Contents

Chapter 1

Basic Probability and Statistics

Problem 1

Think about what it means to scale this data. We have a set of data X and this data has a mean μ_x and standard deviation σ. We want to linearly transform the data which means:

$$Y = a + bX$$
$$\mu_y = a + b\mu_x$$
$$\sigma_y = |b|\sigma_x$$

Now to answer (a) and (b) we can just solve these equations. Let's solve for standard deviation.

$$15 = |b|10$$
$$1.5 = |b|$$

This means that $b = \pm 1.5$ and we can now solve for the mean to get that $a = 47.5$ or $a = 152.5$. We can transform it with either the first set of values or the second. You don't want to use the second because that will reverse the ordering in the data.

Problem 2

(a)

Doing this in R is easy:

```
girls <- c(.4777, .4875, .4874, .4859, .4754, .4864,.4813,.4787,.4895,
          .4797,.4876,.4859,.4857,.4907, .5010,.4903,.4860,.4911,.4871,
```

```
            .4725,.4822,.4870,.4823,.4973)

num_births <- 3903
std <- sd(girls)
avg <- mean(girls)
expected_std <- sqrt(avg * (1 - avg) / num_births)
```

We get that std= .0064 and expected_std= .008

Let's quickly prove this formula for the expected standard error of a proportion.

Proof. We are trying to prove that $SE(X) = \sqrt{\frac{p(1-p)}{n}}$ for binary variables x_i which take on only values 0 and 1.

Begin by assuming that we have a population with m instances where $x = 1$ and $n - m$ instances where $x = 0$ We know that the standard deviation of a population is:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Let's just concern ourselves with the summation portion of this equation and see if we can massage it into the desirable form.

$$\begin{aligned}
\sum (x_i - \bar{x})^2 &= \sum (x_i - \bar{x})(x_i - \bar{x}) \\
&= \sum x_i^2 - 2x_i\bar{x} + \bar{x}^2 \\
&= \sum_{x_i=0} \bar{x}^2 + \sum_{x_i=1} 1 - 2\bar{x} + \bar{x}^2 \\
&= (n-m)\bar{x}^2 + m - 2m\bar{x} + m\bar{x}^2 \\
&= (n-m)\frac{m^2}{n^2} + m - 2m\frac{m}{n} + m\frac{m^2}{n^2} & \bar{x} = \frac{m}{n} \\
&= m + \frac{m^2}{n} - \frac{m^3}{n^2} - \frac{2m^2}{n} + \frac{m^3}{n^2} \\
&= m - \frac{m^2}{n} \\
&= m\left(1 - \frac{m}{n}\right) \\
&= np(1-p) & p = m/n
\end{aligned}$$

So we can finish by substituting this back into our original equation!

$$\begin{aligned}
\sigma &= \sqrt{\frac{np(1-p)}{n}} \\
&= \sqrt{p(1-p)}
\end{aligned}$$

Since the sample proportion is a mean the standard error is calculated like normal yielding

$$\begin{aligned} SE(X) &= \frac{\sigma_x}{\sqrt{n}} \\ &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

□

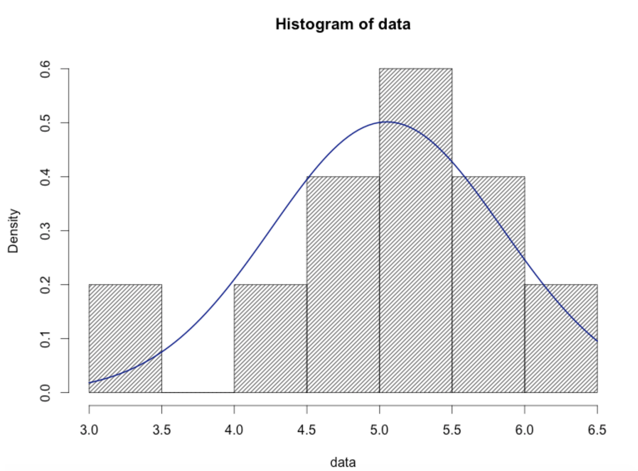
(b)

No they are not significant. If we run a χ^2 test we get $\chi^2 = .0019$ which is no where near what we need for a significant result.

Problem 3

```
cols <- 20
rows <- 1000
x <- replicate(cols, runif(rows))
data <- apply(a, 1, sum)
avg <- mean(data)
std <- sd(data)
hist(data, density=20)
curve(dnorm(x, mean=avg, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

The plot is:

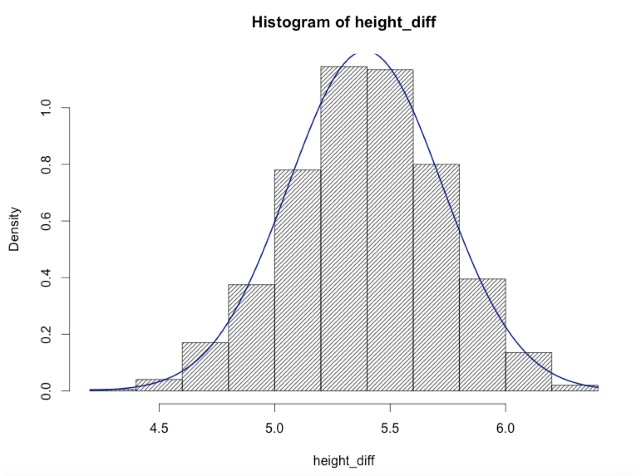


Problem 4

```
i = 1
height_diff = seq(1000) * 0
while (i <= 1000) {
  men_height <- rnorm(100, 69.1, 2)
  women_height <- rnorm(100, 63.7, 2.7)
  height_diff[i] = mean(men_height) - mean(women_height)
  i = i + 1
}

avg <- mean(height_diff)
std <- sd(height_diff)
hist(height_diff, density=20, freq=FALSE)
curve(dnorm(x, mean=avg, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

The mean of the difference is 5.05 and the standard deviation is .79 The plot is:



Problem 5

Expectation is linear so:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

As to standard deviation:

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Cov}[X,Y] \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Corr}[X,Y]\sigma_x\sigma_y\end{aligned}$$