Homework 1 Solution to R Problems

Math 445, Spring 2017

Problem 1 (A revised version of Chapter 2, Exercise 5) Import data from the General Social Survey Case Study in Section 1.6 into R.

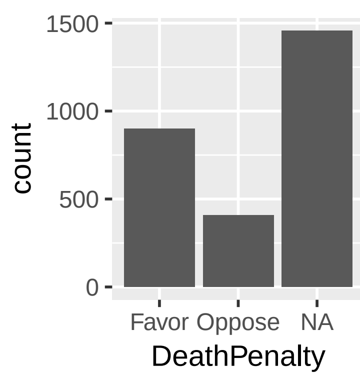
```
library(resampledata) # for textbook data sets
library(dplyr)        # for data wrangling
library(ggplot2)      # for data visualization
```

(a) Create a table and a bar chart summarizing the responses to the question about the death penalty.

```
GSS2002 %>%
  group_by(DeathPenalty) %>%
  summarise(count = n())

ggplot(data = GSS2002, mapping = aes(x = DeathPenalty)) +
  geom_bar()
```

DeathPenalty	count
Favor	899
Oppose	409
NA	1457

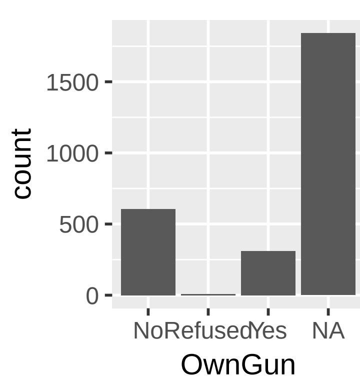


(b) Create a table and a bar chart summarizing the responses to the question about gun ownership.

```
GSS2002 %>%
  group_by(OwnGun) %>%
  summarise(count = n())

ggplot(data = GSS2002, mapping = aes(x = OwnGun)) +
  geom_bar()
```

OwnGun	count
No	605
Refused	9
Yes	310
NA	1841



(c) Create a table comparing the responses to the death penalty and gun ownership questions.

```r
views_tbl <-
  GSS2002 %>%
  group_by(DeathPenalty, OwnGun) %>%
  summarise(count = n())
views_tbl
```

```
## Source: local data frame [11 x 3]
## Groups: DeathPenalty [?]
##
##    DeathPenalty  OwnGun count
##          <fctr>  <fctr> <int>
## 1         Favor      No   375
## 2         Favor Refused     7
## 3         Favor     Yes   243
## 4         Favor      NA   274
## 5        Oppose      No   199
## 6        Oppose Refused     2
## 7        Oppose     Yes    59
## 8        Oppose      NA   149
## 9            NA      No    31
## 10           NA     Yes     8
## 11           NA      NA  1418
```

You can leave the table as it is above, or you can make it display as a contingency table using the following code

```r
library(tidyr)
views_tbl %>%
  na.omit %>%
  spread(DeathPenalty, count)
```

```
## # A tibble: 3  3
##    OwnGun Favor Oppose
## *  <fctr> <int>  <int>
## 1      No   375    199
## 2 Refused     7      2
## 3     Yes   243     59
```

(d) What proportion of gun owners favor the death penalty? Does it appear to be different from the proportion among those who do not own guns?

From the above table, it is easy to see that 80.5% of gun owners favor the death penalty and 65.3% of those who do not own guns favor the death penalty.

Problem 2 (A revised/expanded version of Chapter 2, Exercise 6)

(a) Compute the following numeric summaries for the height changes (`Ht.change`): minimum, .25 quantile (Q1), median, .75 quantile (Q3), mean, standard deviation, and the count.

This can be done using `summary` and the individual functions

```r
summary(Spruce$Ht.change)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   23.20   30.10   30.93   38.17   51.50
```

```r
sd(Spruce$Ht.change)
```

```
## [1] 11.04943
```

```r
length(Spruce$Ht.change)
```

```
## [1] 72
```

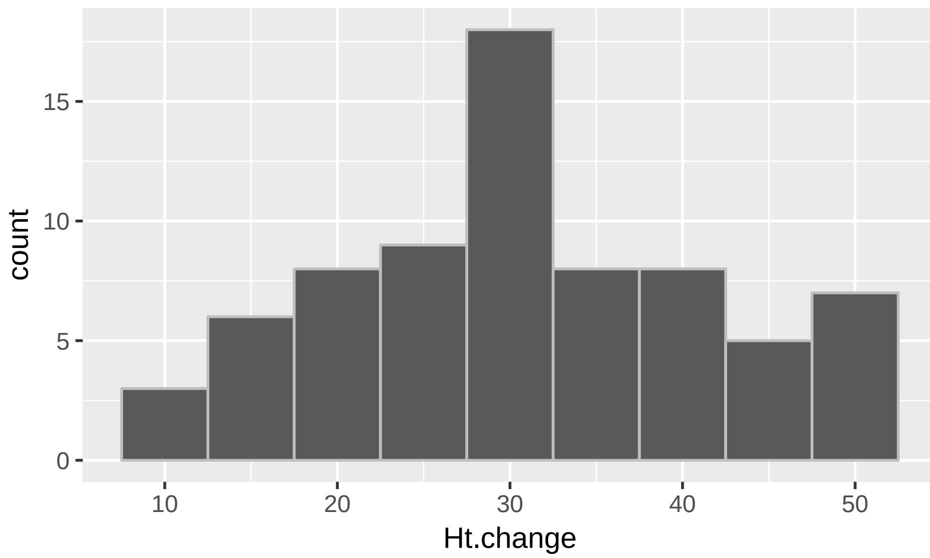
or using the `summarize` function

```
Spruce %>%
  summarize(min = min(Ht.change),
            Q1 = quantile(Ht.change, probs = .25),
            median = median(Ht.change),
            Q3 = quantile(Ht.change, probs = .75),
            mean = mean(Ht.change),
            sd = sd(Ht.change),
            n = n())
```

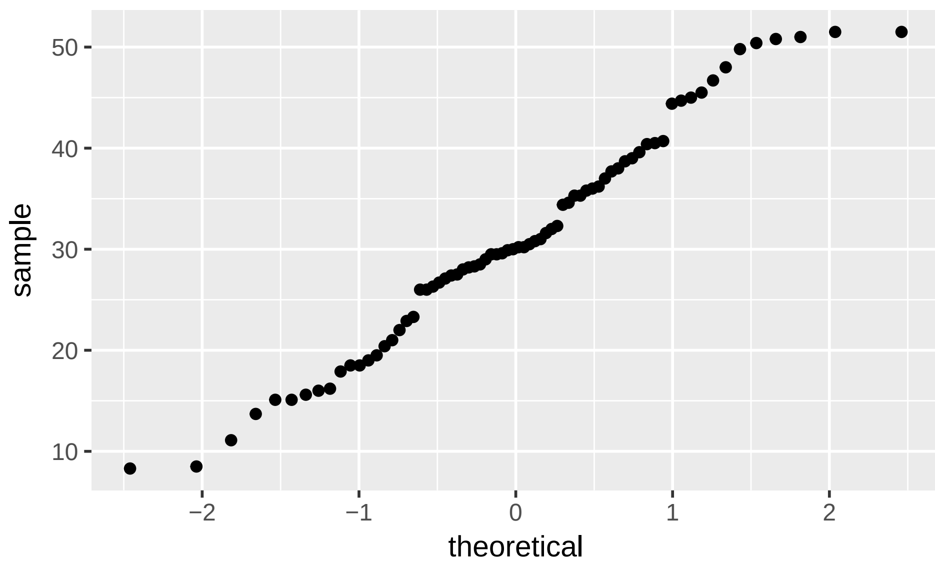
```
##   min   Q1 median     Q3     mean       sd  n
## 1 8.3 23.2   30.1 38.175 30.93333 11.04943 72
```

(b) Create a histogram and normal quantile plot for the height changes of the seedlings. Is the distribution approximately normal?

```
ggplot(data = Spruce, mapping = aes(x = Ht.change)) +
  geom_histogram(binwidth = 5, colour = "gray")
```




```
ggplot(data = Spruce, mapping = aes(sample = Ht.change)) +
  stat_qq()
```



The histogram is approximately unimodal and symmetric, and the normal quantile plot does not show significant deviations from the diagonal, so it appears that the distribution is approximately normal.

(c) Compute the following numeric summaries for the height changes (`Ht.change`) by whether or not they were fertilized plots (`Fertilizer`): minimum, .25 quantile (Q1), median, .75 quantile (Q3), mean, standard deviation, and the count.

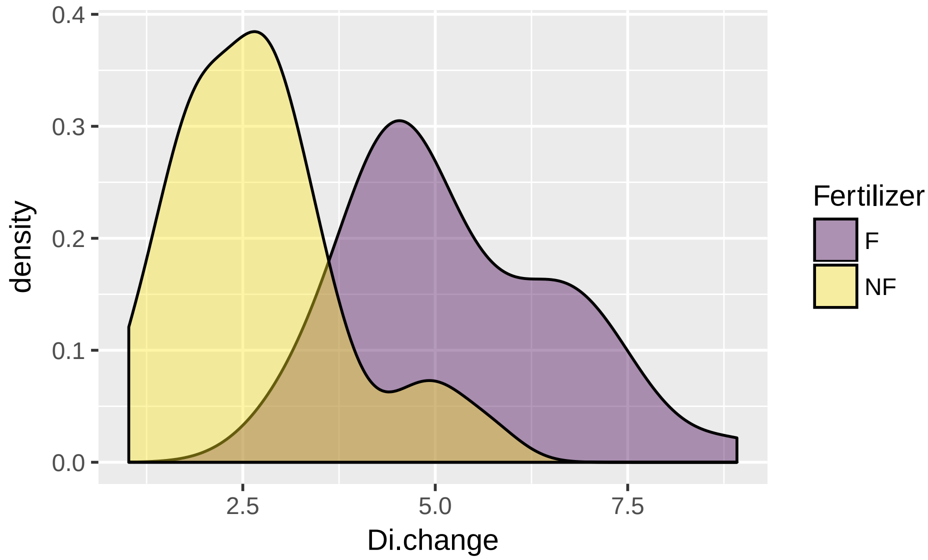
```
Spruce %>%
  group_by(Fertilizer) %>%
  summarise(min = min(Ht.change),
            Q1 = quantile(Ht.change, probs = .25),
            median = median(Ht.change),
            Q3 = quantile(Ht.change, probs = .75),
            mean = mean(Ht.change),
            sd = sd(Ht.change),
            n = n())
```

```
## # A tibble: 2  8
##   Fertilizer   min     Q1 median     Q3     mean       sd     n
##       <fctr> <dbl>  <dbl>  <dbl>  <dbl>    <dbl>    <dbl> <int>
## 1          F  27.4 30.725  37.35 44.775 38.28889 7.980540    36
## 2         NF   8.3 17.475  23.10 28.625 23.57778 8.525193    36
```

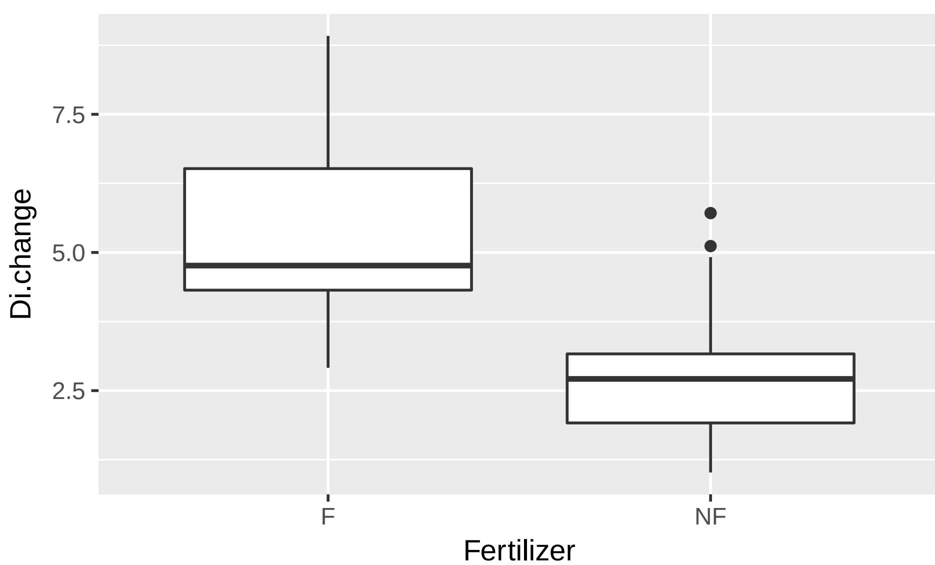
(d) Create a plot to compare the distribution of the change in diameters of the seedlings (`Di.change`) by whether or not they were fertilized plots. What does the plot reveal?

A few options for comparison are below. You should be able to interpret these plots.

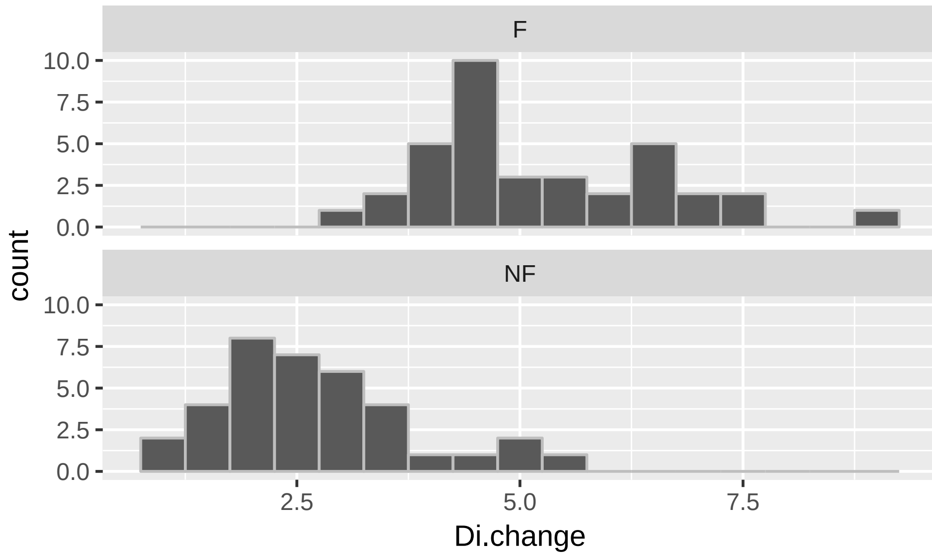
```
library(viridis)
ggplot(data = Spruce, mapping = aes(x = Di.change, fill = Fertilizer)) +
  geom_density(alpha = 0.4) +
  scale_fill_viridis(discrete = TRUE)
```




```r
ggplot(data = Spruce, mapping = aes(x = Fertilizer, y = Di.change)) +
  geom_boxplot()
```




```r
ggplot(data = Spruce, mapping = aes(x = Di.change)) +
  geom_histogram(color = "gray", binwidth = .5) +
  facet_wrap(~ Fertilizer, ncol = 1)
```



(e) Compute the following numeric summaries for the height changes (`Ht.change`) by whether or not they were fertilized plots (`Fertilizer`) and competition (`Competition`) status: minimum, .25 quantile (Q1), median, .75 quantile (Q3), mean, standard deviation, and the count.

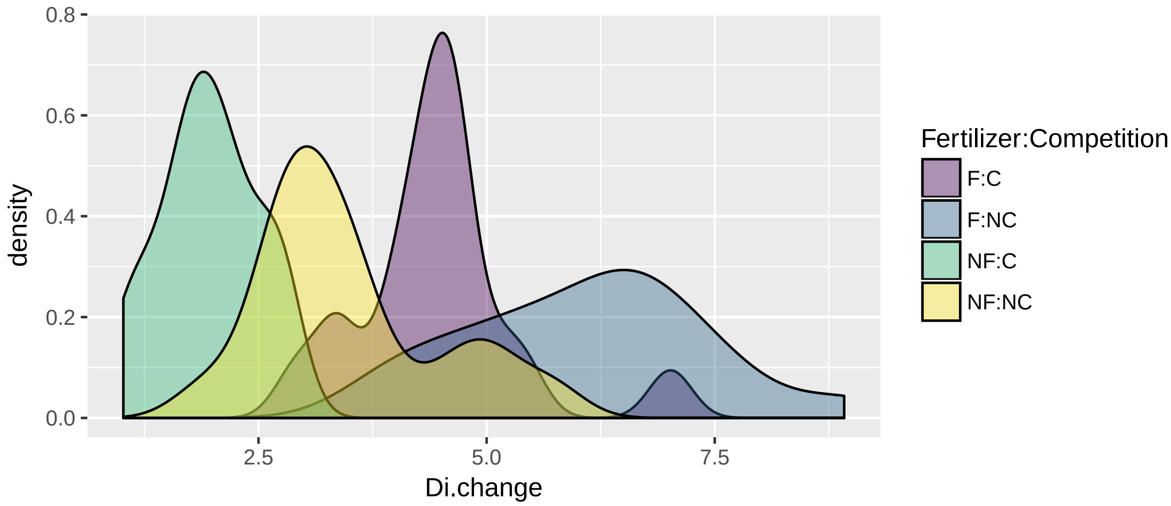
```
Spruce %>%
  group_by(Fertilizer, Competition) %>%
  summarise(min = min(Ht.change),
            Q1 = quantile(Ht.change, probs = .25),
            median = median(Ht.change),
            Q3 = quantile(Ht.change, probs = .75),
            mean = mean(Ht.change),
            sd = sd(Ht.change),
            n = n())
```

```
## Source: local data frame [4 x 9]
## Groups: Fertilizer [?]
##
##   Fertilizer Competition   min     Q1 median     Q3     mean       sd
##       <fctr>      <fctr> <dbl>  <dbl>  <dbl>  <dbl>    <dbl>    <dbl>
## 1          F           C  27.4 29.600  30.90 37.525 33.80556 6.266951
## 2          F          NC  30.0 36.825  44.55 49.350 42.77222 7.020290
## 3         NF           C   8.3 15.100  17.35 20.850 17.60000 5.487848
## 4         NF          NC  17.9 26.400  28.65 33.800 29.55556 6.621553
## # ... with 1 more variables: n <int>
```

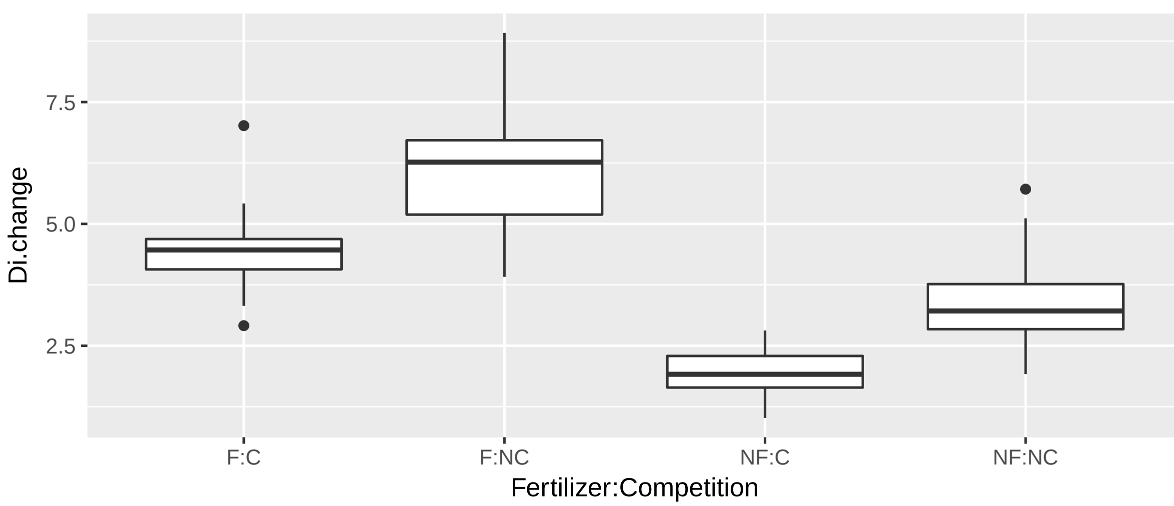
(f) Create a plot to compare the distribution of the change in diameters of the seedlings (`Di.change`) by whether or not they were fertilized plots and competition (`Competition`) status. What does the plot reveal?

A few options for comparison are below. You should be able to interpret these plots.

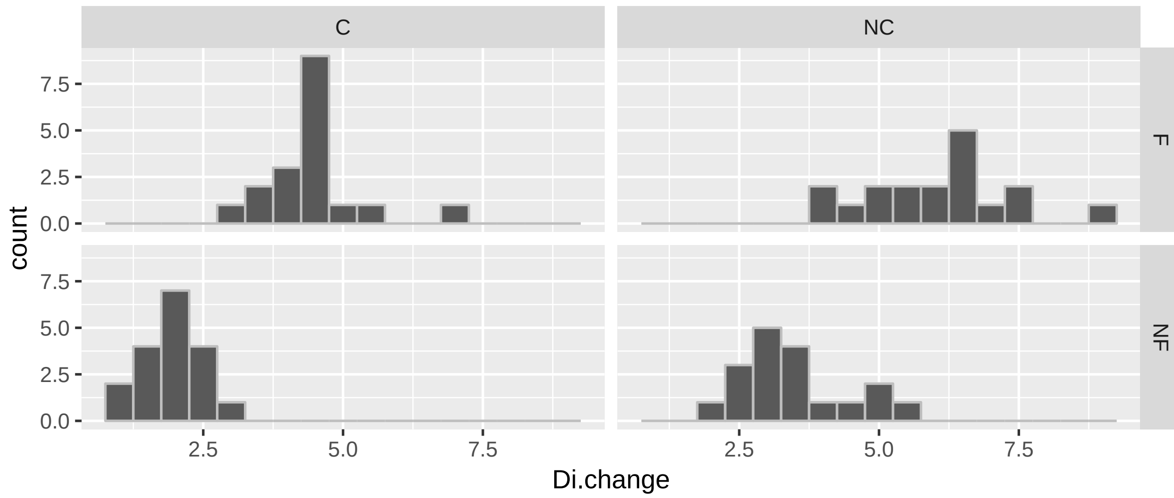
```
library(viridis)
ggplot(data = Spruce, mapping = aes(x = Di.change, fill = Fertilizer:Competition)) +
  geom_density(alpha = 0.4) +
  scale_fill_viridis(discrete = TRUE)
```




```
ggplot(data = Spruce, mapping = aes(x = Fertilizer:Competition, y = Di.change)) +
  geom_boxplot()
```

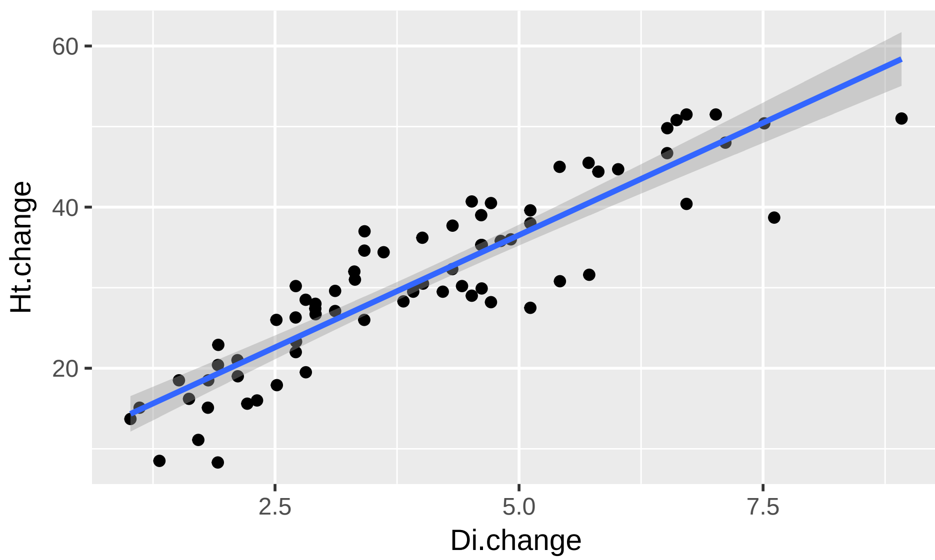



```
ggplot(data = Spruce, mapping = aes(x = Di.change)) +
  geom_histogram(color = "gray", binwidth = .5) +
  facet_grid(Fertilizer ~ Competition)
```



(g) Create a scatter plot of the height changes against the diameter changes and describe the relationship.

```
ggplot(data = Spruce, mapping = aes(x = Di.change, y = Ht.change)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Problem 3 Chapter 2, Exercise 8

(a) The CDF of the exponential distribution is given by $F(x) = 1 - e^{-\lambda x}$. We derive the quantile function (i.e. inverse CDF) by setting a cumulative probability, p, and solving for x:

$$p = 1 - e^{-\lambda x} \implies F^{-1}(p) = -\frac{\log(1-p)}{\lambda}$$

Thus

p — $F^{-1}(p)$ .25 (1st quartile) — $-\dfrac{\log(1-.25)}{\lambda}$ .5 (median) — $-\dfrac{\log(1-.5)}{\lambda}$ .75 (3rd quartile) — $-\dfrac{\log(1-.75)}{\lambda}$

(b) The CDF for the Pareto distribution is found by:

$$F(x) = \int_1^x \alpha/t^{\alpha+1} dt = -t^{-\alpha}\Big|_1^x = 1 - x^{-\alpha}$$

The quantile function is then found by

$$p = 1 - x^{-\alpha} \implies F^{-1}(p) = (1-p)^{-1/\alpha}$$

p — $F^{-1}(p)$ .25 (1st quartile) — $(1 - .25)^{-1/\alpha}$ .5 (median) — $(1 - .5)^{-1/\alpha}$ .75 (3rd quartile) — $(1 - .75)^{-1/\alpha}$

Problem 4 Chapter 2, Exercise 10

(a) You can find percentiles of a normal distribution using the ‘qnorm‘ function:

```
qnorm(.3, mean = 10, sd = 17)

## [1] 1.085191

qnorm(.6, mean = 10, sd = 17)

## [1] 14.3069
```

(b) You can find quantiles of a normal distribution using the ‘qnorm‘ function:

```
qnorm(.1, mean = 25, sd = 32)

## [1] -16.00965

qnorm(.9, mean = 25, sd = 32)

## [1] 66.00965
```

Problem 5 Chapter 2, Exercise 12

To find the pth quantile, solve the following for x

$$p = 1 - 9/x^2 \implies F^{-1}(p) = \frac{3}{\sqrt{1-p}}$$