MECH481A6: Engineering Data Analysis in R

Chapter 9 Homework: Transformations

Ethan Rutledge

08 December, 2023

Load packages

```
# load packages for current session
library(tidyverse)
library(gridExtra) # or library(patchwork) for arranging figures
library(MASS) # for fitting distributions to your data
```

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.

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?

```
salaries_raw <- read_csv("../data/salary_ch9.csv")

salary_ps <- salaries_raw %>%
  mutate(salary = salary/1000) %>%
  filter(salary < 500, salary > 10)
```

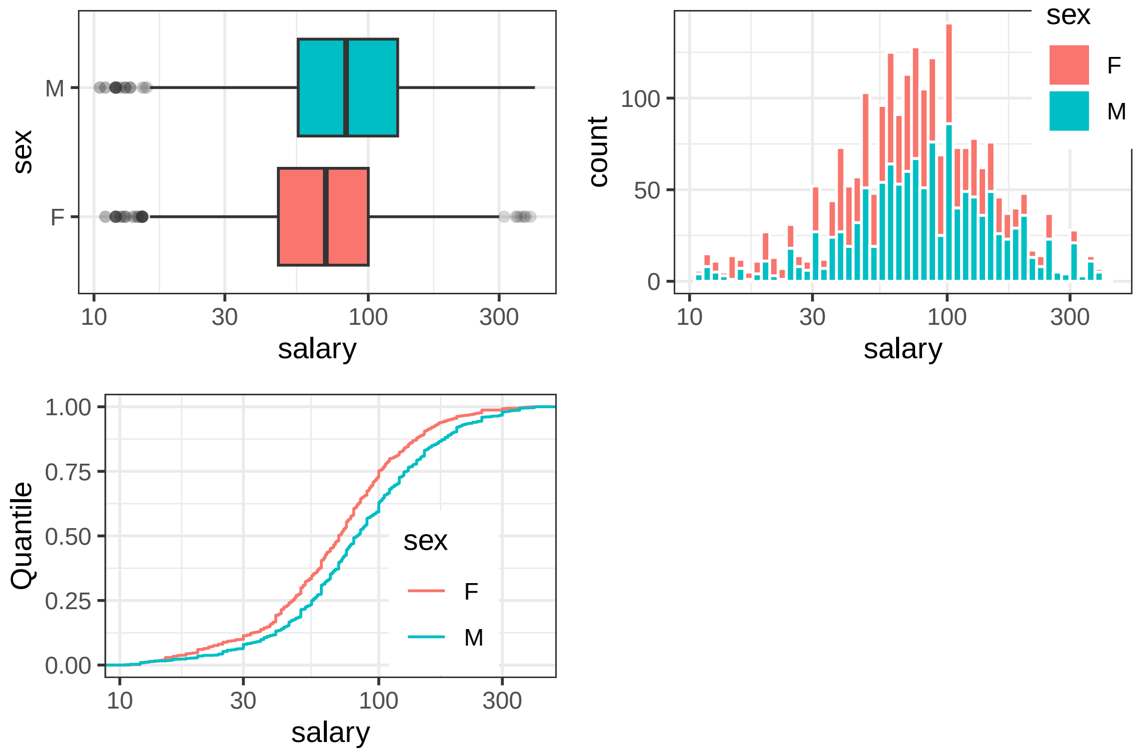
Answer: The data does seem to be more symetric about the median. This is because it has pretty close to a log-normal distribution so this type of plotting makes it appear more symetrical about the median.

```
box1 <- ggplot(data = salary_ps,
       aes(y = sex,
           x = salary,
           fill = sex)) +
  geom_boxplot(outlier.alpha = 0.2) +
  scale_x_log10() +
  theme_bw() +
  theme(legend.position = "none")

hist1 <- ggplot(data = salary_ps,
       aes(x = salary,
           fill = sex)) +
  geom_histogram(color = "white",
                 bins = 50) +
  scale_x_log10() +
  theme_bw() +
  theme(legend.position = c(0.9, 0.8))

cdf1 <- ggplot(data = salary_ps,
       aes(x = salary,
           color = sex)) +
  stat_ecdf() +
  scale_x_log10() +
  theme_bw() +
  ylab("Quantile") +
  theme(legend.position = c(0.75, 0.3))

grid.arrange(box1, hist1, cdf1, nrow = 2, ncol = 2)
```



Question 2

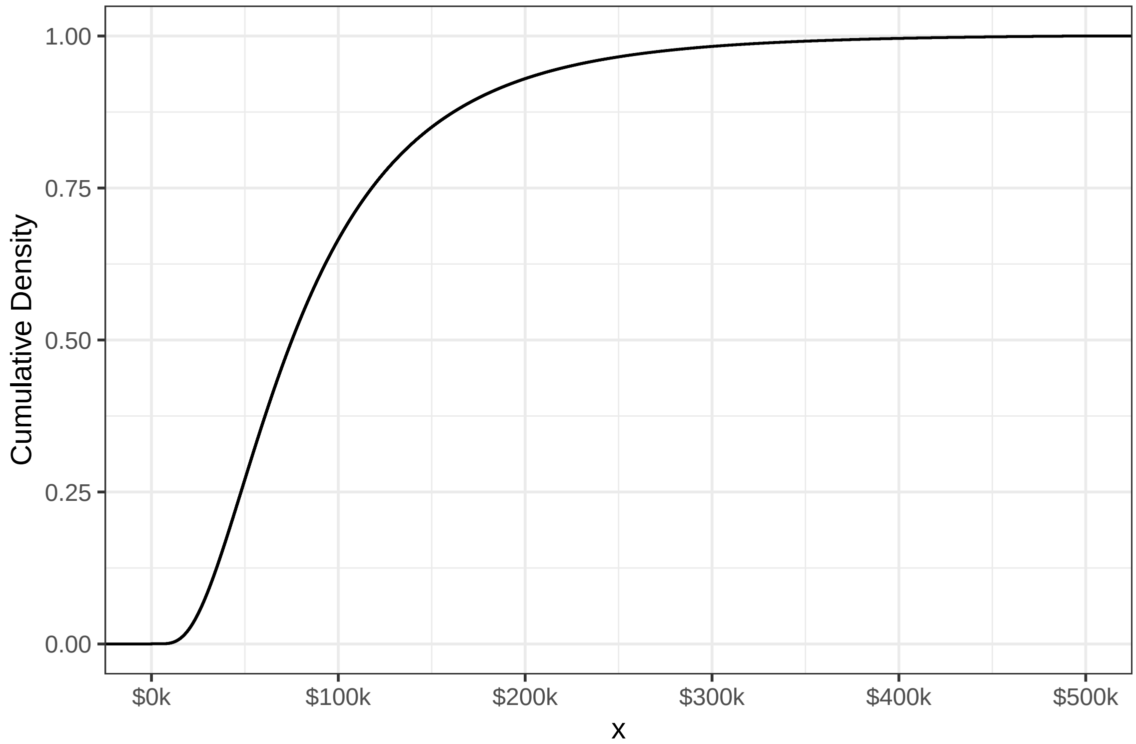
Modify the code that created the `sal_simulate` data frame to create a variable that simulates quantiles from a *cumulative distribution*. Plot these data (instead of a histogram). Hint: instead of `rlnorm()` you will need to use a different log density function that takes a vector of quantiles as input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

```r
#fit the data to a lognormal distribution
fit.lnorm <- fitdistr(salary_ps$salary, densfun = "log-normal")

#simulate quantiles

sal_simulate <- tibble(x = qlnorm(p = seq(0,1,length.out=length(salary_ps$salary)),
                                  meanlog = fit.lnorm$estimate[[1]],
                                  sdlog = fit.lnorm$estimate[[2]]))

ggplot() +
  stat_ecdf(data = sal_simulate, aes(x=x)) +
  ylab("Cumulative Density") +
  scale_x_continuous(labels = scales::label_dollar(suffix = "k"),
                     limits = c(0,500)) +
  theme_bw()
```



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?

```
salary_ps <- mutate(salary_ps, log.salary = log(salary))

fit_norm <- fitdistr(salary_ps$log.salary, densfun = "normal")

mean_est <- fit_norm$estimate[1]
sd_est <- fit_norm$estimate[2]

mean_est
```

```
##     mean
## 4.322399
```

```
sd_est
```

```
##       sd
## 0.669309
```

Answer: They are the exact same as the ones calcualted in the book