STAT 350 Exam 2 Review Problems **Key**

1. A random sample of 26 offshore oil workers took part in a simulated escape exercise, and their times (in seconds) to complete the escape were recorded. The sample mean is 370.69 sec. and the sample standard deviation is 24.36 sec. Construct a 95% lower confidence bound for the true average escape time. Interpret your interval and write down the critical value. Do the problem assuming that you don't know the data (be sure to include the code and the value and code of the critical value) and using the R output below (write down the code that was used to generate the output):

```
            One Sample t-test

data:  escape$time
95 percent confidence interval:
 362.53      Inf
```

The interpretation does not change from either method.

Since population standard deviation σ is not known, use a t-procedure.

$$\mu > \bar{x} - t_{\alpha,n-1} * \frac{s}{\sqrt{n}}$$

Don't know the data:

Code:

```
> xbar <- 370.69
> s <- 24.36
> n <- 26
> C <- 0.95
> t <- qt(1-C,n-1,lower.tail = FALSE)
> t
[1] 1.708141
> xbar - t*s/sqrt(n)
[1] 362.5295
```

The critical value is 1.708141.

Using the output provided:

Code: `t.test(escape$time, conf.level = 0.95,alternative = "greater")`

Answer: 362.53

We are 95% confident that the true mean escape time is greater than 362.53 seconds.

2. The life in hours of a battery is known to be approximately normally distributed. The manufacture claims that the average battery life exceeds 40 hours. A random sample of ten batteries has a mean life of 40.5 hours and sample standard deviation of 1.25 hours. Carry out the appropriate hypothesis test with a significance level of 0.05. Do the problem assuming that you don't know the data (be sure to include the code) and using the R output below (write down the code that was used to generate the output):

```
            One Sample t-test

data:  battery$lifetime
t = 1.2649, df = ?, p-value = 0.1188
```

Since population standard deviation σ is not known, use t-procedure.

STAT 350 Exam 2 Review Problems **Key**

The only thing different is in Step 3:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Don't know the data:

Code:

```
> mu0 <- 40
> xbar <- 40.5
> s <- 1.25
> n <- 10
> tts <- (xbar - mu0)/(s/sqrt(n))
> tts
[1] 1.264911
> pt(tts,n-1,lower.tail = FALSE)
[1] 0.1188363
```

Using the output provided:
Code: `t.test(battery$lifetime, conf.level = 0.95,alternative = "greater")`

tts = 1.2649
df = 10 – 1 = 9
p = 0.1188

Answer:

1: μ is the population average battery life.

2: H_0: μ ≤ 40 H_a: μ > 40

3. tts = 1.2649, df = 10 – 1 = 9, p = 0.1188

4. Fail to reject H_o because 0.1188 > 0.05
The data does not provide support (p = 0.1188) to the claim that the population average battery lifetime exceeds 40.

3. The overall distance traveled by a golf ball is tested by hitting the ball with Iron Byron, a mechanical golfer with a swing that is said to emulate the legendary champion, Byron Nelson. Ten randomly selected balls of two different brands are tested and the overall distance measured to determine if the two brands are different. Please provide the code or work for the appropriate parts. The data (in yards) follows:

Brand 1: 275, 286, 287, 271, 283, 271, 279, 275, 263, 267
Brand 2: 258, 244, 260, 265, 273, 281, 271, 270, 263, 268

The data is summarized in the following table:

	n	$\bar{x}$	s
Brand 1	10	275.7	8.03
Brand 2	10	265.3	10.045
1 - 2	10	10.4	15.005

STAT 350 Exam 2 Review Problems Key

If the situation is two-sample independent, use the following data:

```
distance <- c(275, 286, 287, 271, 283, 271, 279, 275, 263, 267, 258, 244,
   260, 265, 273, 281, 271, 270, 263, 268)
brand <- c(rep("Brand1",10),rep("Brand2",10))
```

If the situation is two-sample paired, use the following data:

```
Brand1 <- c(275, 286, 287, 271, 283, 271, 279, 275, 263, 267)
Brand2 <- c(258, 244, 260, 265, 273, 281, 271, 270, 263, 268)
```

a) Which procedure is the most appropriate, two-sample independent or two-sample paired? Please explain your answer. If you choose a matched pairs procedure, please state the common characteristic that makes these data paired.

Two-sample independent procedure, since there are not any variables that can be matched.

b) Should you use a one-sided or two-sided alternative hypothesis? Please explain your answer.

Since the question asks if the two brands are different, a two-sided alternative hypothesis is the correct choice.

c) Use the four-step procedure to carry out a hypothesis test to determine whether the mean overall distance for brand 1 and brand 2 are different. Assume a significance level of 0.05.

Since the data is provided, we use function t.test()
Code and output:

```
> distance <- c(275, 286, 287, 271, 283, 271, 279, 275, 263, 267, 258, 244,
260, 265, 273, 281, 271, 270, 263, 268)
> brand <- c(rep("Brand1",10),rep("Brand2",10))
> t.test(distance ~ brand, conf.level = 0.95, paired = FALSE,
        alternative = "two.sided", var.equal = FALSE)

        Welch Two Sample t-test

data:  distance by brand
t = 2.5575, df = 17.166, p-value = 0.02028
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.82698 18.97302
sample estimates:
mean in group Brand1 mean in group Brand2
               275.7                265.3
```

1: μ_1 is the population mean distance for brand 1
μ_2 is the population mean distance for brand 2
OR μ_2 " brand 2

2. H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 \neq 0$

3. $t_{ts} = 2.5575$, df = 17.166, p = 0.02028
The following equation is not required.

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

4. Reject H_o because 0.0202 < 0.05
The data does provide support (p = 0.0202) to the claim that the population average difference between the brand1 and brand2 of golf balls is different (or not 0)

d) Find and interpret the appropriate 95% confidence interval or bound that corresponds with part c). Be sure to provide the value and code for the critical value.

Using the same code/output as before: (the equation is not required)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,v}\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

The interval is (1.82698, 18.97302)

The critical value is:

```
> qt(0.05/2,17.166,lower.tail = FALSE)
[1] 2.108262
```

We are 95% confident that the difference between the true means of distance that brand1 and brand2 travel is covered by the interval (1.82, 18.97).

e) Why are parts c) and d) saying the same thing?

In part c) 0 is not in the interval and in part d), we reject the null hypothesis.

f) In practical terms, are the two different brands different? Additional information is required.

I don't play golf, but I would assume that 1.82 yards is practically different from 0. Therefore, we need to look at the effect size to see if the difference is measurable.

To get the two standard deviations, use the following code:

```
> tapply(distance,brand,sd)
   Brand1    Brand2
 8.028422 10.044899
```

Effect Size

$$\frac{1.82698}{8.028422} = 0.2276$$

The effect size is moderate, so again, I believe that 1.82 years is practically different from zero with a measurable amount so the brands are practically different.

4. The Indiana State Police wish to estimate the average mph being traveled on the Interstate Highways which cross the state. If the estimate needs to be within ±5 mph of the true mean with 95% confidence and the estimated population standard deviation is 25 mph, how large a sample size must be taken? Please provide the code or work.

$$n = \left(\frac{z_{\alpha/2}\sigma}{ME}\right)^2 = \left(\frac{(1.96)(25)}{5}\right)^2 = 9.8^2 = 96.04 \Rightarrow 97$$

Code:

```
> ME <- 5
> sigma = 25
> z <- qnorm(0.05/2,lower.tail = FALSE)
> z
[1] 1.959964
> (z*sigma/ME)^2
[1] 96.03647
```

So the answer is 97. Remember to also round up on these problems.

5. A laboratory is testing the concentration level in mg/mL for the active ingredient found in a pharmaceutical product. In a random sample of ten vials of the product, the mean and the sample standard deviation of the concentrations are 2.58 mg/mL and 0.09 mg/mL, respectively. Find a 95% confidence interval for the mean concentration level in mg/mL for the active ingredient found in this product. Please interpret your result. Do the problem assuming that you don't know the data (be sure to include the code and the value and code of the critical value) and using the R output below (write down the code that was used to generate the output):

```
            One Sample t-test

data:  escape$time
95 percent confidence interval:
2.515617  2.64438
```

Use one-sample t, as σ is unknown.
This is a confidence interval so it is 2-sided.

Since we don't know the data, we can't use t.test()

Code:

```
> xbar <- 2.58
> s <- 0.09
> n <- 10
> C <- 0.95
> t <- qt((1-C)/2,n-1,lower.tail = FALSE)
> t
[1] 2.262157
> c(xbar - t*s/sqrt(n),xbar + t*s/sqrt(n))
[1] 2.515618 2.644382
```

Using the output provided:
Code: `t.test(escape$time ,conf.level = 0.95, alternative = "two.sided")`

answer: (2.515617, 2.64438)

We are 95% confident that the true mean concentration is covered by the interval (2.52, 2.64).

6. An investigator wishes to estimate the difference between two population mean lifetimes of two different brands of batteries under specified conditions. If the population standard deviations are both roughly 2 hr and the sample size from the first brand is twice the sample size from the second brand, what values of the sample sizes will be necessary to estimate the half-width to within 0.5 hours with 99% confidence?
This is a difficult question.

This has to be done by hand except for the z critical value:

```
> qnorm(0.01/2,lower.tail = FALSE)
[1] 2.575829
```

$\sigma_1 = \sigma_2 = 2$
$n_1 = 2n_2$

STAT 350 Exam 2 Review Problems **Key**

Two-sample means CI for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

$$Margin\ of\ Error = 0.5 = 2.5758 \sqrt{\frac{2^2}{2n_2} + \frac{2^2}{n_2}}$$

$$2.5758 \sqrt{\frac{6}{n_2}} = 0.5$$

$n_2 = 6*(2.5758/0.5)^2 = 159.2 = 160$
$n_1 = 2n_2 = 320$
so $n_1 = 320$ and $n_2 = 160$

7. The following are summary data on the proportional stress limits for two different types of woods, red oak and Douglas fir. We are interested if the proportional stress limits are different. Note that only the summary data is available for this question.

Type of Wood	Sample Size	Sample Mean	Sample Standard Deviation
Red oak	50	8.51	1.52
Douglas fir	62	7.69	3.25

a) Which procedure is the most appropriate, two-sample independent or two-sample paired? Please explain your answer. Stating that the sample sizes are different or there is no information for the paired situation is an incorrect answer. If you choose a matched pairs procedure, please state the common characteristic that makes these data paired.

two-sample independent. When you are comparing the different types of trees, there are no listed variables that can be matched.

b) Should you use a one-sided or two-sided alternative hypothesis? Explain.

Since the question asks if the two trees are different, a two-sided alternative hypothesis is the correct choice.

c) Perform a hypothesis test $\alpha = 0.05$ to determine if the stress limits are different for the two types of woods.

This is a t-distribution because the sample standard deviations are provided

Code for step 3:

```
> n1 <- 50
> x1bar <- 8.51
> s1 <- 1.52
> n2 <- 62
> x2bar <- 7.69
> s2 <- 3.25
> delta0 <- 0
> SE <- sqrt(s1^2/n1+s2^2^2/n2)
> tts <- (x1bar - x2bar - delta0)/SE
> tts
[1] 1.762032
> SE1 <- s1^2/n1
> SE2 <- s2^2/n2
> df <- (SE1+SE2)^2/(SE1^2/(n1-1)+SE2^2/(n2-1))
> df
[1] 90.30732
> 2*pt(-abs(tts),df)
[1] 0.08144778
```

1: μ_{oak} is the population mean proportional stress for red oak.
μ_{fir} is the population mean proportional stress for Douglas fir.

2. H_0: $\mu_{oak} - \mu_{fir} = 0$ H_a: $\mu_{oak} - \mu_{fir} \neq 0$

3. tts = 1.762032, df = 90.307, p = 0.0814

The equation is not required.

$$t = \frac{\bar{x}_{oak} - \bar{x}_{fir} - 0}{\sqrt{\frac{s_{oak}^2}{n_{oak}} + \frac{s_{fir}^2}{n_{fir}}}}$$

4. Fail to reject H_0 because 0.0814 > 0.05

The data might not provide support (p = 0.0814) to the claim that the population mean difference between the proportional stress for Red Oak and Douglas Fir is not the same.

This double negative is required since the conclusion needs to be in terms of the alternative hypothesis.

d) Find the appropriate 95% confidence interval or bound that corresponds to part c). Please interpret your result and write down the critical value.

This is an interval because the hypothesis test is two-sided.

Code:

```
> t <- qt(0.05/2,df,lower.tail = FALSE)
> t
[1] 1.986582
> c(x1bar - x2bar - t*SE,x1bar - x2bar + t*SE)
[1] -0.1044994  1.7444994
```

The critical value is 1.986582.

STAT 350 Exam 2 Review Problems **Key**

Equation is not required.

$$(\bar{x}_{oak} - \bar{x}_{fir}) \pm t_{\alpha/2,\upsilon}\sqrt{\frac{s_{oak}^2}{n_{oak}} + \frac{s_{fir}^2}{n_{fir}}}$$

We are 95% confident that the difference of the average population proportional stress limits for Red Oak versus Douglas Fir is covered by the interval of (-0.104, 1.744)

e) Why are parts c) and d) saying the same thing?

In c) we are failing to reject the null hypothesis that they are the same and in d) because 0 is in the interval.

f) What practical answer would you tell your supervisor concerning the different between the average proportional stress limits for the two types of trees?

Since this is a fail to reject, there is no evidence that the proportional stress is different. However, if in these measurements -0.10 is close to zero, further experimentation might need to be performed.

8. The accompanying summary data on the ratio of strength to cross-sectional area for knee extensors is from the article "Knee Extensor and Knee Flexor Strength: Cross Sectional Area Ratios in Young and Elderly Men": I assuming that these are self-identified men. Note that only the summary data is available for this question.

Group	Sample Size	Sample Mean	Sample Standard Deviation
Young Men	50	7.47	0.44
Elderly Men	45	6.71	0.56

a) Which procedure is the most appropriate, two-sample independent or two-sample paired? Explain. Stating that the sample sizes are different or there is no information for the paired situation is an incorrect answer. If you choose a matched pairs procedure, please state the common characteristic that makes these data paired.

two-sample independent. When you are comparing young men and elderly men, we are not provided with any variable that can be matched.

b) Does the data suggest that the true average ratio for young men exceeds that for elderly men? Carry out a test of significance using $\alpha = 0.01$.

code for step 3:

```
> n1 <- 50
> x1bar <- 7.47
> s1 <- 0.44
> n2 <- 45
> x2bar <- 6.71
> s2 <- 0.56
> delta0 <- 0
> SE <- sqrt(s1^2/n1+s2^2^2/n2)
> tts <- (x1bar - x2bar - delta0)/SE
> tts
[1] 7.299299
> SE1 <- s1^2/n1
> SE2 <- s2^2/n2
> df <- (SE1+SE2)^2/(SE1^2/(n1-1)+SE2^2/(n2-1))
> df
[1] 83.36715
> pt(tts,df,lower.tail = FALSE)
[1] 7.771377e-11
```

1: μ_Y is the population mean cross-sectional area for knee extensor for young men.
μ_E is the population mean cross-sectional area for knee extensor for elderly men.

2. 1. H_0: $\mu_Y - \mu_E \leq 0$ H_a: $\mu_Y - \mu_E > 0$

3. tts = 7.30, df = 83.37, p = 7.77e-11

Equation is not required.

$$t = \frac{\bar{x}_Y - \bar{x}_E - 0}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_E^2}{n_E}}}$$

4. reject H_0 because 7.77e-11 < 0.01
The data does provide strong support (p = 7.77e-11) to the claim that the population mean difference between the cross-sectional area for the knee extensor is larger in younger men versus elderly men.

c) Find and interpret the appropriate confidence interval or bound at a 99% confidence level. Please calculate and write down the critical value.

Because the hypothesis test is an upper tail, we want the lower bound.

code:

```
> t <- qt(0.01,df,lower.tail = FALSE)
> t
[1] 2.371913
> x1bar - x2bar - t*SE
[1] 0.5130374
```

The critical value is 2.371913.

Equation is not required.

$$\mu > (\bar{x}_Y - \bar{x}_E) - t_{\alpha,v}\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_E^2}{n_E}}$$

We are 99% confident that the mean population difference of mean cross-sectional area for knee extensor for young men versus elderly men is greater than 0.513.

d) Why are parts a) and b) saying the same thing?

Since 0.513 (lower bound) is greater than 0, we know that the hypothesis test is a reject H_0.

e) What practical answer would you tell the researcher concerning the difference in the ratio for young men versus elderly men?

The question here is 0.513 practically greater than 0 in this situation. As I do not know what is large in this case, we will continue to look at the effect size to see if the effect is measurable.

Effect Size

$$\frac{0.513}{0.44} = 1.166$$

The effect size is large, so again, I believe that this depends totally on the if value of 0.513 is practically different or not for the difference in the knee extensors for young and elderly men.

9. Coronary heart disease (CHD) begins in young adulthood and is the fifth leading cause of death among adults aged 20 to 24 years. Studies of serum cholesterol levels among college students, however, are very limited. A 1999 study looked at a large sample of students from a large southeastern university and reported that the mean serum cholesterol level among women is **168 mg/dL** with a **population standard deviation** of **27 mg/dL**.
A more recent study at a southern university investigated the lipid levels in a cohort of sedentary university students. The mean total cholesterol level among $\boldsymbol{n = 71}$ women was $\boldsymbol{\bar{x} = 173.7}$. Is there evidence that the mean cholesterol level among sedentary university women has increased from the average in 1999?

a) Use the four-step procedure to carry out a test of significance at $\boldsymbol{\alpha = 0.01}$. Note that the summary data is all that is available for this question.

The population standard deviation is given so we will use a one-sample z-test

code

```
> mu_0 <- 168
> sigma <- 27
> n <- 71
> xbar <- 173.7
> zts <- (xbar - mu_0)/(sigma/sqrt(n))
> zts
[1] 1.778854
> pnorm(z_ts, lower.tail = FALSE)
[1] 0.03763186
```

1. μ is the mean cholesterol level among sedentary self-identified female students.

2. H_0: $\mu \leq 168$ H_a: $\mu > 168$

3. zts = 1.779, p = 0.0377

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$P(Z > 1.778854) = 0.0377$

4. fail to reject H_o because 0.0377 > 0.01
The data does not provide support (p = 0.0377) to the claim that the population mean cholesterol level among sedentary female students in college is greater from female students in 1999 (or > 168).

b) Researchers are concerned that the power may be too low to detect an increase of 5 mg/dl.

i. Determine the cutoff value for $\alpha = 0.01$.

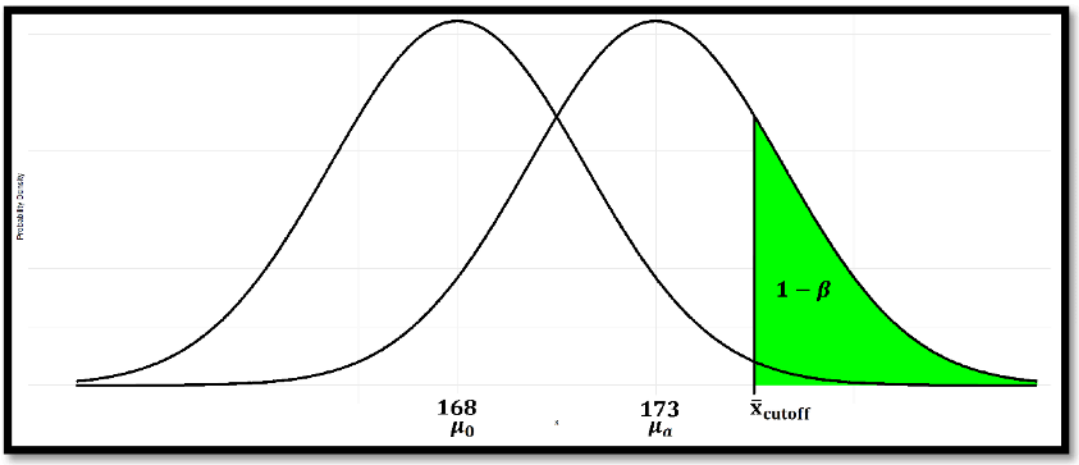
code

```
> z<-qnorm(0.01, lower.tail = FALSE)
> z
[1] 2.326348
> xbar_cutoff <- mu_0+z*sigma/sqrt(n)
> xbar_cutoff
[1] 175.4543
```

$$\bar{\mathbf{x}}_{\mathbf{cutoff}} = \mathbf{175.4543}$$

ii. Determine the power associated with an alternative of 173 mg/dL.

To compute power we need to find the probability of being greater than the cutoff under the alternative value of 173 mg/dL.



$$\mathbf{1 - \beta = P(\bar{X} > \bar{x}_{cutoff} | \mu = 173)}$$

code

```
> mu_a <- 173
> pnorm((xbar_cutoff-mu_a)/(sigma/sqrt(n)),lower.tail=FALSE)
[1] 0.2218531
```

iii. How can the researchers obtain a larger power in a follow up study if they require the same significance level?

The researchers would need to conduct the study using a larger sample size.

10.

STAT 350 Exam 2 Review Problems **Key**

10. In continuation of the investigation into cholesterol levels among university students, an additional follow-up study is planned to be conducted with greater statistical power.

Given that the study wants to detect an **increase of 5 mg/dL (alternative of 173 mg/dL)** from the mean cholesterol level among sedentary university women from the 1999 average of **168 mg/dL**, with a **power of 90%**. **Determine the sample size** required to achieve this level of statistical power? You may assume the **population standard deviation of 27 mg/dL** remains unchanged and a significance level of $\alpha = 0.01$.

To achieve a power of 90% we need the following to hold true. We shall write this out symbolically first and plug in at the end.

$P(\bar{X} > \bar{x}_{cutoff} | \mu = \mu_a) = 0.9$

Formula for the cutoff value is: $\bar{x}_{cutoff} = \mu_0 + z_{0.01}\sigma/\sqrt{n}$

Notice that $\bar{x}_{cutoff}$ **depends on the sample size.**

Standardize with respect to the alternative:

$$P\left(Z > \frac{\bar{x}_{cutoff} - \mu_a}{\sigma/\sqrt{n}}\right) = 0.9$$

For this to hold true we need for $\frac{\bar{x}_{cutoff} - \mu_a}{\sigma/\sqrt{n}}$ **to be equivalent to the normal quantile with 90% of the area under the curve to the right of this value. In other words we need it to equal** $-z_{0.1}$

```
z<-qnorm(0.1, lower.tail = FALSE
>z
[1] 1.281552
```

$$\frac{\bar{x}_{cutoff} - \mu_a}{\sigma/\sqrt{n}} = -z_{0.1}$$

$$\frac{\mu_0 + z_{0.01}\sigma/\sqrt{n} - \mu_a}{\sigma/\sqrt{n}} = -z_{0.1}$$

$$\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_{0.01} = -z_{0.1}$$

$$\sqrt{n} = \frac{(z_{0.1} + z_{0.01})^2}{\left(\frac{\mu_0 - \mu_a}{\sigma}\right)^2}$$

$$n = \left\lceil \left(\frac{z_{0.1} + z_{0.01}}{\frac{\mu_0 - \mu_a}{\sigma}} \right)^2 \right\rceil$$

Where $\lceil \ \rceil$ **is the ceiling function indicating round up to next integer if the value is a decimal.**

Solve for n using:

$\mu_0 = 168$

$\mu_a = 173$

$z_{0.01} = 2.326348$

$z_{0.1} = 1.281552$

$\sigma = 27$

$$n = \left\lceil \left(\frac{1.281552 + 2.326348}{\frac{168 - 173}{27}} \right)^2 \right\rceil$$

$$n = 380$$

To test that this indeed results in a power of approximately 90% we can check in R.

```
> pnorm((mu_0-mu_a)/(sigma/sqrt(380))+z_0.01,lower.tail = FALSE)
[1] 0.9003548
```

11. Fifteen self-identified adult males between the ages of 35 and 45 participated in a study to evaluate the effect of diet and exercise on blood cholesterol levels. The total cholesterol was measured in each subject initially. Each subject then spent three months participating in an aerobic exercise program and switched to a low-fat diet. After the three months, the total cholesterol level was again measured. It is hoped that the cholesterol level decreased after the three months is over. Please provide the code or work for the appropriate parts. The data are shown in the accompanying tables.

Table I: Blood Cholesterol Levels for 15 Adult self-identified Males

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	265	240	258	296	251	245	287	314	260	279	283	240	238	225	247
After	229	231	227	240	238	241	234	256	247	239	246	218	219	226	233

	N	Mean	StDev
Before	**15**	**261.80**	**24.96**
After	**15**	**234.93**	**10.48**
Diff (Before - After)	**15**	**26.87**	**19.04**

If the situation is two-sample independent, use the following code:

```
cholesterol <- c(275, 286, 287, 271, 283, 271, 279, 275, 263, 267, 258,
   244, 260, 265, 273, 281, 271, 270, 263, 268)
time <- c(rep("before",15),rep("after",15))
```

If the situation is two-sample paired, use the following code:

```
before <- c(265, 240, 258, 296, 251, 245, 287, 314, 260, 279, 283, 240,
   238, 225, 247)
after <- c(229, 231, 227, 240, 238, 241, 234, 256, 247, 239, 246, 218, 219,
   226, 233)
```

a) Which procedure is the most appropriate, two-sample independent or two-sample paired? Please explain your answer. Stating that the sample sizes are different or there is no information for the paired situation is an incorrect answer. If you choose a matched pairs procedure, please state the common characteristic that makes these data paired.

The data is paired because you would not expect that the cholesterol levels of the 15 subjects to be similar for all of the self-identified men. Therefore, the matching variable is the self-identified men themselves.

b) Should you use a one-sided or two-sided alternative hypothesis? Please explain your answer.

Since we want to know if the cholesterol level has decreased, this is a one-sided alternative hypothesis.

c) Carry out a hypothesis test to determine if the data support the claim that the low-fat diet and aerobic exercise are of value in reducing the mean blood cholesterol levels? Use α=0.05.

code:

```
> before <- c(265, 240, 258, 296, 251, 245, 287, 314, 260, 279, 283, 240, 238,
225, 247)
> after <- c(229, 231, 227, 240, 238, 241, 234, 256, 247, 239, 246, 218, 219,
226, 233)
> t.test(before,after,paired = TRUE,alternative = "greater",conf.level = 0.95)

        Paired t-test

data:  before and after
t = 5.4488, df = 14, p-value = 4.287e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 18.22722      Inf
sample estimates:
mean of the differences
               26.93333
```

1. μ_d is the true mean difference of blood cholesterol level before the experiment and after the experiment.

2. H_0: $\mu_d \leq 0$ H_a: $\mu_d > 0$
Note: Since this is before – after, a decrease would have a value greater than 0.

3. tts = 5.4488, df = 15 – 1 = 14, p = 4.287e-5

$$t = \frac{\bar{x}_d - 0}{SE_d}$$

4. reject H_o because 4.287e-5 ≤ 0.05
The data does provide strong support (p = 4.287e-5) to the claim that the population mean cholesterol level is lower after the experiment than the before.

d) Find and interpret the appropriate confidence interval or bound at a 95% confidence level. Please calculate and write down the critical value.

code (the rest of the code is given above)

```
> df <- 15 - 1
> qt(0.05,df,lower.tail = FALSE)
[1] 1.76131
```

The critical value is 1.76131.

$$\mu > \bar{x}_d - t_{0.1,n-1}\frac{s_d}{\sqrt{n}}$$

We are 95% confident that the mean population lowering due to the change of conditions of cholesterol of adult self-identified males between 35 and 46 is greater than 18.227.

e) What practical answer would you tell the researcher concerning the effect of aerobic exercise and a low fat diet on the cholesterol level?

I would think that 18.227 is practically greater than 0, now we have to check the effect size to be sure that the effect is measurable.

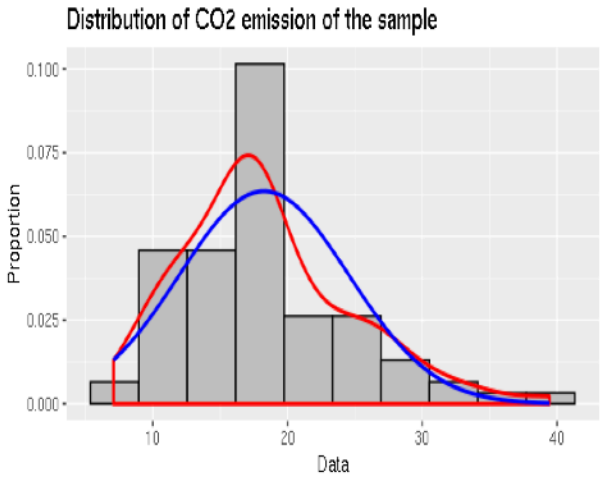
Effect Size

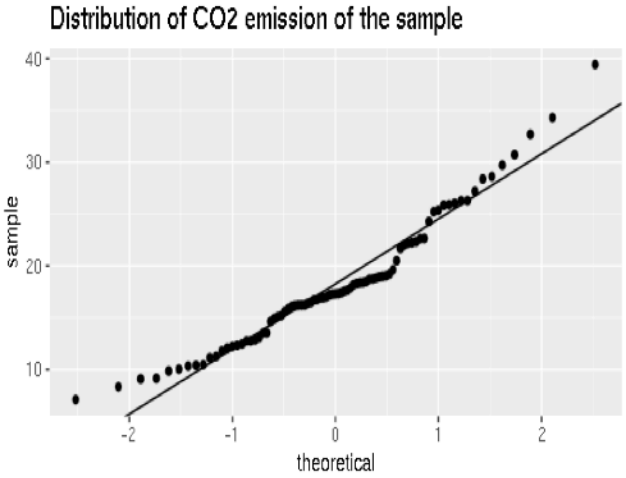
```
> sd(before-after)
[1] 19.14407
```

$$\frac{18.22722}{19.14407} = 0.952$$

This is relatively large, so I would believe that aerobic exercise and a low fat diet did practically decrease the cholesterol level in middle aged men because the difference is practically different and the effect is measurable.

11. A simple random sample of 85 automobiles was obtained and the CO_2 emissions from each was measured (in hectogram/mi). The following graphs were generated from the above sample.





Using the above figures, is the appropriate assumption valid so that you can perform statistical inference for the population mean CO_2 emission? Please explain your answer. Be sure to state the assumption that is being shown from the graphs.

The appropriate assumption that can be checked from the graphs is normality.

From the graphs the distribution is right skewed. Since there are no outliers and the sample size is 85 (which is greater than 40), we can say that the normality assumption is met because of CLT.