, but rather about shape generally, so you have to be careful explicitly to compare your conclusion from this one with a normal distribution: it is approximately normal in shape, or you could be picky (looking at the highest and lowest values) and describe it as *short*-tailed compared to the normal. (I don't know whether I explicitly talked about "short-tailed", but the idea here is that it doesn't go up as far as the normal, and it also doesn't go down as far as the normal.)

As with rejecting a null hypothesis, it's not enough to say "not normal", if that's what you think it is; you need to say *how* it's not normal (in this case, that the tails are too short compared to the normal).

All of the above is really saying that there aren't any big problems with the shape, and that a t-test for the mean would actually have been OK (and thus we should really be doing one instead of a sign test). But I got this from the section of a textbook that talked about the sign test (among other things). Maybe the argument is that we are specifically interested in the *median* wage, and if there had been an unusually high one, we wouldn't have wanted our conclusion to be unduly affected by it.

(d) (4 marks) Use `smmr` to run a suitable sign test on these data. What do you conclude, in the context of the data? Is it consistent with your guess from part (b)? Explain briefly.

Solution:

The expected thing: data frame, column name and null median. This is the default two-sided, so you don't need an `alternative`:

```
library(smmr)
sign_test(millwrights, wage, 15)

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

The P-value 0.388 is not small, so we do not reject the null hypothesis that the median wage is $15. In other words, there is no evidence that the median hourly wage for all construction millwrights differs from $15. (The official's belief was justified.)

Compare what you concluded here with what you guessed in part (b). I guessed that I would not reject the null, so I was consistent, but your results might have been inconsistent. Say what you saw, whatever it was. Being inconsistent is full marks here, if that's what you were. (But it is an invitation to consider *why* your guess was wrong before. See my discussion at the end of part (b).)

(e) (2 marks) Obtain a 90% confidence interval for the median. (You can use something from `smmr` for this.) Is it reasonable that the interval contains 15? Explain briefly.

Solution: Add `conf.level` and a decimal number to get a non-default confidence interval:

```
ci_median(millwrights, wage, conf.level = 0.90)

## [1] 14.37552 17.49862
```

We didn't reject a median of 15, so it is entirely reasonable that the confidence interval contains $15, and it does.

Extra: if you want to be precise, the P-value was larger than 0.10, so the corresponding (to $\alpha = 0.10$) 90% confidence interval should contain the hypothesized median, and does. The P-value is even larger than 0.30, so it should be inside a 70% confidence interval:

```
ci_median(millwrights, wage, conf.level = 0.70)

## [1] 14.46350 17.24625
```

but outside a 60% one (why?):

```
ci_median(millwrights, wage, conf.level = 0.60)

## [1] 15.50324 16.99975
```

I have never actually used this level of confidence interval in practice, but that's what the theory says.