Exercise 5.74: In an area having sandy soil, 50 small trees of a certain type were planted, and another 50 trees were planted in an area having clay soil. Let X = the number of trees planted in sandy soil that survive 1 year and Y = the number of trees planted in clay soil that survive 1 year. If the probability that a tree planted in sandy soil will survive 1 year is .7 and the probability of 1-year survival in clay soil is .6, compute an approximation to $P(-5 \leq X - Y \leq 5)$ (do not bother with the continuity correction).

Solution:
Recall that by the CLT since the number of trees is $n = 50 > 30$ we know that X and Y are approximately normal. Since $X_i \sim binom(50, .7)$ and $Y_i \sim binom(50, .6)$ we get,

$$\mu_X = 50(.7) = 35,$$

$$\mu_Y = 50(.6) = 30.$$

Similarly we can calculate variance of each sample variable,

$$\rho_X^2 = 50(.7)(.3) = 10.5,$$

$$\rho_Y^2 = 50(.6)(.4) = 12.$$

Calculating the $E(X - Y)$ and $V(X - Y)$,

$$E(X - Y) = E(X) - E(Y) = 35 - 30 = 5$$

$$V(X - Y) = V(X) + (-1)^2 V(Y) = 10.5 + 12 = 22.5.$$

Then finally we can calculate the probability by standardizing $X - Y$,

$$\begin{aligned} P(-5 \leq X - Y \leq 5) &= P(\frac{0-5}{\sqrt{22.5}} \leq \frac{(X-Y)-5}{\sqrt{22.5}} \leq 0) \\ &= P(-2.11 \leq Z \leq 0), \\ &= P(Z \leq 0) - P(Z \leq -2.11), \\ &= .4826. \end{aligned}$$

Exercise 5.92: 1. Show that $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$.

Solution:
By the definition of Covariance, we have the following formula,

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

Applying this formula and using the linearity of expectations we get the following,

$$\begin{aligned}
Cov(X, Y+Z) &= E[X(Y+Z)] - E(X)E(Y+Z), \\
&= E[XY + XZ] - E(X)E(Y+Z), \\
&= E(XY) + E(XZ) - E(X)E(Y+Z), \\
&= E(XY) + E(XZ) - E(X)E(Y) - E(X)E(Z), \\
&= E(XY) - E(X)E(Y) + E(XZ) - E(X)E(Z), \\
&= Cov(XY) + Cov(XZ).
\end{aligned}$$

2. Let X_1 and X_2 be quantitative and verbal scores on one aptitude exam, and let Y_1 and Y_2 be corresponding scores on another exam. If $Cov(X_1, Y_1) = 5$, $Cov(X_1, Y_2) = 1$, $Cov(X_2, Y_1) = 2$, and $Cov(X_2, Y_2) = 8$, what is the covariance between the two total scores $X_1 + X_2$ and $Y_1 + Y_2$?

   Solution:
   By the previous problem we know that,

   $$Cov((X_1 + X_2), (Y_1 + Y_2)) = Cov((X_1 + X_2), Y_1) + Cov((X_1 + X_2), Y_2).$$

   Applying the formula again we get that,

   $$Cov((X_1 + X_2), (Y_1 + Y_2)) = Cov(X_1, Y_1) + Cov(X_2, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_2).$$

   By substitution we know that,

   $$Cov((X_1 + X_2), (Y_1 + Y_2)) = 5 + 2 + 1 + 8 = 16.$$

Exercise 6.2: The National Health and Nutrition Examination Survey (NHANES) collects demographic, socioeconomic, dietary, and healthrelated information on an annual basis. Here is a sample of 20 observations on HDL cholesterol level (mg/dl) obtained from the 2009–2010 survey (HDL is “good” cholesterol; the higher its value, the lower the risk for heart disease)

1. Calculate a point estimate of the population mean HDL cholesterol level.

   Solution:
   A sample mean, $\overline{x}$ is a point estimate of the population mean.

   Console:

```
> x <- c(35, 49, 52, 54, 65, 51, 51, 47, 86, 36,
         46, 33, 39, 45, 39, 63, 95, 35, 30, 48)
> mean(x)
[1] 49.95
```

2. Making no assumptions about the shape of the population distribution, calculate a point estimate of the value that separates the largest 50% of HDL levels from the smallest 50%

Solution:
By definition we know that the median is the value that separates the largest 50% and smallest 50%. The point estimate for the population median is the sample median.

Console:

```
> x <- c(35, 49, 52, 54, 65, 51, 51, 47, 86, 36,
         46, 33, 39, 45, 39, 63, 95, 35, 30, 48)
> median(x)
[1] 47.5
```

3. Calculate a point estimate of the population standard deviation.

Solution:
The sample variance S^2 is calculated by,

$$S^2 = \frac{\sum_{i=1}^{20}(x_i - \overline{X})^2}{20 - 1}.$$

Taking the square root we get the sample standard deviation, through r we see,

Console:

```
> x <- c(35, 49, 52, 54, 65, 51, 51, 47, 86, 36,
         46, 33, 39, 45, 39, 63, 95, 35, 30, 48)
> sqrt(var(x))
[1] 16.81001
```

4. An HDL level of at least 60 is considered desirable as it corresponds to a significantly lower risk of heart disease. Making no assumptions about the shape of the population distribution, estimate the proportion p of the population having an HDL level of at least 60.

Solution:
Using our sample data we can estimate that approximately 20% of the values are above 60.

Console:

```
> x <- c(35, 49, 52, 54, 65, 51, 51, 47, 86, 36,
         46, 33, 39, 45, 39, 63, 95, 35, 30, 48)
> y <- c()
> for (val in x){
    if (val > 59)
        y <- c(y,val)
}

> length(y)/length(x)
[1] 0.2
```

Exercise 6.4: Prior to obtaining data, denote the beam strengths by $X_1, ..., X_m$ and the cylinder strengths by $Y_1, ..., Y_n$. Suppose that the $X_i's$ constitute a random sample from a distribution with mean μ_1 and standard deviation σ_1 and that the $Y_i's$ form a random sample (independent of the $X_i's$) from another distribution with mean μ_2 and standard deviation σ_2.

1. Use rules of expected value to show that $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. Calculate the estimate for the given data.

Solution:
From the rules of expected values we know,

$$\begin{aligned}
E(\overline{X} - \overline{Y}) &= E(\overline{X}) - E(\overline{Y}), \\
&= \frac{1}{m}\sum_{i=1}^{m} E(X_i) - \frac{1}{n}\sum_{i=1}^{n} E(Y_i), \\
&= \frac{1}{m} mE(X_1) - \frac{1}{n} nE(Y_1), \\
&= \mu_1 - \mu_2.
\end{aligned}$$

Therefore $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. Estimating $\mu_1 - \mu_2$ with r,
Console:

```
> x <- c(5.9, 7.2, 7.3, 6.3, 8.1, 6.8, 7.0, 7.6, 6.8, 6.5,
         7.0, 6.3, 7.9, 9.0, 8.2, 8.7, 7.8, 9.7, 7.4, 7.7,
         9.7, 7.8, 7.7, 11.6, 11.3, 11.8, 10.7)

> y <- c(6.1, 5.8, 7.8, 7.1, 7.2, 9.2, 6.6, 8.3, 7.0, 8.3,
         7.8, 8.1, 7.4, 8.5, 8.9, 9.8, 9.7, 14.1, 12.6, 11.2)

> mean(x) - mean(y)
[1] 0.4342593
```

2. Use rules of variance from Chapter 5 to obtain an expression for the variance and standard deviation (standard error) of the estimator in part (a), and then compute the estimated standard error.

Solution:
Since the X and Y are independent we get the following through the rule of variances,

$$\begin{aligned} V(\overline{X} - \overline{Y}) &= V(\overline{X}) + V(\overline{Y}), \\ &= \frac{1}{m^2}\sum_{i=1}^{m} V(X_i) + \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i), \\ &= \frac{1}{m^2} mV(X_1) + \frac{1}{n^2} nV(Y_1), \\ &= \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}. \end{aligned}$$

Using r to estimate $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$ we get, the sample variience and the square root gives us the standard deviation,
Console:

```
> var(x)/length(x) + var(y)/length(y)
[1] 0.3233635
> sqrt(var(x)/length(x) + var(y)/length(y))
[1] 0.5686506
```

3. Calculate a point estimate of the ratio $\frac{\sigma_1}{\sigma_2}$ of the two standard deviations.

Solution:
We use the sample variance S^2 as a point estimate of the variance so we can get the standard deviation by simply square rooting the sample variance then computing the ratio.

Console:

```
> sqrt(var(x))/sqrt(var(y))
[1] 0.7887133
```

Exercise 6.10: Using a long rod that has length μ, you are going to lay out a square plot in which the length of each side is μ. Thus the area of the plot will be μ^2. However, you do not know the value of μ, so you decide to make n independent measurements $X_1, X_2, ..., X_n$ of the length. Assume that each X_i has mean μ (unbiased measurements) and variance s^2

1. Show that $\overline{X}^2$ is not an unbiased estimator for μ^2.

Solution:
By our variance and expected value definition of $E(\overline{X}^2)$,

$$E(\overline{X}^2) = V(\overline{X}) - E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2.$$

Therefore since,

$$E(\overline{X}^2) \neq \mu^2$$

we know that $E(\overline{X}^2)$ is a biased estimator.

2. For what value of k is the estimator $\overline{X}^2 - kS^2$ an unbiased for m^2?

Solution:
By the linearity of the expected value we know that,

$$E(\overline{X}^2 - kS^2) = E(\overline{X}^2) - kE(S^2).$$

By the previous problem and since S^2 is an unbiased estimator for σ^2 we know that,

$$E(\overline{X}^2 - kS^2) = \frac{\sigma^2}{n} + \mu^2 - k\sigma^2.$$

Setting the right hand side to μ^2 and solving for k,

$$\begin{aligned}\frac{\sigma^2}{n} + \mu^2 - k\sigma^2 &= \mu^2,\\ k &= \frac{1}{n}.\end{aligned}$$

Exercise 6.12: Suppose a certain type of fertilizer has an expected yield per acre of μ_1 with variance σ^2, whereas the expected yield for a second type of fertilizer is μ_2 with the same variance σ^2. Let S_1^2 and S_2^2 denote the sample variances of yields based on sample sizes n_1 and n_2, respectively, of the two fertilizers. Show that the pooled (combined) estimator is unbiased,

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}.$$

Solution:
Through the linearlity of expectations we can pull all the constants out and separate the sample variences. We can also substitute σ^2 since S_i^2 are unbiased estimators,

$$\begin{aligned}E(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}) &= \frac{(n_1 - 1)}{(n_1 + n_2 - 2)}E(S_1^2) + \frac{(n_2 - 1)}{(n_1 + n_2 - 2)}E(S_2^2),\\ &= \frac{(n_1 + n_2 - 2)}{(n_1 + n_2 - 2)}\sigma^2,\\ &= \sigma^2.\end{aligned}$$

Thus our estimator is unbiased.