Solutions to Homework 6

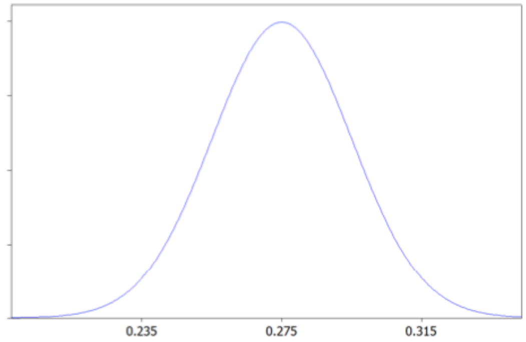
Statistics 302 Professor Larget

Textbook Exercises

5.29 (Graded for Completeness) What Proportion Have College Degrees? According to the US Census Bureau, about 27.5% of US adults over the age of 25 have a bachelor's level (or higher) college degree. For random samples of $n = 500$ US adults over the age of 25, the sample proportions, $\hat{p}$, with at least a bachelor's degree follow a normal distribution with mean 0.275 and standard deviation 0.02. Draw a sketch of this normal distribution and label at least three points on the horizontal axis.

<u>Solution</u>

A $N(0.275, 0.02)$ curve is centered at its mean, 0.275. The values two standard deviations away, 0.235 and 0.315, are labeled so that approximately 95% of the area falls between these two values. They should be out in the tails, with only about 2.5% of the distribution beyond them on each side. See the figure.



5.32 (Graded for Completeness) Critical Reading on the SAT Exam In the table from Exercise 5.30, we see that scores on the Critical Reading portion of the SAT (Scholastic Aptitude Test) exam are normally distributed with mean 501 and standard deviation 112. Use the normal distribution to answer the following questions:

(a) What is the estimated percentile for a student who scores 690 on Critical Reading?

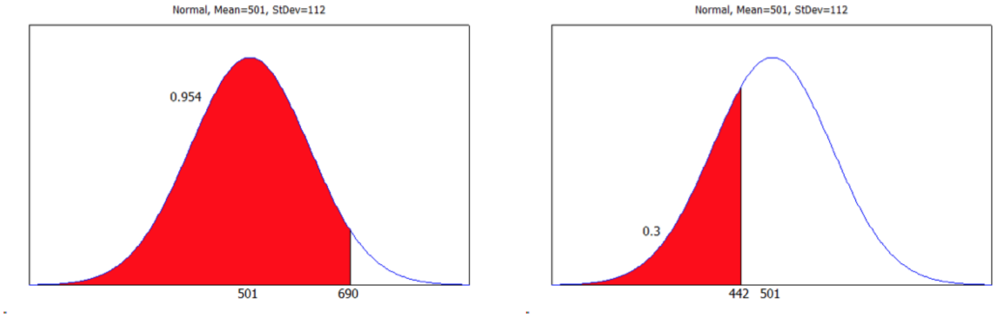
(b) What is the approximate score for a student who is at the 30th percentile for Critical Reading?

<u>Solution</u>

The plots below show the required endpoint(s) and/or probabilities for the given normal distributions. Note that a percentile always means the area to the left. Using technology, we can find the endpoints and areas directly, and we obtain the answers below. (Alternately, we could convert to a standard normal and use the standard normal to find the equivalent area.)

(a) The area below 690 is 0.95, so that point is the 95th percentile of a $N(501, 112)$ distribution.

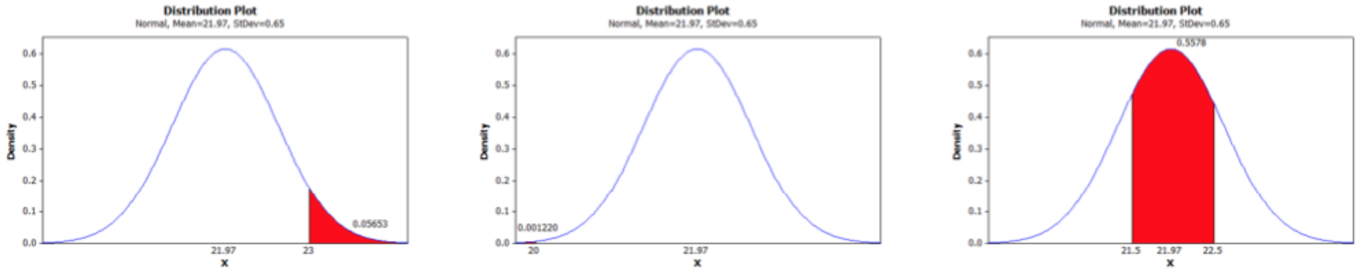
(b) The point where 30% of the scores are below it is a score of 442.



(a) 690 on Critical Reading (b) 30th percentile Critical Reading

5.36 (Graded for Accurateness) Commuting TImes in St. Louis A bootstrap distribution of mean commute times (in minutes) based on a sample of 500 St. Louis workers stored in **CommuteStLouis** is shown in the book. The pattern in this dot plot is reasonably bell-shaped so we use a normal curve to model this distribution of bootstrap means. The mean for this distribution is 21.97 minutes and the standard deviation is 0.65 minutes. Based on this normal distribution, what proportion of bootstrap means should be in each of the following regions?
(a) More than 23 mintues
(b) Less than 20 minutes
(c) Between 21.5 and 22.5 minutes

Solution
The plots below show the three required regions as areas in a $N(21.97, 0.65)$ distribution. We see that the areas are 0.0565, 0.0012, and 0.5578, respectively.



(a) more than 23 (b) less than 20 (c) between 21.5 and 22.5

If converting to a standard normal, the relevant z-scores and areas are shown below.

(a) $z = \frac{23-21.97}{0.65} = 1.585$. The area above 1.585 for $N(0, 1)$ is 0.0565.

(b) $z = \frac{20-21.97}{0.65} = -3.031$. The area below -3.031 for $N(0, 1)$ is 0.0012.

(c) $z = \frac{21.5-21.97}{0.65} = -0.7231$ and $z = \frac{22.5-21.97}{0.65} = 0.8154$. The area between -0.7231 and 0.8154 for $N(0, 1)$ is 0.5578.

5.62 (Graded for Accurateness) To Study Effectively, Test Yourself! Cognitive science consistently shows that one of the most effective studying tools is to self-test. A recent study reinforced this finding. In the study, 118 college student studied 48 pairs of Swahili and English words. All students had an initial study time and then three blocks of practice time. During the practice time, half the students studied the words by reading them side by side, while the other half gave themselves quizzes in which they were shown one word and had to recall it partner. Students were randomly assigned to the two groups, and total practice times was the same for both groups. On the final test one week later, the proportion of items correctly recalled was 15% for the reading-study group and 42% for the self-quiz group. The standard error for the difference in proportions is about 0.07. Test whether giving self-quizzes is more effective and show all details of the test. The sample size is large enough to use the normal distribution.

Solution

The relevant hypotheses are $H_0 : p_Q = p_R$ vs $H_a : p_Q > p_R$, where p_Q and p_R are the proportions of words recalled correctly after quiz studying or studying by reading alone, respectively. Based on the sample information the statistic of interest is

$$\hat{p}_Q - \hat{p}_R = 0.42 - 0.15 = 0.27$$

The standard error of this statistic is given as $SE = 0.07$ and the null hypothesis is that the difference in the proportions for the two group is zero. We compute the standardized test statistic with

$$z = \frac{SampleStatistic - NullParameter}{SE} = \frac{0.27 - 0}{0.07} = 3.86$$

Using technology, the area under a $N(0, 1)$ curve beyond $z = 3.86$ is only 0.000056. This very small p-value provides very strong evidence that the proportion of words recalled using self-quizzes is more than the proportion recalled with reading study alone.

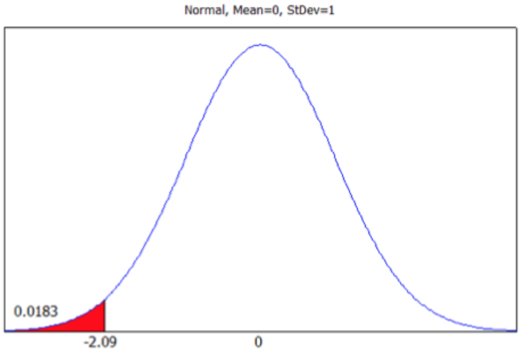
5.64 (Graded for Completeness) Penalty Shots in World Cup Soccer A study of 138 penalty shots in World Cup Finals games between 1982 and 1994 found that the goalkeeper correctly guessed the direction of the kick only 41% of the time. The article notes that this is "slightly worse than random chance". We use these data as a sample of all World Cup penalty shots ever. Test at a 5% significance level to see whether there is evidence that the percent guessed correctly is less than 50%. The sample size is large enough to use the normal distribution. The standard error from a randomization distribution under the null hypothesis is SE=0.043.

Solution

We test $H_0 : p = 0.5$ vs $H_a : p < 0.5$ where p is the proportion of all World Cup penalty shots for which the goalkeeper guesses the correct direction. The statistic from the original sample is $\hat{p} = 0.41$ and the null parameter is $p = 0.5$. The standard error is $SE = 0.043$. We use this information to find the standardized test statistic:

$$\begin{aligned} z &= \frac{Samplestatistic - Nullparameter}{SE} \\ &= \frac{0.41 - 0.5}{0.043} \\ &= -2.09. \end{aligned}$$

This is a lower tail test, so we find the area below $z = -2.09$ in the lower tail of the standard normal distribution. We see in the figure below that this area is 0.0183. The p-value for this test is 0.0183. At a 5% significance level, we find evidence that the proportion of World Cup penalty shots in which the goalkeeper guesses correctly is significantly less than half.



5.75 (Graded for Completeness) Malevolent Uniforms in Football The figure in the book shows a bootstrap distribution of correlations between penally yards and uniform malevolence using the data on 28 NFL teams in **MalevolentUniformsNFL**. We wee from the percentiles of the bootstrap distribution that a 99% confidence interval for the correlation is -0.224 to 0.788. The correlation between the two variables for the original sample is $r = 0.37$.
(a) Use the original sample correlation and the standard deviation of the bootstrap distribution shown in the figure to compute a 99% confidence interval for the correlation using z^* from a normal distribution.
(b) Why is the normal-based interval somewhat different front he percentile interval? Hint: Look at the shape of the bootstrap distribution.

Solution
(a) For a 99% confidence interval the standard normal value leaving 0.5% in each tail is $z^* = 2.576$. From the bootstrap distribution we estimate the standard error of the correlations to be 0.205. Thus the 99% confidence interval based on a normal distribution would be

$$0.37 \pm 2.576 \times 0.205 = 0.37 \pm 0.528 = (-0.158, 0.898)$$

(b) The bootstrap distribution of correlations is somewhat right skewed, while the normal-based interval assumes the distribution is symmetric and bell-shaped.

Disjoint, Independent, and Complement For Exercise 11.30, state whether the two events (A and B) described are disjoint, independent, and/or complements. (It is possible that the two events fall into more than one of the three categories, or none of them.)

11.30 (Graded for Accurateness) Roll two (six-sided) dice. Let A be the event that the first die is a 3 and B be the event that the sum of the two dice is 8.

Solution
The two events are not disjoint or complements, as it is possible to have the rolls be $\{3, 5\}$ where the first die is a 3 and the sum is 8. To check independence we need to find

$$P(A) = 1/6 \text{ and } P(B) = P(\{2, 6\} \text{ or } \{3, 5\} \text{ or } \{4, 4\} \text{ or } \{5, 3\} \text{ or } \{6, 3\}) = 5/36$$

There is only one possibility for the intersection so $P(A \text{ and } B) = P(\{3,5\}) = 1/36$. We then check that

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{1/36}{1/6} = 1/6 \neq P(B) = 5/36$$

so A and B are not independent. We can also verify that

$$P(A \text{ and } B) = 1/36 \neq P(A) \times P(B) = (1/6) \times (5/36) = 5/216$$

11.36 (Graded for Accurateness) Peanut M & Ms In a bag of peanut M & M's, there are 80 M & Ms with 11 red ones, 12 orange ones, 20 blue ones, 11 green ones, 18 yellow ones, and 8 brown ones. They are mixed up so that each candy piece is equally likely to be selected if we pick one.
(a) If we select one at random, what is the probability that it is red?
(b) If we select one at random, what is the probability that it is not blue?
(c) If we select one at random, what is the probability that is is red or orange?
(d) If we select one at random, then put it back, mix them up well (so the selections are independent) and select another one, what is the probability that both the first and second ones are blue?
(e) If we select one, keep it, and then select a second one, what is the probability that the first one is red and the second one is green.

Solution
(a) There are 11 red ones out of a total of 80, so the probability that we pick a red one is $11/80 = 0.1375$.

(b) The probability that it is blue is $20/80 = 0.25$ so the probability that it is not blue is $1 - 0.25 = 0.75$.

(c) The single piece can be red or orange, but not both, so these are disjoint events. The probability the randomly selected candy is red or orange is $11/80 + 12/80 = 23/80 = 0.2875$.

(d) The probability that the first one is blue is $20/80 = 0.25$. When we put it back and mix them up, the probability that the next one is blue is also 0.25. By the multiplication rule, since the two selections are independent, the probability both selections are blue is $0.25 \times 0.25 = 0.0625$.

(e) The probability that the first one is red is $11/80$. Once that one is taken (since we don't put it back and we eat it instead), there are only 79 pieces left and 11 of those are green. By the multiplication rule, the probability of a red then a green is $(11/80) \times (11/79) = 0.0191$.

11.39 (Graded for Completeness) Color Blindness in Men and Women The most common form of color blindness is an inability to distinguish red from green. However, this particular form of color blindness is much more common in men than in women (this is because the genes corresponding to the red and green receptors are located on the X-chromosomes). Approximately 7% of American men and 0.4% of American women are red-green color-blind.
(a) If an American male is selected at random, what is the probability that he is red-green color-blind?
(b) If an American female is selected at random, what is the probability that she is NOT red-green color-blind?
(c) If one man and one woman are selected at random, what is the probability that neither are

red-green color-blind?
(d) If one man and one woman are selected at random, what is the probability that at least one of them is red-green color-blind?

<u>Solution</u>
Let CBM and CBW denote the events that a man or a woman is colorblind, respectively.

(a) As 7% of men are colorblind, P(CBM) = 0.07.

(b) As 0.4% of women are colorblind, P(not CBW) = 1 - P(CBW) = 1 - 0.004 = 0.996.

(c) The probability the woman is not colorblind is 0.996, and the probability that the man is not color- blind is 1 - 0.07 = 0.93. As the man and woman are selected independently, we can multiply their probabilities:

P(Neither is Colorblind) = P(not CBM) × P(not CBW) = 0.93 × 0.996 = 0.926.

(d) The event that "At least one is colorblind" is the complement of part (d) that "Neither is Colorblind" so we have

P (At least one is Colorblind) = 1 - P (Neither is Colorblind) = 1 - 0.926 = 0.074

We could also do this part as

P(CBM or CBW) = P(CBM)+P(CBW)-P(CBM and CBW) = 0.07+0.004-(0.07)(0.004) = 0.074

11.58 (Graded for Accurateness) Mammograms and Breast Cancer The mammogram is helpful for detecting breast cancer in its early stages. However, it is an imperfect diagnostic tool. According to one study, 86.6 of every 1000 women between the ages of 50 and 59 that do not have cancer are wrongly diagnosed (a "false" positive), while 1.1 of every 1000 women between the ages of 50 and 59 that do have cancer are not diagnosed (a "false" positive). One in 38 women between the ages of 50 and 59 will develop breast cancer. If a woman between the ages of 50 and 59 has a positive mammogram, what is the probability that she will have breast cancer?

<u>Solution</u>
We are given

$$P(\text{Positive if no Cancer}) = 86.6/1000 = 0.0866,$$
$$P(\text{Positive if Cancer}) = 1 - 1.1/1000 = 0.9989, \text{ and}$$
$$P(Cancer) = 1/38 = 0.0263.$$

Applying Bayes rule we have

$$\begin{aligned} P(\text{Cancer if Positive}) &= \frac{P(\text{Cancer})P(\text{Positive if Cancer})}{P(\text{no Cancer})P(\text{Positive if no Cancer}) + P(\text{Cancer})P(\text{Positive if Cancer})} \\ &= \frac{(0.0263)(0.9989)}{(1-0.0263)(0.0866)+(0.0263)(0.9989)} \\ &= 0.2375 \end{aligned}$$

Identifying Spam Text Messages Bayes' rule can be used to identify and filter spam emails and text messages. Exercise 11.60 refers to a large collation of real SMS text messages from participating cellphone users. In this collection, 747 of the 5574 total messages (13.40%) are identified as spam.

11.60 (Graded for Completeness) The word “free” is contained in 4.75% of all messages, and 3.57% of all messages both contain the word “free” and are marked as spam.
(a) What is the probability that a message contains the word “free”, given that it is spam?
(b) What is the probability that a message is spam given that it contains the word “free”?

Solution
(a) Using the formula for conditional probability,

$$P(\text{Free if Spam}) = \frac{P(\text{Free and Spam})}{P(\text{Spam})} = \frac{0.0357}{0.134} = 0.266$$

(b) Using the formula for conditional probability,

$$P(\text{Spam if Free}) = \frac{P(\text{Free and Spam})}{P(\text{Free})} = \frac{0.0357}{0.0475} = 0.752$$

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 (Graded for Accurateness) The function `pnorm()` finds probabilities from normal distributions. By default, it returns the area to the left from the standard normal density, but the second and third arguments can be used to specify a different mean or standard deviation. So, here are various ways to calculate the area to the right of 650 from a $N(500, 100)$ distribution.

```
1 - pnorm(650, 500, 100)
## [1] 0.06681
1 - pnorm(650, mean = 500, sd = 100)
## [1] 0.06681
1 - pnorm((650 - 500)/100)
## [1] 0.06681
```

The function `qnorm()` finds quantiles from a normal distribution. Again, without other arguments, it uses the standard normal distribution.

```
qnorm(0.9)
## [1] 1.282
qnorm(0.9, mean = 500, sd = 100)
## [1] 628.2
```

Write an expression using `pnorm()` and or `qnorm()` to find each of the following values. Note, we are using the notation $N(\mu, \sigma)$ to represent a normal distribution with parameters μ for the mean and σ for the standard deviation (and not using $N(\mu, \sigma^2)$).

For all problems, use the $N(250, 30)$ distribution.

1. $P(X < 200)$.

Solution

Using R, we determine that $P(X < 200) = 0.0478$

```
R Code
pnorm(200,250,30)
```

2. $P(X > 260)$.

Solution

Using R, we determine that $P(X > 260) = 1 - P(X < 260) = 0.36944$.

```
R Code
1-pnorm(260,250,30)
```

3. $P(|X - 250| > 40)$.

Solution

Using R, we find that

$$\begin{aligned} P(|X-250|>40) &= P((X-250<-40)\cup(40<X-250)) \\ &= P(X-250<-40)+P(40<X-250) \\ &= P(X<-40+250)+P(40+250<X) \\ &= P(X<210)+P(290<X) \\ &= P(X<210)+(1-P(290<X)) \\ &= 0.18242 \end{aligned}$$

```
R Code
(1-pnorm(290,250,30))+(pnorm(210,250,30))
```

4. $P(260 < X < 300)$.

Solution

Using R, we find that

$$\begin{aligned} P(260<X<300) &= P(X<300)-P(X<260) \\ &= 0.32165 \end{aligned}$$

```
R Code
pnorm(300,250,30)-pnorm(260,250,30)
```

5. The number c so that $P(X < c) = 0.9$.

Solution

Using R, we find that the value c such that $P(X < c) = 0.9$ is 288.45.

```
R Code
qnorm(0.9,250,30)
```

6. The number c so that $P(X > c) = 0.24$.

Solution
Using R, we find that

$$\begin{aligned}
P(X > c) &= 0.24\\
\Rightarrow 1 - P(X > c) &= 1 - 0.24\\
\Rightarrow P(X < c) &= 0.76\\
\Rightarrow c &= 271.1891
\end{aligned}$$

```
R Code
qnorm(0.76,250,30)
```

7. The number c so that $P(|X - 250| > c) = 0.18$.

Solution
Using R, we find that

$$\begin{aligned}
P(|X - 250| > c) &= 0.18\\
\Rightarrow P((X - 250 < -c) \cup (c < X - 250)) &= 0.18\\
\Rightarrow P(X - 250 < -c) + P(c < X - 250)) &= 0.18\\
\Rightarrow P(X - 250 < -c) + P(X - 250 < -c) &= 0.18\\
\Rightarrow 2P(c < X - 250)) &= 0.18\\
\Rightarrow P(c + 250 < X)) &= 0.09\\
\Rightarrow P(c + 250 > X)) &= 1 - 0.09\\
\Rightarrow P(c + 250 > X)) &= 0.91\\
\Rightarrow c + 250 &= 290.2227\\
\Rightarrow c &= 40.22
\end{aligned}$$

```
R Code
qnorm(0.91,250,30)
```

8. The number c so that $P(|X - 250| < c) = 0.9$.

Solution
First consider that

$$\begin{aligned}
P(-c < X - 250 < c) &= 0.9\\
P(X - 250 < c) - P(X - 250 > -c) &= 0.9\\
P(X - 250 < c) - (1 - P(X - 250 < c)) &= 0.9\\
P(X - 250 < c) - 1 + P(X - 250 < c) &= 0.9\\
2P(X - 250 < c) - 1 &= 0.9\\
2P(X - 250 < c) &= 1.9\\
P(X - 250 < c) &= \frac{1.9}{2}\\
P(X < c + 250) &= 0.95
\end{aligned}$$

From R, we determine that

$$\begin{aligned} c + 250 &= 299.3456 \\ \Rightarrow c &= 49.35 \end{aligned}$$

```
R Code
qnorm(0.95,250,30)
```

R problem 2 (Graded for Accurateness) The height of the density function of a normal curve can be computed with the R function dnorm(). Write a function called gnorm() that will draw a sketch of a normal density and shade in the two tails with probability $\alpha/2$ if one passes in a mean, sd, and alpha value. Modify this function which calculates and draws $P(X \leq a)$.

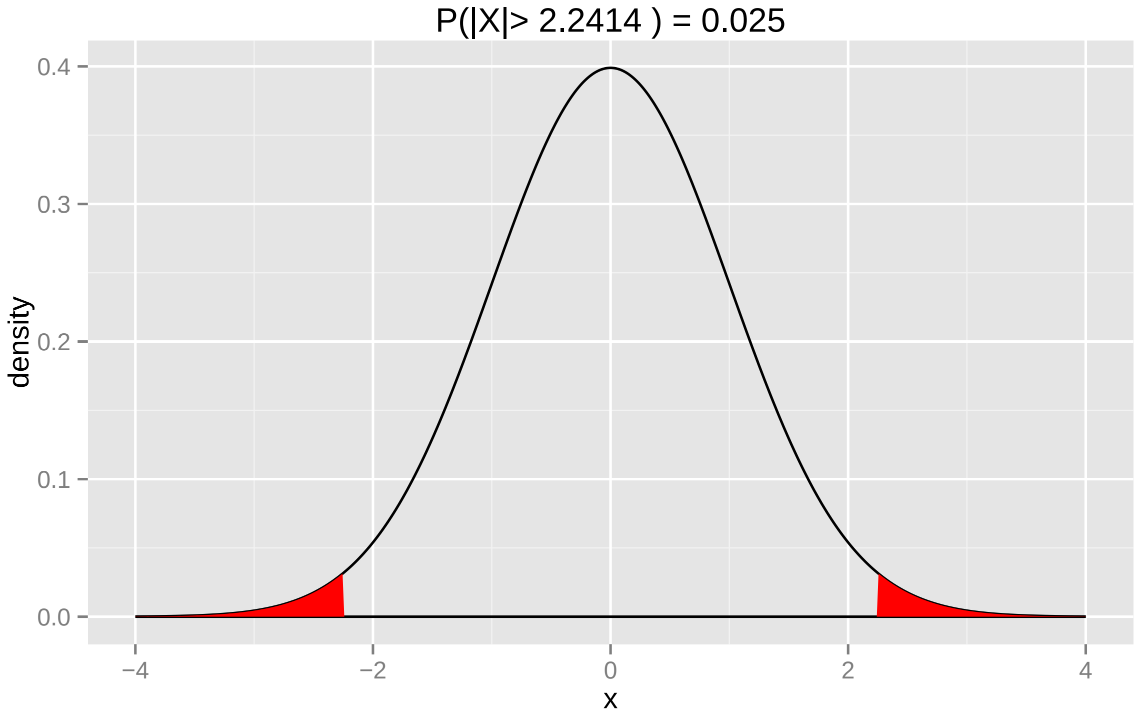
```
gnorm = function(a,mu=0,sigma=1) {
# create an array of x values of length 501
# from 4 SDs below to 4 SDs above the mean
x = seq(mu-4*sigma,mu+4*sigma,length=501)
# calculate the height of the normal density at these points
y = dnorm(x,mean=mu,sd=sigma)
# put x and y into a data frame
d = data.frame(x,y)
# create a plot that graphs the normal density
# and overlays this with a horizontal line for the x axis
# store the plot as the object p and later add more to it
require(ggplot2)
p = ggplot(d, aes(x=x,y=y)) + geom_line() +
        geom_segment(aes(x=mu-4*sigma,xend=mu+4*sigma,y=0,yend=0),
            data=data.frame(mu,sigma)) +
ylab(’density’)
# shade in the area
# add points to the data frame d that are the bottom of the segment to shade
# and extract only those x and y from d where x <= a
d2 = data.frame(x = c(mu-4*sigma,x[x<=a],a), y = c(0,y[x<=a],0))
p = p + geom_polygon(aes(x=x,y=y),data=d2,fill="red") +
        ggtitle(paste("P(X <",a,") =",round(pnorm(a,mu,sigma),4)))
plot(p)
}
```

<u>Solution</u>
Below is code for a function that will perform the desired task.

```
anorm = function(alpha,mu=0,sigma=1) {
  x = seq(mu-4*sigma,mu+4*sigma,length=501)
  y = dnorm(x,mean=mu,sd=sigma)
  d = data.frame(x,y)
  require(ggplot2)
  p = ggplot(d, aes(x=x,y=y)) + geom_line() +
    geom_segment(aes(x=mu-4*sigma,xend=mu+4*sigma,y=0,yend=0),
                 data=data.frame(mu,sigma)) + ylab(’density’)
  a <- qnorm(alpha/2,mu,sigma)
  d2 = data.frame(x = c(mu-4*sigma,x[x<=a],a), y = c(0,y[x<=a],0))
  b <- qnorm(1-alpha/2,mu,sigma)
  d3 = data.frame(x = c(b,x[x>=b],mu+4*sigma), y = c(0,y[x>=b],0))
  p = p + geom_polygon(aes(x=x,y=y),data=d2,fill="red") +
    geom_polygon(aes(x=x,y=y),data=d3,fill="red")+
    ggtitle(paste("P(|X|>",round(b,4),") =",alpha))
  plot(p)
}
```

For example, this code can be used in the following manner to create the image below.

```
gnorm(.025,0,1)
```



P(|X|> 2.2414) = 0.025
density
0.4
0.3
0.2
0.1
0.0
x
−4
−2
0
2
4