Unit 1 – Review of BIOSTATS 540
Practice Problems
SOLUTIONS - R

Syllabus questions

Please see syllabus; I really want you to have read it. Thank you! cb.

Preliminary – R Users

```
# KEY
# A leading hashtag denotes a comment.  Nothing is executed.  Good for documentation.

# Preliminary - Install Packages Used (one time).  Don't forget to enclose package name in quotes.
# If not installed, remove leading # in the following two commands
# install.packages("DescTools")
# install.packages("ggplot2")
```

#1. (Reviews BIOSTATS 540 Unit 1).

The following table lists length of stay in hospital (days) for a sample of 25 patients.

5	10	6	11	5	14	30	11	17	3
9	3	8	8	5	5	7	4	3	7
9	11	11	9	4					

Using **R**, construct a frequency/relative frequency table for these data using 5-day class intervals. Include columns for the frequency counts and relative frequencies.

```
# Create vector los containing data.
los <- c(5,10,6,11,5,14,30,11,17,3,9,3,8,8,5,5,7,4,3,7,9,11,11,9,4)

# Create los_cat, grouped measure of los using 5-day intervals.
los_cat <- cut(los, c(0,5,10,15,20,25,30), right=TRUE)        # right=TRUE for closed intervals right

# Solution I - User codes
n <- length(los)                              # total sample size
los_freq <- table(los_cat)                    # interval frequency
los_relfreq <- los_freq/n                     # interval relative frequency
los_cum <- cumsum(los_freq)                   # interval cumulative frequency
los_cumrel <- cumsum(los_relfreq)             # interval cumulative relative frequency

# Create q1table:  Bind together objects
q1table <- cbind(los_freq, los_relfreq, los_cum, los_cumrel)

# Label columns
colnames(q1table) <- c("Freq", "Rel Freq", "Cum Freq", "Cum Rel Freq").
```

```
# Show
q1table

##          Freq Rel Freq Cum Freq Cum Rel Freq
## (0,5]      9     0.36        9          0.36
## (5,10]     9     0.36       18          0.72
## (10,15]    5     0.20       23          0.92
## (15,20]    1     0.04       24          0.96
## (20,25]    0     0.00       24          0.96
## (25,30]    1     0.04       25          1.00

# Solution II - Using package {DescTools}.  Command is Freq( )
# Tip - Turn of scientific notation first using options(scipen=100)
options(scipen=1000)
library(DescTools)
Freq(los_cat)

##      level  freq   perc  cumfreq  cumperc
## 1    (0,5]     9  36.0%        9    36.0%
## 2   (5,10]     9  36.0%       18    72.0%
## 3  (10,15]     5  20.0%       23    92.0%
## 4  (15,20]     1   4.0%       24    96.0%
## 5  (20,25]     0   0.0%       24    96.0%
## 6  (25,30]     1   4.0%       25   100.0%
```

#2. (Reviews BIOSTATS 540 Unit 2).

The following table lists fasting cholesterol levels (mg/dl) for two groups of men.

Group 1:									
233	291	312	250	246	197	268	224	239	239
254	276	234	181	248	252	202	218	212	325
Group 2:									
344	185	263	246	224	212	188	250	148	169
226	175	242	252	153	183	137	202	194	213

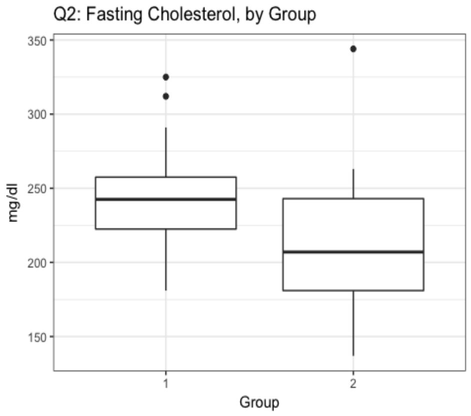
Using **R**, construct the following graphical comparisons of the two groups:

```
# STEP 1:  Get data into R

# Solution I - direct entry
### create variable group
group <- c(rep(1,times=10), rep(2,times=10))

###  create outcome variable ychol
ychol <- c(233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218,
212, 325, 344, 185, 263, 246, 224, 212, 188, 250, 148, 169, 226, 175, 242, 252, 153, 183, 137, 202,
194, 213)
```

```
### create R dataset using data.frame( )
q2data <-  data.frame(group, ychol)




# Solution II - Load R data from course website
# Note - User must change desktop path from /Users/cbigelow/Desktop to User’s own
# Note - I then created a copy of this that I named q2data for ease of use in solutions below!

setwd("/Users/cbigelow/Desktop")
load(file="BIOSTATS640_hw01_q2.Rdata")
q2data <- BIOSTATS640_hw01_q2
```

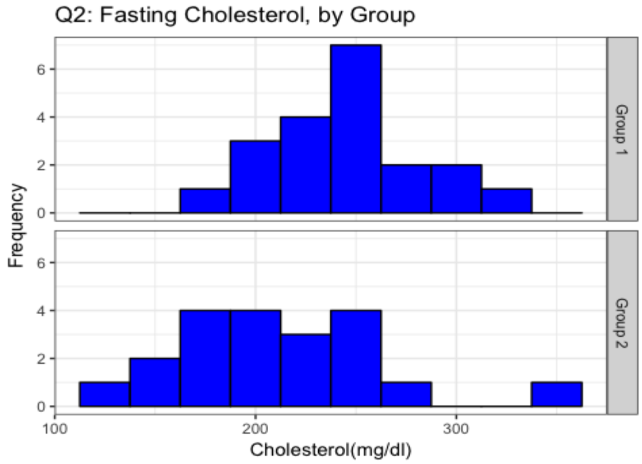
2a. Side-by-side box plot

```
# Preliminary:  Must declare group variable to be of type=factor
q2data$group <- as.factor(q2data$group)

# Side-by-side box plot using package {ggplot2}.
# KEY:
# qplot(groupvar, yvar, data=dataframename, geom="boxplot",OTHEROPTIONS)
library(ggplot2)
qplot(group, ychol, data=q2data,
      geom="boxplot",
      main="Q2: Fasting Cholesterol, by Group",
      xlab="Group",
      ylab="mg/dl") + theme_bw()
```



2b. Side-by-side histograms with same definitions (starting value, ending value, tick marks, etc) of the horizontal and vertical axes.

```
# Preliminary:  Attach labels to group values
q2data$group <- factor(q2data$group, levels=c(1,2), labels=c("Group 1", "Group 2"))

# KEY:
ggplot(data=dataframe, aes(x=VARTOPLOT)) + geom_histogram( ) + options
ggplot(data=q2data, aes(x=ychol)) +
        geom_histogram(color="black", fill="blue", binwidth=25) +
        facet_grid(group ~ .) +
        theme_bw() +
        labs(title="Q2: Fasting Cholesterol, by Group",
             x="Cholesterol(mg/dl)",
             y="Frequency")
```



In 1-2 sentences, compare the two distributions. What conclusions do you draw?

Comparison of the distributions: Men in Group 1 tend to have higher fasting cholesterol values, as reflected in the upward shift in location of data points. The variation in fasting cholesterol is slightly greater for men in Group 2; this is most easily seen in the side-by-side box plot, which shows a larger interquartile range (the size of the box itself) and a more distant outlier.

#3. (Reviews BIOSTATS 540 Unit 6 – Bernoulli and Binomial).

Consider the following setting.
Seventy-nine firefighters were exposed to burning polyvinyl chloride (PVC) in a warehouse fire in Plainfield, New Jersey on March 20, 1985. A study was conducted in an attempt to determine whether or not there were short- and long-term respiratory effects of the PVC. At the long term follow-up visit at 22 months after the exposure, 64 firefighters who had been exposed during the fire and 22 firefighters who where not exposed reported on the presence of various respiratory conditions. Eleven of the PVC exposed firefighters had moderate to severe shortness of breath compared to only 1 of the non-exposed firefighters.

Calculate the probability of finding 11 or more of the 64 exposed firefighters reporting moderate to severe shortness of breath if the rate of moderate to severe shortness of breath is 1 case per 22 persons. **Show your work**.

Answer: 0.000136

```
# dbinom(x,ntrials,pevent) yields Pr[X = x]
# pbinom(x,ntrials,pevent) yields Pr[X <= x]
# sum(dbinom(a:b,ntrials,pevent)) yields Pr[a <= X <= b]
#
# Get answer with way too many digits printed.  Display.
q3full <- sum(dbinom(11:64,64,1/22))
q3full

## [1] 0.0001360821



# Too many digits shown?  Round the answer to 6 places after the decimal point. Display.
q3round6 <- round(q3full, 6)
q3round6

## [1] 0.000136

# Print out the rounded answer so as not to inadvertently communicate unrealistic precision!
paste("Pr[11 or more events, ntrials=64,p=1/22] = ", q3round6)
## [1] "Pr[11 or more events, ntrials=64,p=1/22] =  0.000136"
```

#4. (Reviews BIOSTATS 540 Unit 7- Normal Distribution).

The Air Force uses ACES-II ejection seats that are designed for men who weigh between 140 lb and 211 lb. Suppose it is known that women's weights are distributed Normal with mean 143 lb and standard deviation 29 lb.

4a. What proportion of women have weights that are ***outside*** the ACES-II ejection seat acceptable range?

Answer: 46.8%

```
# pnorm(x, mean=##, sd=##) yields Pr[X <= x]
# 1-pnorm(x, mean=##, sd=##) yields Pr[X >= x]
# pnorm(x,mean=##,sd=##, lower.tail=FALSE) also yields Pr[X >= x]
q4afull <- pnorm(140, mean=143, sd=29) + (1-pnorm(211,mean=143,sd=29))
q4around <- round(q4afull, 6)


paste("Pr[X<=140 OR X>=211 | mean=143, sd=29]  =  ", q4around)
## [1] "Pr[X<=140 OR X>=211 | mean=143, sd=29]  =   0.468322"
```

4b. In a sample of 1000 women, how many are expected to have weights below the 140 lb threshold?

Answer: 459 since
459 = 1000 women x (.4588 likelihood of weight below 140 lb)

```
q4b <- 1000*pnorm(140,mean=143, sd=29)
q4bround <- round(q4b, 0)

paste("Expected # with weight below 140 lb  =  ", q4bround)
## [1] "Expected # with weight below 140 lb  =   459"
```

#5. (Reviews BIOSTATS 540 Units 8 and 9).

Consider the setting of a single sample of n=16 data values that are a random sample from a normal distribution. Suppose it is of interest to perform a type I error α = 0.01 statistical hypothesis test of H_O: $\mu \geq 100$ versus H_A: $\mu < 100$, $\alpha = 0.01$. Suppose further that σ is unknown.

5a. State the appropriate test statistic

Answer: Student t with degrees of freedom = 15.

Solution:
This is a one-sample setting of normally distributed data where the population variance is not known and interest is in the mean. Because the sample size is 16, the degrees of freedom is (16-1)=15.

5b. Determine the critical region for values of the sample mean $\bar{X}$.

Answer: $\bar{X} \leq (S/4)(-2.602) + 100$

Solution:

Given: $n=16,\ \alpha=0.01, \text{one sided(left)}, \mu_O=100 \ \rightarrow$

(1) Solution for $\hat{SE} = \frac{S}{\sqrt{16}} = \frac{S}{4}$

(2) Solution for $t_{CRITICAL} = t_{.01;df=15} = -2.602$

(3) Solution for $\bar{X}_{\text{CRITICAL}}$ is obtained by its solution in the following expression:

$$t_{\text{OBSERVED}} \leq t_{\text{CRITICAL}} \rightarrow$$

$$\frac{\bar{X} - \mu_{\text{NULL}}}{\hat{SE}} \leq -2.602 \rightarrow$$

$$\bar{X} - \mu_{\text{NULL}} \leq (\hat{SE})(-2.602) \rightarrow$$

$$\bar{X} \leq (\hat{SE})(-2.602) + \mu_{\text{NULL}} \rightarrow$$

$$\bar{X} \leq (S/4)(-2.602) + \mu_{\text{NULL}} \rightarrow$$

$$\bar{X} \leq (S/4)(-2.602) + 100$$

#6. (Reviews BIOSTATS 540 Unit 9).

An investigator is interested in the mean cholesterol level μ of patients with myocardial infarction. S/he drew a simple random sample of n=50 patients and from these data constructed a 95% confidence interval for the mean μ. In these calculations, it was assumed that the data are a simple random sample from a normal distribution with known variance. The resulting width of the confidence interval was 10 mg/dl.

How large a sample size would have been required if the investigator wished to obtain a confidence interval width equal to 5 mg/dl?

Answer: 200

Solution:

(Step 1) Solution for value of confidence coefficient:
95% CI and σ known → Desired confidence coefficient is 97.5th percentile of Normal(0,1) = 1.96

(Step 2) Expression for CI width:

$$\text{width} = (\text{upper limit}) - (\text{lower limit})$$

$$= \left(\bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) - \left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}\right)$$

$$= \frac{1.96\sigma}{\sqrt{n}} - \left(-\frac{1.96\sigma}{\sqrt{n}}\right)$$

$$= \frac{(2)(1.96)\sigma}{\sqrt{n}}$$

(Step 3) Using known width=10 and known n=50, obtain σ = 18.0384

$$10 = \frac{(2)(1.96)\sigma}{\sqrt{n}} \rightarrow$$

$$\frac{(10)(\sqrt{n})}{(2)(1.96)} = \sigma \rightarrow$$

$$\frac{(10)(\sqrt{50})}{(2)(1.96)} = \sigma \rightarrow$$

$$\sigma = 18.0384$$

(Step 4) Using known width=5 and σ = 18.0384 known, obtain n=200

$$5 = \frac{(2)(1.96)\sigma}{\sqrt{n}} \rightarrow$$

$$\sqrt{n} = \frac{(2)(1.96)\sigma}{5} \rightarrow$$

$$\sqrt{n} = \frac{(2)(1.96)(18.0384)}{5} \rightarrow$$

$$\sqrt{n} = 14.1421 \rightarrow$$

$$n = 199.99$$

Or 200, by rounding up.

#7. (Reviews BIOSTATS 540 Units 9 and 10).

In (a) – (d) below, you may assume that the data are a simple random sample (or samples) from a normal distribution (or distributions). Each setting is a different setting of confidence interval estimation. In each, state the values of the confidence coefficients *(recall – these will be the values of specific percentiles from the appropriate probability distribution)*.

7a.

For a single sample size of n=15 and the estimation of the population mean μ when the variance is unknown using a 90% confidence interval, what are the values of the confidence coefficients?

Answer: -1.7613 and + 1.7613

```
# qt(lefttailprobability,df=##) yields percentile of Student-t
# For desired confidence = 90%, split the remaining 10% in the tails -> want 95th percentile
q7a <- qt(.95, df=14)
q7around <- round(q7a, 4)

paste("For 90% CI, want 95th percentile of T (df=14)  =  ", q7around)
## [1] "For 90% CI, want 95th percentile of T (df=14)  =   1.7613"
```

7b.

For a single sample size n=35 and the estimation of a variance parameter σ^2 using a 95% confidence interval, what are the values of the confidence coefficients?

Answer: 19.81 and 51.97

```
# qchisq(lefttailprobability,df=##) yields percentile of Chi Square
q7bleft <- qchisq(.025, df=34)
q7blr <- round(q7bleft, 4)
q7bright <- qchisq(.975, df=34)
q7brr <- round(q7bright, 4)


paste("For 95% CI, want 2.5th percentile of Chi Square (df=34)  =  ", q7blr)
## [1] "For 95% CI, want 2.5th percentile of Chi Square (df=34)  =   19.8063"

paste("and we want the 97.5th percentile of Chi Square (df=34)  =  ", q7brr)
## [1] "and we want the 97.5th percentile of Chi Square (df=34)  =   51.966"
```

7c.

For a single sample size of n=25 and the estimation of the population mean μ when the variance is known using a 80% confidence interval, what are the values of the confidence coefficients?

Answer: -1.282 and + 1.282

```
# qnorm(lefttailprobability,mean=0, sd=1) yields percentile of Normal(0,1)
# for desired confidence = 80% split the remaining 20% in the tails → want 90th percentile
q7c <- qnorm(.90, mean=0, sd=1)
q7cround <- round(q7c, 4)
paste("For 80% CI, want 90th percentile of Normal(0,1)  =  ", q7cround)
## [1] "For 80% CI, want 90th percentile of Normal(0,1)  =   1.2816"
```

7d.

For the setting of two independent samples, one with sample size n_1 = 13 and the other with sample size n_2 = 22, it is of interest to construct a 90% confidence interval estimate of the ratio of the two population variances, $[\ \sigma_1^2/\sigma_2^2\ ]$. What are the values of the confidence coefficients?

Answer: 0.39 and 2.25

```
# qf(lefttailprobability,df1=##, df2=##) yields percentile of F
q7dleft <- qf(.05, df1=12, df2=21)
q7dright <- qf(.95, df1=12, df2=21)
q7dleftr <- round(q7dleft, 4)
q7drightr <- round(q7dright, 4)


paste("For 90% CI, want 5th percentile of F (df=12, 21)  =  ", q7dleftr)
## [1] "For 90% CI, want 5th percentile of F (df=12, 21)  =   0.3948"
```

```
paste("and we want the 95th percentile of F (df=12, 21)  =  ", q7drightr)
## [1] "and we want the 95th percentile of F (df=12, 21)  =   2.2504"
```

#8. (Reviews BIOSTATS 540 Unit 10).

A study was investigated of length of hospital stay associated with seat belt use among children hospitalized following motor vehicle crashes. The following are the observed sample mean and sample standard deviations for two groups of children: 290 children who were ***not*** wearing a seat belt at the time of the accident plus 123 children who ***were*** wearing a seat belt at the time of the accident.

Group	Sample size, n	Sample mean	Sample standard deviation
Seat belt = no	$n_{NO} = 290$	$\bar{X}_{NO} = 1.39$ days	$S_{NO} = 3.06$ days
Seat belt = yes	$n_{YES} = 123$	$\bar{X}_{YES} = 0.83$ days	$S_{YES} = 2.77$ days

You may assume normality. You may also assume that the unknown variances are equal. Construct a 95% confidence interval estimate of the difference between the two population means. In developing your answer, you may assume that the population variances are unknown but EQUAL.

8a.

What is the value of the point estimate?

Answer: 0.56

Solution:

$$\text{Point estimate of } [\mu_1 - \mu_2] = [\bar{X}_1 - \bar{X}_2] = [1.39 - 0.83] = 0.56$$

8b.

What is the value of the estimated standard error of the point estimate?

Answer: 0.32

Solution:

(1) Preliminary: Obtain S^2_{pool}

$$S^2_{pool} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1) + (n_2-1)} = \frac{(289)(9.3636) + (122)(7.6729)}{(289) + (122)} = 8.8617377$$

(2) Solution for $\hat{SE}[\bar{X}_1 - \bar{X}_2]$

$$\hat{SE}[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} = \sqrt{\frac{8.8617377}{290} + \frac{8.8617377}{123}} = 0.320319$$

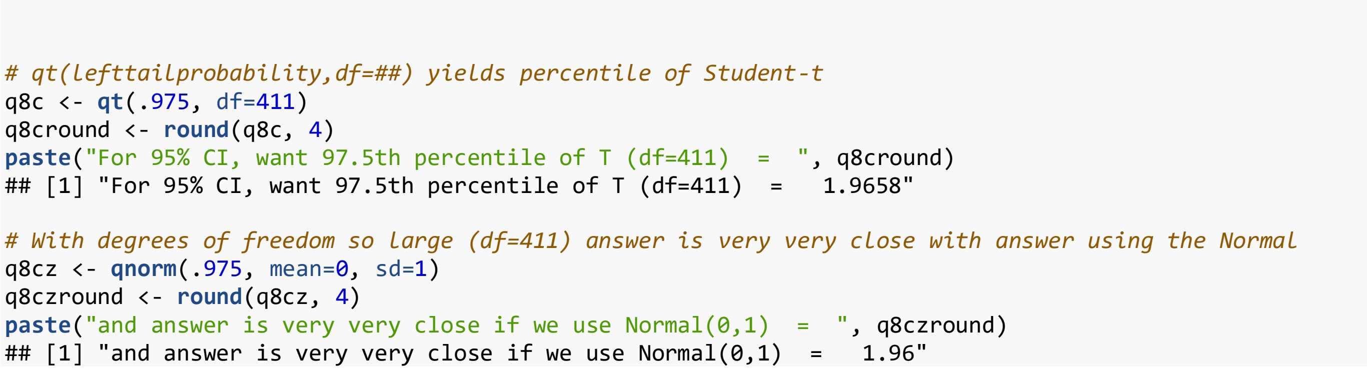
(3) Solution for Degrees of Freedom, df

$df = (n_1 - 1) + (n_2 - 1) = (290-1) + (123-1) = 411$

8c.

What is the value of the confidence coefficient?

Answer: 1.966 using Student-t.
1.96 using Normal(0,1). Close!



```
# qt(lefttailprobability,df=##) yields percentile of Student-t
q8c <- qt(.975, df=411)
q8cround <- round(q8c, 4)
paste("For 95% CI, want 97.5th percentile of T (df=411)  =  ", q8cround)
## [1] "For 95% CI, want 97.5th percentile of T (df=411)  =   1.9658"

# With degrees of freedom so large (df=411) answer is very very close with answer using the Normal
q8cz <- qnorm(.975, mean=0, sd=1)
q8czround <- round(q8cz, 4)
paste("and answer is very very close if we use Normal(0,1)  =  ", q8czround)
## [1] "and answer is very very close if we use Normal(0,1)  =   1.96"
```

8d.

What are values of the lower and upper limits of the confidence interval?

Answer: [-0.07, + 1.19]

Solution:

$$\begin{aligned} CI &= [\bar{X}_1 - \bar{X}_2] \pm \left(t_{.975;df=411}\right)\hat{SE}(\bar{X}_1 - \bar{X}_2) \\ &= 0.56 \pm (1.966)(0.32) \\ &= [\ -0.06912\ ,\ +1.18912\ ] \end{aligned}$$

8e.

Write a clear interpretation of the confidence interval.

With 95% confidence, from these data, it is estimated that the difference in average length of stay (non-seat belt wearers minus seat belt wearers) is between -0.07 days and +1.19 days. Since this interval includes 0, these data do not provide statistically significant evidence that the length of hospital stay for children in motor vehicle crashes who were not wearing seat belts is different than the length of hospital stay for children in motor vehicle crashes who were wearing seat belts.

#9. **(Reviews BIOSTATS 540 Unit 6).**
A test consists of multiple-choice questions, each having four possible answers, one of which is correct. What is the probability of getting exactly four correct answers when six guesses are made?

Answer: .03

This is a binomial probability calculation.
N=6 π=.24 Want Pr[X = 4]

$$\Pr[X=4] = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \binom{6}{4} .25^4 (.75)^2 = .032959$$

```
# dbinom(x,ntrials,pevent) yields Pr[X = x]
q9 <- dbinom(4, 6, .25)
q9round <- round(q9, 4)


paste("Pr[X=4 | ntrials=6, p=.25] = ", q9round)
## [1] "Pr[X=4 | ntrials=6, p=.25] =  0.033"
```

#10. **(Reviews BIOSTATS 540 Unit 6).**
After being rejected for employment, woman "A" learns that company "X" has hired only 2 women among the last 20 new employees. She also learns that the pool of applicants is very large, with an approximately equal number of qualified men and women. Help her address the charge of gender discrimination by finding the probability of getting 2 or fewer women when 20 people hired under the assumption that there is no discrimination based on gender. Does the resulting probability really support such a charge?

Answer: .0002
This is a very small probability. As such, it would support a charge of gender discrimination but only under the circumstances where, for each position filled, there an equal number of men and women applicants.

Solution:

This is also a binomial probability calculation.
Here, n=20 π=.50 Want Pr[X $\leq$ 2]

$$\Pr[X \leq 2] = \sum_{x=0}^{2} \binom{n}{x} \pi^x (1-\pi)^{n-x} = \sum_{x=0}^{2} \binom{20}{x} .50^x (.50)^{20-x} = .00020122$$

```
# pbinom(x,ntrials,pevent) yields Pr[X <= x]
q10 <- pbinom(2, 20, .50)
q10round <- round(q10, 4)
```

```
paste("Pr[X<=2 | ntrials=20, p=.50] = ", q10round)
## [1] "Pr[X<=2 | ntrials=20, p=.50] =  0.0002"
```

#11. (Reviews BIOSTATS 540 Unit 7).

Suppose the length of newborn infants is distributed normal with mean 52.5 cm and standard deviation 4.5 cm. What is the probability that the mean of a sample of size 15 is greater than 56 cm?

Answer: .0013

Solution:

This is a normal distribution probability calculation.

$$\bar{X}_{n=15} \text{ is distributed Normal with } \mu_{\bar{X}}=52.5 \text{ and } se(\bar{X}_{n=15}) = \frac{\sigma}{\sqrt{15}} = \frac{4.5}{\sqrt{15}} = 1.1619$$

$$\text{Want } \Pr[\bar{X}_{n=15} \geq 56] = \Pr\left[\frac{\bar{X}_{n=15} - \mu_{\bar{X}}}{se\left(\bar{X}_{n=15}\right)} \geq \frac{56 - 52.5}{1.1619}\right] = \Pr\left[\text{Z-score} \geq 3.0123\right] = .001296$$

```
# 1-pnorm(x, mean=##, sd=##) yields Pr[X >= x]
# pnorm(x,mean=##,sd=##, lower.tail=FALSE) also yields Pr[X >= x]
q11 <- pnorm(56, mean=52.5, sd=1.1619, lower.tail=FALSE)
q11round <- round(q11,4)
paste("Pr[Xbar >= 56 |  mean=52.5, SE(xbar)=1.1619]  =  ", q11round)
## [1] "Pr[Xbar >= 56 |  mean=52.5, SE(xbar)=1.1619]  =   0.0013"
```

#12. (Reviews BIOSTATS 540 Unit 7).

Suppose that 25 year-old males have a remaining life expectancy of an additional 55 years with a standard deviation of 6 years. Suppose further that this distribution of additional years life is normal. What proportion of 25 year-old males will live past 65 years of age?

Answer: 99.4%

This is also a normal distribution probability calculation.

$$X \text{ is distributed Normal with } \mu = 55 \text{ and } \sigma = 6$$

"*Living past 65 years of age*" corresponds to a remaining life expectancy of an *additional 40+ years*.

Thus, want:

$$\Pr[X \geq 40] = \Pr\left[\frac{X - m}{s} \geq \frac{40 - 55}{6}\right] = \Pr\left[\text{Z-score} \geq -2.5\right] = .9938$$

```
# 1-pnorm(x, mean=##, sd=##) yields Pr[X >= x]
# pnorm(x,mean=##,sd=##, lower.tail=FALSE) also yields Pr[X >= x]
q12 <- pnorm(40, mean=55, sd=6, lower.tail=FALSE)
q12round <- round(q12,4)
paste("Proportion living an additional 40 years  =  ", q12round)
## [1] "Proportion living an additional 40 years  =   0.9938"
```