Textbook Exercises

6.12 Impact of the Population Proportion on SE Compute the standard error for sample proportions from a population with proportions $p = 0.8, p = 0.5, p = 0.3$, and $p = 0.1$ using a sample size of $n = 100$. Comment on what you see. For which proportion is the standard error the greatest? For which is it the smallest?

Solution
We compute the standard errors using the formula:

$$p = 0.8 : SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.8(0.2)}{100}} = 0.040$$
$$p = 0.5 : SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(0.5)}{100}} = 0.050$$
$$p = 0.3 : SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(0.7)}{100}} = 0.046$$
$$p = 0.1 : SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1(0.9)}{100}} = 0.030$$

The largest standard error is at a population proportion of 0.5 (which represents a population split 50-50 between being in the category we are interested in and not begin in). The farther we get from this 50-50 proportion, the smaller the standard error is. Of the four we computed, the smallest standard error is at a population proportion of 0.1.

Standard Error from a Formula and a Bootstrap Distribution In exercise 6.20, use Statkey or other technology to generate a bootstrap distribution of sample proportions and find the standard error for that distribution. Compare the result to the standard error given by the Central Limit Theorem, using the sample proportion as an estimate of the population proportion.

6.20 Proportion of home team wins in soccer, with $n = 120$ and $\hat{p} = 0.583$.

Solution
Using StatKey or other technology to create a bootstrap distribution, we see for one set of 1000 simulations that $SE = 0.045$. (Answers may vary slightly with other simulations.) Using the formula from the Central Limit Theorem, and using $\hat{p} = 0.583$ as an estimate for p, we have

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.583(1-.583)}{120}} = 0.045$$

We see that the bootstrap standard error and the formula match very closely.

6.38 Home Field Advantage in Baseball There were 2430 Major League Baseball (MLB) games played in 2009, and the home team won in 54.9% of the games. If we consider the games played in 2009 as a sample of all MLB games, find and interpret a 90% confidence interval for the proportion of games the home team wins in Major League Baseball.

Solution
To find a 90% confidence interval for p, the proportion of MLB games won by the home team, we

use $z^* = 1.645$ and $\hat{p} = 0.549$ from the sample of $n = 2430$ games. The confidence interval is

$$\begin{array}{rcl} \text{Sample statistic} & \pm & z^* \cdot SE \\ \hat{p} & \pm & z^*\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \\ 0.549 & \pm & 1.645\sqrt{\dfrac{0.549(0.451)}{2430}} \\ 0.549 & \pm & 0.017 \\ 0.532 & \text{to} & 0.566 \end{array}$$

We are 90% confident that the proportion of MLB games that are won by the home team is between 0.532 and 0.566. This statement assumes that the 2009 season is representative of all Major League Baseball games. If there is reason to assume that that season introduces bias, then we cannot be confident in our statement.

6.50 What Proportion Favor a Gun Control Law? A survey is planned to estimate the proportion of voters who support a proposed gun control law. The estimate should be within a margin of error of ±2% with 95% confidence, and we do not have any prior knowledge about the proportion who might support the law. How many people need to be included in the sample?

Solution
The margin of error we desire is $ME = 0.02$, and for 95% confidence we use $z^* = 1.96$. Since we have no prior knowledge about the proportion in support p, we use the conservative estimate of $\tilde{p} = 0.5$. We have:

$$\begin{array}{rcl} n & = & \left(\dfrac{z^*}{ME}\right)^2 \tilde{p}(1-\tilde{p}) \\ & = & \left(\dfrac{1.96}{0.02}\right)^2 0.5(1-0.5) \\ & = & 2401 \end{array}$$

We need to include 2, 401 people in the survey in order to get the margin of error down to within ±2%.

6.64 Home Field Advantage in Baseball There were 2430 Major League Baseball (MLB) games played in 2009, and the home team won the game in 54.9% of the games. If we consider the games played in 2009 as a sample of all MLB games, test to see if there is evidence, at the 1% level, that the home team wins more than half the games. Show all details of the test.

Solution
We are conducting a hypothesis test for a proportion p, where p is the proportion of all MLB games won by the home team. We are testing to see if there is evidence that $p > 0.5$, so we have

$$H_0 : p = 0.5$$

$$H_a : p > 0.5$$

This is a one-tail test since we are specifically testing to see if the proportion is greater than 0.5.

The test statistic is:

$$z = \frac{\text{Sample Statistic} - \text{Null parameter}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.549 - 0.5}{\sqrt{\frac{0.5(0.5)}{2430}}} = 4.83.$$

Using the normal distribution, we find a p-value of (to five decimal places) zero. This provides very strong evidence to reject H_0 and conclude that the home team wins more than half the games played. The home field advantage is real!

6.70 Percent of Smokers The data in **Nutrition Study**, introduced in Exercise 1.13 on page 13, include information on nutrition and health habits of a sample of 315 people. One of the variables is *Smoke*, indicating whether a person smokes or not (yes or no). Use technology to test whether the data provide evidence that the proportion of smokers is different from 20%.

Solution
We use technology to determine that the number of smokers in the sample is 43, so the sample proportion of smokers is $\hat{p} = 43/315 = 0.1365$. The hypotheses are:

$$H_0 : p = 0.20$$

$$H_a : p \neq 0.20$$

The test statistic is:

$$z = \frac{\text{Sample Statistic} - \text{Null Parameter}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1365 - 0.20}{\sqrt{\frac{0.2(0.8)}{325}}} = -2.82$$

This is a two-tail test, so the p-value is twice the area below -2.82 in a standard normal distribution. We see that the p-value is $2(0.0024) = 0.0048$. This small p-value leads us to reject H_0. We find strong evidence that the proportion of smokers is not 20%.

6.84 How Old is the US Population? From the US Census, we learn that the average age of all US residents is 36.78 years with a standard deviation of 22.58 years. Find the mean and standard deviation of the distribution of sample means for age if we take random samples of US residents of size:

(a) $n = 10$
(b) $n = 100$
(c) $n = 1000$

Solution
(a) The mean of the distribution is 36.78 years old. The standard deviation of the distribution of sample means is the standard error:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{22.58}{\sqrt{10}} = 7.14$$

(b) The mean of the distribution is 36.78 years old. The standard deviation of the distribution of sample means is the standard error:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{22.58}{\sqrt{100}} = 2.258$$

(c) The mean of the distribution is 36.78 years old. The standard deviation of the distribution of sample means is the standard error:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{22.58}{\sqrt{1000}} = 0.714$$

Notice that as the sample size goes up, the standard error of the sample means goes down.

Standard Error from a Formula and a Bootstrap Distribution In Exercises 6.96 to 6.99, use *StatKey* or other technology to generate a bootstrap distribution of sample means and find the standard error for that distribution. Compare the result to the standard error given by the Central Limit Theorem, using the sample standard deviation as an estimate of the population standard deviation.

6.97 Mean commute time in Atlanta, in minutes, using the data in CommuteAtlanta with $n = 500, \bar{x} = 29.11$, and $s = 20.72$.

Solution
Using StatKey or other technology to create a bootstrap distribution, we see for one set of 1000 simulations that $SE \approx 0.92$. (Answers may vary slightly with other simulations.) Using the formula from the Central Limit Theorem, and using $s = 20.72$ as an estimate for σ, we have

$$SE = \frac{s}{\sqrt{n}} = \frac{11.11}{\sqrt{25}} = 2.22.$$

We see that the bootstrap standard error and the formula match very closely.

6.120 Bright Light at Night Makes Even Fatter Mice Data A.1 on page 136 introduces a study in which mice that had a light on at night (rather than complete darkness) ate most of their calories when they should have been resting. These mice gained a significant amount of weight, despite eating the same number of calories as mice kept in total darkness. The time of eating seemed to have a significant effect. Exercise 6.119 examines the mice with dim light at night. A second group of mice had bright light on all the time (day and night). There were nine mice in the group with bright light at night and they gained an average of 11.0g with a standard deviation of 2.6. The data are shown in the figure in the book. Is it appropriate to use a t-distribution in this situation? Why or why not? If not, how else might we construct a confidence interval for mean weight gain of mice with a bright light on all the time?

Solution
The sample size of $n = 9$ is quite small, so we require a condition of approximate normality for the underlying population in order to use the t-distribution. In the dotplot of the data, it appears that the data might be right skewed and there is quite a large outlier. It is probably more reasonable to use other methods, such as a bootstrap distribution, to compute a confidence interval using this data.

6.130 Find the sample size needed to give, with 95% confidence, a margin of error within ± 10. Within ± 5. Within ± 1. Assume that we use $\tilde{\sigma} = 30$ as our estimate of the standard deviation in each case. Comment on the relationship between the sample size and the margin of error.

Solution

We use $z^* = 1.96$ for 95% confidence, and we use $\tilde{\sigma} = 30$. For a desired margin of error of $ME = 10$, we have:

$$n = \left(\frac{z^* \cdot \tilde{\sigma}}{ME}\right)^2 = \left(\frac{1.96 \cdot 30}{10}\right)^2 = 34.6$$

We round up to $n = 35$.

For a desired margin of error of $ME = 5$, we have:

$$n = \left(\frac{z^* \cdot \tilde{\sigma}}{ME}\right)^2 = \left(\frac{1.96 \cdot 30}{5}\right)^2 = 138.3$$

We round up to $n = 139$.

For a desired margin of error of $ME = 1$, we have:

$$n = \left(\frac{z^* \cdot \tilde{\sigma}}{ME}\right)^2 = \left(\frac{1.96 \cdot 30}{1}\right)^2 = 3457.4$$

We round up to $n = 3,458$.

We see that the sample size goes up as we require more accuracy. Or, put another way, a larger sample size gives greater accuracy.

6.145 The Chips Ahoy! Challenge in the mid-1990s a Nabisco marketing campaign claimed that there were at least 1000 chips in every bag of Chips Ahoy! cookies. A group of Air Force cadets collected a sample of 42 bags of Chips Ahoy! cookies, bought from locations all across the country to verify this claim. The cookies were dissolved in water and the number of chips (any piece of chocolate) in each bag were hand counted by the cadets. The average number of chips per bag was 1261.6, with standard deviation 117.6 chips.

(a) Why were the cookies bought from locations all over the country?

(b) Test whether the average number of chips per bag is greater than 1000. Show all details.

(c) Does part (b) confirm Nabisco's claim that every bag has at least 1000 chips? Why or why not?

Solution

(a) The cookies were bought from locations all over the country to try to avoid sampling bias.

(b) Let μ be the mean number of chips per bag. We are testing $H_0 : \mu = 1000$ vs $H_a : \mu > 1000$. The test statistic is

$$t = \frac{1261.6 - 1000}{117.6/\sqrt{42}} = 14.4$$

We use a t-distribution with 41 degrees of freedom. The area to the left of 14.4 is negligible, and p-value ≈ 0. We conclude, with very strong evidence, that the average number of chips per bag of Chips Ahoy! cookies is greater than 1000.

(c) No! The test in part (b) gives convincing evidence that the average number of chips per bag is greater than 1000. However, this does not necessarily imply that every individual bag has more than 1000 chips.

6.150 Are Florida Lakes Acidic or Alkaline? The pH of a liquid is a measure of its acidity or

alkalinity. Pure water has a pH of 7, which is neutral. Solutions with a pH less than 7 are acidic while solutions with a pH greater than 7 are basic or alkaline. The dataset **FloridaLakes** gives information, including pH values, for a sample of lakes in Florida. Computer output of descriptive statistics for the pH variable is shown in the book.

(a) How many lakes are included in the dataset? What is the mean pH value? What is the standard devotion?

(b) Use the descriptive statistics above to conduct a hypothesis test to determine whether there is evidence that average pH in Florida lakes is different from the neutral value of 7. Show all details of the test and use a 5% significance level. If there is evidence that it is not neutral, does the mean appear to be more acidic or more alkaline?

(c) Compare the test statistic and p-value found in part (b) to the computer output shown in the book for the same data.

Solution

(a) We see that $n = 53$ with $\bar{x} = 6.591$ and $s = 1.288$.

(b) The hypotheses are:

$$H_0 : \mu = 7$$

$$H_a : \mu \neq 7$$

?where μ represents the mean pH level of all Florida lakes. We calculate the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{6.591 - 7}{1.288/\sqrt{53}} = -2.31$$

We use a t-distribution with 52 degrees of freedom to see that the area below ?2.31 is 0.0124. Since this is a two-tail test, the p-value is $2(0.0124) = 0.0248$. We reject the null hypothesis at a 5% significance level, and conclude that average pH of Florida lakes is different from the neutral value of 7. Florida lakes are, in general, somewhat more acidic than neutral.

(c) The test statistic matches the computer output exactly and the p-value is the same up to rounding off.

Computer Exercises

For each R problem, turn in answers to questions with the written portion of the homework. Send the R code for the problem to Katherine Goode. The answers to questions in the written part should be well written, clear, and organized. The R code should be commented and well formatted.

R problem 1 Ideally, a 95% confidence interval will be as tightly clustered around the true value as possible, and will have a 95% coverage probability. When the possible data values are discrete, (such as in the case of sample proportions which can only be a count over the sample size), the true coverage or capture probability is not exactly 0.95 for every p. This problem examines the true coverage probability for three different methods of making confidence intervals.

To compute the coverage probability of a method, recognize that each possible value x from 0 to n for a given method results in a confidence interval with a lower bound $a(x)$ and an upper bound $b(x)$. The interval will capture p if $a(x) < p < b(x)$. To compute the capture probability of

a given p, we need to add up all of the binomial probabilities for the x values that capture p in the interval. For a sample size n and true population proportion p, this coverage probability is

$$P(p \text{ in interval }) = \sum_{x:a(x)<p<b(x)} \binom{n}{x} p^x (1-p)^{n-x}$$

To compute this in R, you need to find the lower and upper bounds of the confidence interval for each possible outcome x, and add the probabilities of the outcomes that capture p. Here is some code to get you started using the textbook method for an example where $n = 10$ and $p = 0.3$.

```
x = 0:10
p.hat = x/10 # will be a vector
se = sqrt(p.hat*(1-p.hat)/10) # also a vector
z = qnorm(0.975)
a = p.hat - z*se # also a vector
b = p.hat + z*se #also a vector
x[ (a < 0.3) & (0.3 < b) ] # x that capture p
## [1] 2 3 4 5 6
```

For each of the following methods, find which outcomes x result in confidence intervals that capture p and compute the coverage probability from a sample of size $n = 60$ when $p = 0.4$.

1. Normal from maximum likelihood estimate, $\hat{p} = X/n, SE = \sqrt{\hat{p}(1-\hat{p})/n}$, with the interval

   $$\hat{p} \pm 1.96SE$$

   Solution
   Since $n = 60$, it is possible for x to range from 0 to 60. Thus, we calculate all possible values of $\hat{p}$, which are

   $$\left\{ \frac{0}{60}, \frac{1}{60}, ..., \frac{60}{60} \right\}.$$

   We then calculate the standard errors associated with each $\hat{p}$ using the formula shown above and R. With the standard error, we are able to calculate both the lower and upper bounds for the confidence intervals using the formula $\hat{p} \pm 1.96SE$. Several of the values we calculated are shown in the table below.

X	$\hat{p}$	SE	Lower Bound	Upper Bound
0	$\frac{0}{60}$	0	0	0
1	$\frac{1}{60}$	0.0165	-0.0157	0.0491
2	$\frac{2}{60}$	0.0232	-0.0121	0.0788
3	$\frac{3}{60}$	0.0281	-0.0051	0.1051
⋮	⋮	⋮	⋮	⋮
59	$\frac{59}{60}$	0.0165	0.9509	1.0157
60	$\frac{60}{60}$	0	1	1

   Now, we need to determine for which values of X, 0.4 is captured by X's confidence interval. Using R, we find that this is the case when

   $$X \in \{18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31\}.$$

Thus, we can now compute the coverage probability as shown below.

$$P(0.4 \text{ in interval }) = \sum_{x=18}^{31} \binom{60}{x} (0.4)^x (1-0.4)^{n-x} = 0.9337$$

We find that the coverage probability in this case is in fact a bit less than 95%.

The following R code was used to complete this problem.

```
x.1 = 0:60
p.hat.1 = x.1/60
se.1 = sqrt(p.hat.1*(1-p.hat.1)/60)
z = qnorm(0.975)
a.1 = p.hat.1 - z*se.1
b.1 = p.hat.1 + z*se.1
x[ (a.1 < 0.4) & (0.4 < b.1) ]
sum(dbinom(18:31,60,0.4))
```

2. Normal from adjusted maximum likelihood estimate, $\tilde{p} = (X+2)/(n+4), SE = \sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$, with the interval

$$\tilde{p} \pm 1.96SE$$

Solution

We perform the same process again, but this time we use the new equations presented for calculating $\tilde{p}, SE$, and the confidence intervals. The table below shows some of the values that we calculate using R.

X	$\tilde{p}$	SE	Lower Bound	Upper Bound
0	$\frac{0+2}{60+4} = \frac{2}{64}$	0.0217	-0.0114	0.0739
1	$\frac{1+2}{60+4} = \frac{3}{64}$	0.0264	-0.0049	0.0987
2	$\frac{2+2}{60+4} = \frac{4}{64}$	0.0303	0.0032	0.1218
3	$\frac{3+2}{60+4} = \frac{5}{64}$	0.0335	0.0124	0.1439
⋮	⋮	⋮	⋮	⋮
59	$\frac{59+2}{60+4} = \frac{61}{64}$	0.026	0.9013	1.0049
60	$\frac{60+2}{60+4} = \frac{62}{64}$	0.0217	0.9261	1.0114

We use R to determine that 0.4 is captured by the confidence intervals when

$$X \in \{17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31\}.$$

Hence, the coverage probability is

$$P(0.4 \text{ in interval }) = \sum_{x=17}^{31} \binom{60}{x} (0.4)^x (1-0.4)^{n-x} = 0.9529$$

We find that this method for calculating confidence intervals gives us a coverage probability that is much closer to 95% than the method in part 1.

The following R code was used to complete this problem.

```
x.2 = 0:60
p.tilde.2 = (x.2+2)/(60+4)
se.2 = sqrt(p.tilde.2*(1-p.tilde.2)/(60+4))
z = qnorm(0.975)
a.2 = p.tilde.2 - z*se.2
b.2 = p.tilde.2 + z*se.2
x[ (a.2 < 0.4) & (0.4 < b.2) ]
sum(dbinom(17:31,60,0.4))
```

3. Within $z^2/2$ of the maximum likelihood loglikelihood. For this method, the file `logl.R` has a function `logl.ci.p()` which returns the lower and upper bounds of a 95% confidence interval given n and x. You can graph the loglikelihood using `glogl.p()` for n, x, and $z = 1.96$ to see if the returned values make sense.

   Solution
   Using R and the code presented by the professor, we calculate a confidence interval based on the maximum loglikelihood for each value of X between 0 and 60. Some of the confidence intervals are shown below.

X	Lower Bound	Upper Bound
0	0	0.0315
1	0.0010	0.0713
2	0.0056	0.0994
3	0.0127	0.1246
4	0.0212	0.1481
⋮	⋮	⋮
59	0.9287	0.9990
60	0.9685	60

   We use R to determine that 0.4 is captured by the confidence intervals when

$$X \in \{17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31\}.$$

   Thus, once again, the capture probability is

$$P(0.4 \text{ in interval }) = \sum_{x=17}^{31} \binom{60}{x} (0.4)^x (1-0.4)^{n-x} = 0.9529.$$

   The following R code was used to complete this problem.

```
CIs <- matrix(,nrow=61,ncol=2)
for(i in 1:61)
{
   CIs[i,1]<-logl.ci.p(60,i-1,conf=0.95)[1]
   CIs[i,2]<-logl.ci.p(60,i-1,conf=0.95)[2]
}
x.3 <- 0:60
x.3[ (CIs[,1] < 0.4) & (0.4 < CIs[,2]) ]
```

R Problem 2 Repeat the previous problem, but for $n = 60$ and $p = 0.1$.

Solution
We go through the same process that we did for problem 1.

For the normal maximum likelihood estimate, we determine that 0.1 is captured by a confidence interval when

$$X \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

Thus, the capture probability is

$$P(0.1 \text{ in interval }) = \sum_{x=3}^{12} \binom{60}{x} (0.1)^x (1 - 0.1)^{n-x} = 0.9413$$

The following R code was used to complete this problem.

```
x.1 = 0:60
p.hat.1 = x.1/60
se.1 = sqrt(p.hat.1*(1-p.hat.1)/60)
z = qnorm(0.975)
a.1 = p.hat.1 - z*se.1
b.1 = p.hat.1 + z*se.1
x.1[ (a.1 < 0.1) & (0.1 < b.1) ]
sum(dbinom(3:12,60,0.1))
```

For the normal from adjusted maximum likelihood estimate, we determine that 0.1 is captured by a confidence interval when

$$X \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Thus, the capture probability is

$$P(0.1 \text{ in interval }) = \sum_{x=2}^{10} \binom{60}{x} (0.1)^x (1 - 0.1)^{n-x} = 0.9520$$

The following R code was used to complete this problem.

```
x.2 = 0:60
p.tilde.2 = (x.2+2)/(60+4)
se.2 = sqrt(p.tilde.2*(1-p.tilde.2)/(60+4))
z = qnorm(0.975)
a.2 = p.tilde.2 - z*se.2
b.2 = p.tilde.2 + z*se.2
x.2[ (a.2 < 0.1) & (0.1 < b.2) ]
sum(dbinom(2:10,60,0.1))
```

For the confidence intervals derived from the maximum likelihood loglikelihood, we determine that 0.1 is captured by a confidence interval when

$$X \in \{3, 4, 5, 6, 7, 8, 9, 10, 11\}.$$

Thus, the capture probability is

$$P(0.1 \text{ in interval }) = \sum_{x=3}^{11} \binom{60}{x} (0.1)^x (1-0.1)^{n-x} = 0.9324$$

The following R code was used to complete this problem.

```
CIs <- matrix(,nrow=61,ncol=2)
for(i in 1:61)
{
   CIs[i,1]<-logl.ci.p(60,i-1,conf=0.95)[1]
   CIs[i,2]<-logl.ci.p(60,i-1,conf=0.95)[2]
}
x.3 <- 0:60
x.3[ (CIs[,1] < 0.1) & (0.1 < CIs[,2]) ]
sum(dbinom(3:11,60,0.1))
```

R Problem 3 This problem examines a t distribution with 4 degrees of freedom.

Here is some sample code to draw graphs of continuous distributions.

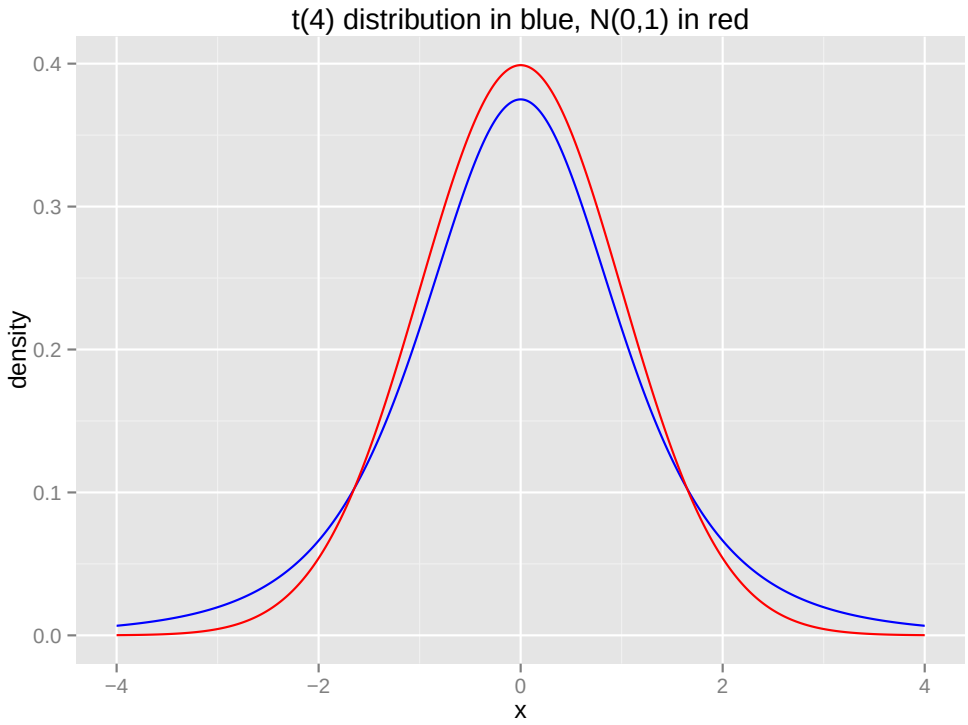
```
x = seq(-4,4,0.001)
z = dnorm(x)
y.10 = dt(x, df=10)
d = data.frame(x,z,y.10)
require(ggplot2)
## Loading required package:  ggplot2
ggplot(d) +
geom_line(aes(x=x,y=y.10),color="blue") +
geom_line(aes(x=x,y=z),color="red") +
ylab('density') +
ggtitle("t(10) distribution in blue, N(0,1) in red")
```

1. Draw a graph of a t distribution with 4 degrees of freedom and a standard normal curve from -4 to 4.

   Solution
   Below is the graph that was drawn in R.

   This is the R code used to create the above graph.

```
x = seq(-4,4,0.001)
z = dnorm(x,0,1)
y.10 = dt(x, df=4)
d = data.frame(x,z,y.10)
require(ggplot2)
ggplot(d) +
   geom_line(aes(x=x,y=y.10),color="blue")+
  geom_line(aes(x=x,y=z),color="red")+
   ylab('density')+
   ggtitle("t(4) distribution in blue, N(0,1) in red")
```



2. Find the area to the right of 2 under each curve.

   Solution
   The area to the right of 2 under the t distribution curve is as follows.

   $$P(t > 2) = 0.0581$$

   The area to the right of 2 under the standard normal distribution curve is as follows.

   $$P(z > 2) = 0.0228$$

   The following is the R code used to achieve these answers.

   ```
   1-pt(2,4)
   1-pnorm(2,0,1)
   ```

3. Find the 0.975 quantile of each curve.

   Solution
   The 0.975 quantile for the t distribution is 2.7764.
   The 0.975 quantile for the standard normal distribution is 1.9600.

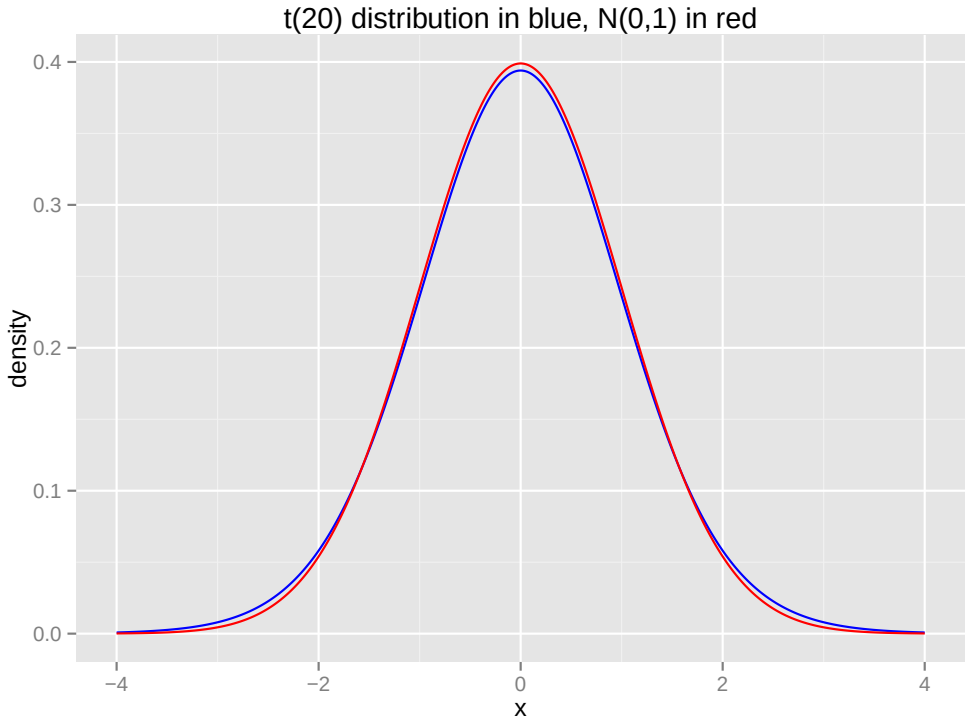
   The following is the R code used to achieve these answers.

   ```
   qt(0.975,4)
   qnorm(0.975,0,1)
   ```

R Problem 4 Repeat the previous problem, but for a t distribution with 20 degrees of freedom.

1. Draw a graph of a t distribution with 20 degrees of freedom and a standard normal curve from -4 to 4.

   Solution
   Below is the graph that was drawn in R.



This is the R code used to create the above graph.

```
x = seq(-4,4,0.001)
z = dnorm(x,0,1)
y.10 = dt(x, df=20)
d = data.frame(x,z,y.10)
require(ggplot2)
ggplot(d) +
   geom_line(aes(x=x,y=y.10),color="blue")+
  geom_line(aes(x=x,y=z),color="red")+
   ylab('density')+
   ggtitle("t(20) distribution in blue, N(0,1) in red")
```

2. Find the area to the right of 2 under each curve.

   Solution
   The area to the right of 2 under the t distribution curve is as follows.

$$P(t > 2) = 0.0296$$

   The area to the right of 2 under the standard normal distribution curve is as follows.

$$P(z > 2) = 0.0228$$

   The following is the R code used to achieve these answers.

```
1-pt(2,20)
1-pnorm(2,0,1)
```

3. Find the 0.975 quantile of each curve.

   Solution
   The 0.975 quantile for the t distribution is 2.0860.
   The 0.975 quantile for the standard normal distribution is 1.9600.

   The following is the R code used to achieve these answers.

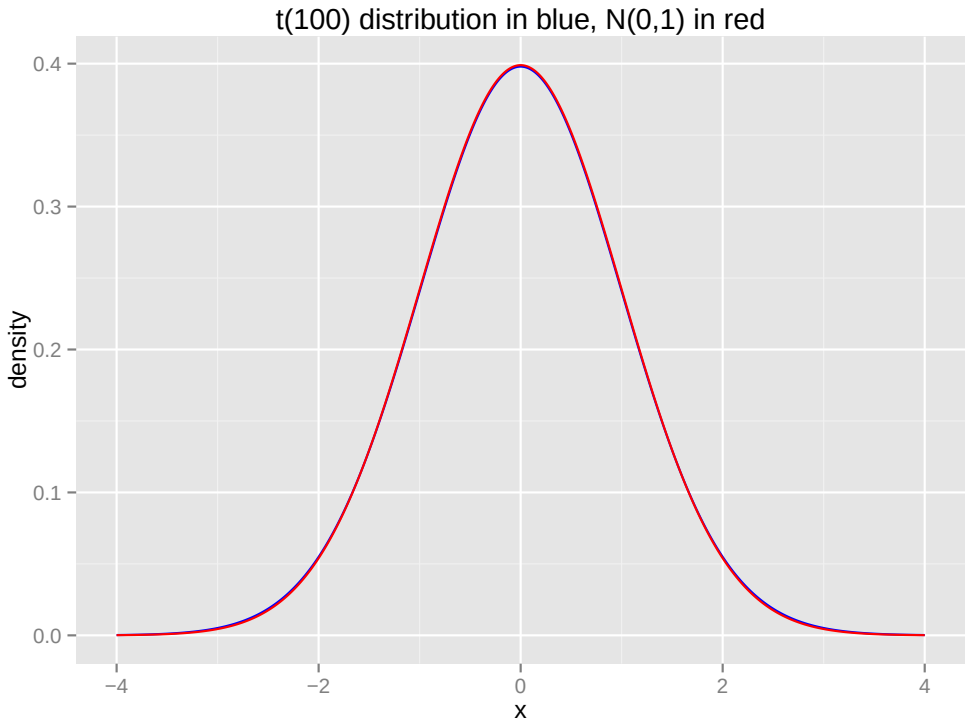
```
qt(0.975,20)
qnorm(0.975,0,1)
```

R Problem 5 Repeat the previous problem, but for a t distribution with 100 degrees of freedom.

1. Draw a graph of a t distribution with 100 degrees of freedom and a standard normal curve from -4 to 4.

   Solution
   Below is the graph that was drawn in R.

   This is the R code used to create the above graph.




```
x = seq(-4,4,0.001)
z = dnorm(x,0,1)
y.10 = dt(x, df=100)
d = data.frame(x,z,y.10)
require(ggplot2)
ggplot(d) +
   geom_line(aes(x=x,y=y.10),color="blue")+
  geom_line(aes(x=x,y=z),color="red")+
   ylab('density')+
   ggtitle("t(100) distribution in blue, N(0,1) in red")
```

2. Find the area to the right of 2 under each curve.

   Solution
   The area to the right of 2 under the t distribution curve is as follows.

$$P(t > 2) = 0.0241$$

   The area to the right of 2 under the standard normal distribution curve is as follows.

$$P(z > 2) = 0.0228$$

   The following is the R code used to achieve these answers.

```
1-pt(2,100)
1-pnorm(2,0,1)
```

3. Find the 0.975 quantile of each curve.

   Solution
   The 0.975 quantile for the t distribution is 1.9840.
   The 0.975 quantile for the standard normal distribution is 1.9600.

   The following is the R code used to achieve these answers.

```
qt(0.975,100)
qnorm(0.975,0,1)
```