MECH481A6: Engineering Data Analysis in R

Chapter 9 Homework: Transformations

Brad Portouw

10 November, 2022

Load packages

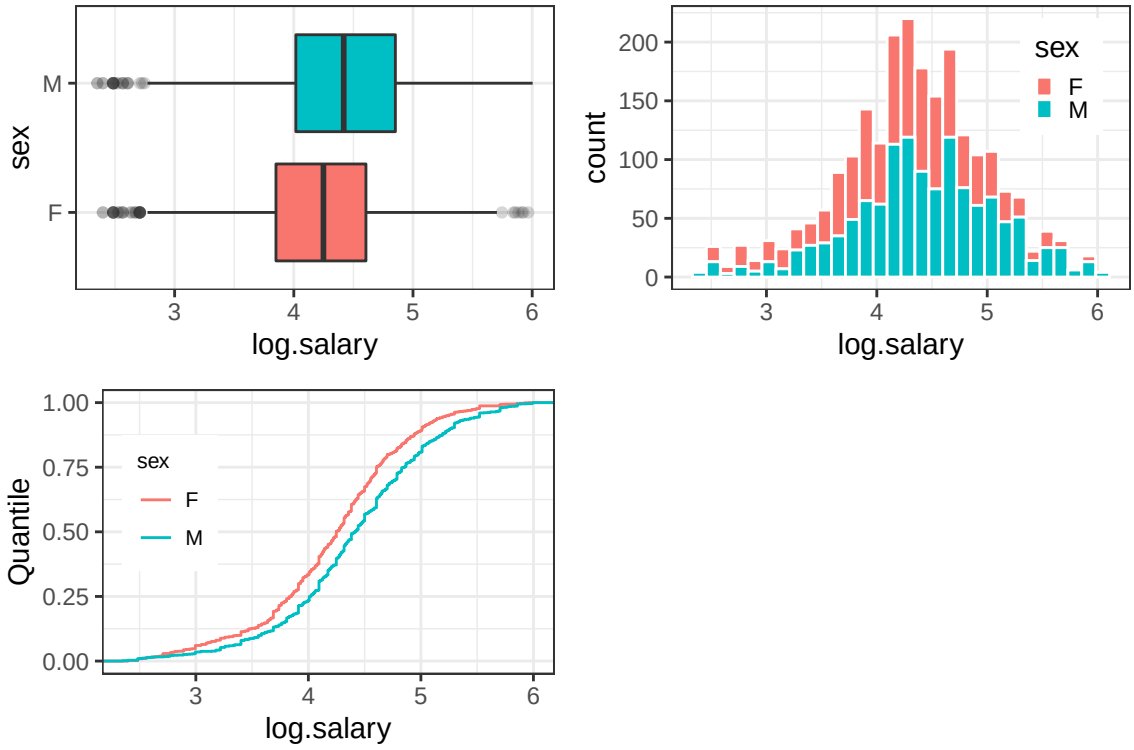
Chapter 9 Homework

This homework will give you practice at transforming and visualizing data and fitting a distribution to a set of data. Note that much of the code needed to complete this homework can be adapted from the Coursebook Exercises in Chapter 9.

When a question asks you to make a plot, remember to set a theme, title, subtitle, labels, colors, etc. It is up to you how to personalize your plots, but put in some effort and make the plotting approach consistent throughout the document. For example, you could use the same theme for all plots. I also like to use the subtitle as a place for the main summary for the viewer.

Question 1

Recreate Figure 9.8 (the three EDA plots based on `salary_ps2$salary`), but show the plots on a log-scale x-axis. Plot the histogram with 30 bins and move the legends so that they don't block the data. Does the data in these plots appear more symmetric about the median? Why or why not?

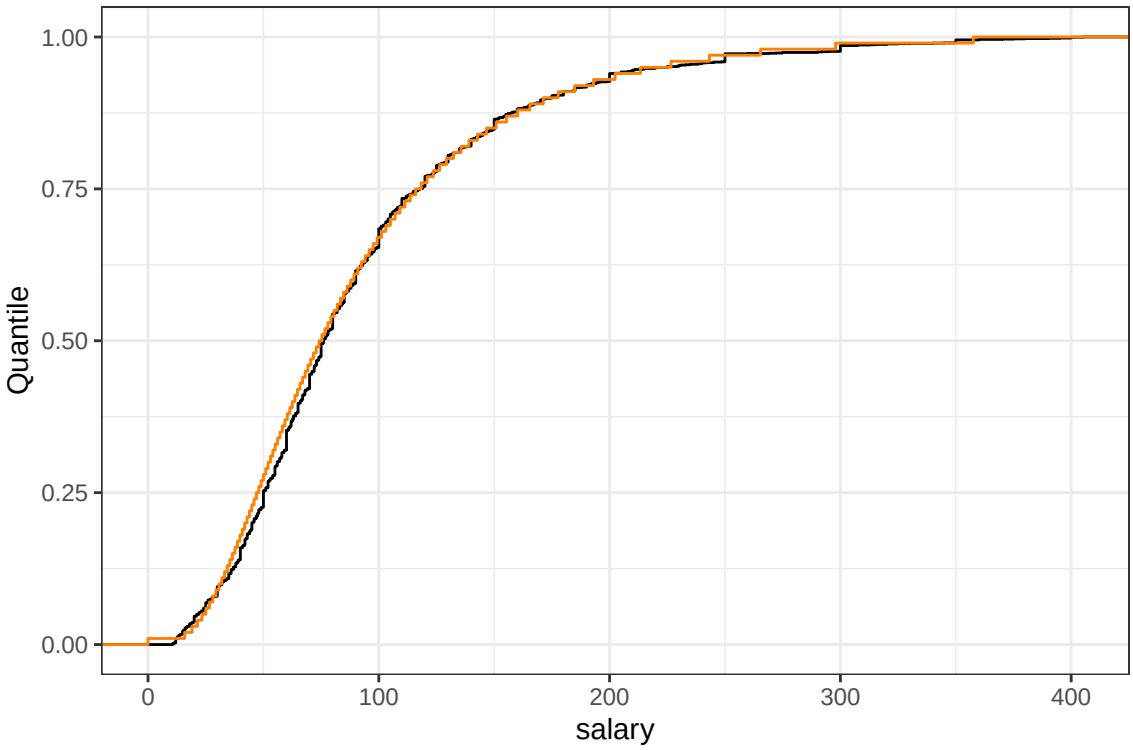


In a logarithmic scale, the distribution of slaries for both men and women seem to have a symmetric distribution about the median. This does keep in mind that the data for men and women were plotted separately, with different medians and averages. Without this logarithmic scaling, the data would have appeared to be skewed toward employees with lower salaries.

Question 2

Modify the code that created the `sal_simulate` data frame to create a variable that simulates quantiles from a *cumulative distribution*. Plot these data (instead of a histogram). Hint: instead of `rlnorm()` you will need to use a different log density function that takes a vector of quantiles as input (you will need to specify the quantile vector). Type ?Lognormal into the Console for help.

```
## meanlog   sdlog
##       4       1
```



Question 3

Mutate the `salary_ps2` data frame to create a new column variable that takes the log of the salary data (call that variable `log.salary`). Then use `fitdistr()` to fit a *normal distribution* to `log.salary`. What are the resultant parameter estimates for the mean and sd? Hint: the output of `fitdistr()` is a list; look in the `estimate` entry for these parameters. How close are these estimates to those calculated in section 9.6.4 of the Coursebook?

```
##     mean       sd
## 4.322399 0.669309
```

```
## mean   sd
## 4.32 0.67
```

Looking at section 9.6.4 of the course text the estimates of the log normal distribution of the mean and standard deviation are:

mean log = 4.32 sdlog = 0.67

These values were also rounded to 2 decimal places. If the values for the normal distribution for the log data were also rounded to 2 decimal places, the results are identical.