Conference 2

March. 25. 2019

Jee-Young Moon

Outline

- Sampling distribution of sample mean
- Hypothesis Testing and Confidence Interval
  - T-test
    - One sample t-test
    - Paired t-test
    - Two sample t-test – equal variance, unequal variance
  - Distribution-free test
- Study Design Issues
  - Type I error, Type II error, and sample size

Sampling distribution of sample mean

Suppose that we draw all possible **samples** of size n from a given population.
Suppose further that we compute a **sample mean** for each **sample**.
The probability distribution of these sample means is called a **sampling distribution of sample mean**.

P1. Explain the Central Limit Theorem (CLT).

P1. Explain the Central Limit Theorem (CLT).

Let $X_1, X_2, \ldots, X_n$ be an independent random sample of size n from a distribution with mean μ and standard deviation σ (which may or may not be normal).
If a sample size (n) is large enough (about 30 or more), the distribution of sample mean ($\bar{X}$) is approximately a normal distribution with mean μ and standard deviation = $\frac{\sigma}{\sqrt{n}}$.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

P2. Why is it important that a sample drawn from a population be random?

P2. Why is it important that a sample drawn from a population be random?

If the sample is not drawn at random from the target population, the sample does not represent the population. For example, if you only select samples that are convenient to collect (e.g. reaching out by landline phone during the day to ask political opinion), the inference based on your sample will not represent the target population.

When the sample is drawn at random, you can use it to estimate the population parameter, and the sampling distribution of the estimator (based on CLT.)

P3. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid has mean 29.5 mg/100 ml and standard deviation 9.25 mg/100 ml. Suppose that you select repeated samples of size 50 from this population and calculate the mean for each sample.

a. If you were to select a large number of random samples of size 50, what would be the mean of the sample means?

b. What would be their standard deviation? What is another name for this standard deviation of the sample means?

c. How does the standard deviation of the sample means compare with the standard deviation of the albumin levels themselves?

d. If you were to take all the different sample means and use them to construct a histogram, what would be the shape of their distribution?

e. What proportion of the means of samples of size 50 is larger than 33 mg/100 ml?

f. What proportion of the means is between 29 and 30 mg/100 ml?

P3. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid has mean 29.5 mg/100 ml and standard deviation 9.25 mg/100 ml. Suppose that you select repeated samples of size 50 from this population and calculate the mean for each sample.

a. If you were to select a large number of random samples of size 50, what would be the mean of the sample means?

Mean of sample means would be the population mean, 29.5

b. What would be their standard deviation? What is another name for this standard deviation of the sample means?

The standard deviation of sample means is $\frac{9.25}{\sqrt{50}} = 1.31$

It is also called as a standard error.

P3. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid has mean 29.5 mg/100 ml and standard deviation 9.25 mg/100 ml. Suppose that you select repeated samples of size 50 from this population and calculate the mean for each sample.

c. How does the standard deviation of the sample means compare with the standard deviation of the albumin levels themselves?
The standard deviation ($\sigma/\sqrt{n}$) of the sample means is smaller than the standard deviation (σ) of albumin levels themselves as long as n>1.

d. If you were to take all the different sample means and use them to construct a histogram, what would be the shape of their distribution?
By CLT, it would look like a normal distribution with mean = 29.5, sd = 1.31.

P3. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid has mean 29.5 mg/100 ml and standard deviation 9.25 mg/100 ml. Suppose that you select repeated samples of size 50 from this population and calculate the mean for each sample.

e. What proportion of the means of samples of size 50 is larger than 33 mg/100 ml?

By CLT,

$\bar{X}$ ~ N(mean=29.5, sd=1.31)

$$P(\bar{X} > 33) = P\left(Z > \frac{33 - 29.5}{1.31}\right) = P(Z > 2.67)$$

$$= 1 - 0.9962 = 0.0038$$

> 1-pnorm(33, mean = 29.5, sd=9.25/sqrt(50))

P3. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid has mean 29.5 mg/100 ml and standard deviation 9.25 mg/100 ml. Suppose that you select repeated samples of size 50 from this population and calculate the mean for each sample.

f. What proportion of the means is between 29 and 30 mg/100 ml?

$$P(29 < \bar{X} < 30)$$
$$= P\left(\frac{29 - 29.5}{1.31} < Z < \frac{30 - 29.5}{1.31}\right)$$
$$= P(-0.38 < Z < 0.38) = 0.6480 - (1 - 0.6480) = 0.296$$

```
> pnorm(30, mean = 29.5, sd=9.25/sqrt(50)) - pnorm(29, mean = 29.5, sd=9.25/sqrt(50))
```

For statistical inference on mean

If population distribution is normal or sample size is large

- when σ is known,

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

A normal distribution is used.

- when σ is unknown,

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

A t-distribution is used.

P4. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners expend an average of 42 kilowatt-hours per year with a standard deviation of 11.9 kilowatt-hours, does this suggest at the 0.05 level of significance that vacuum cleaners expend, on the average, less than 46 kilowatt-hours annually? Assume the population of kilowatt-hours to be normal.

P4. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners expend an average of 42 kilowatt-hours per year with a standard deviation of 11.9 kilowatt-hours, does this suggest at the 0.05 level of significance that vacuum cleaners expend, on the average, less than 46 kilowatt-hours annually? Assume the population of kilowatt-hours to be normal.

H_0: μ = 46 kilowatt-hours

H_A: μ < 46 kilowatt-hours

We perform a one-sided test.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{42-46}{11.9/\sqrt{12}} = -1.16. \text{ (d.f. = 11)}$$

p-value = P(T< -1.16)=0.135 > 0.05 (significance level)

Conclusion: At the significance level at 0.05, we do not have enough evidence to reject the null hypothesis.

Blood cholesterol levels among heart attack survivors

A study was conducted at a major north eastern American medical center regarding blood cholesterol levels and heart-attack incidents. A total of 28 heart-attack patients had their cholesterol levels measured two days, 4 days, and 14 days after the attack. In addition, cholesterol levels were recorded for a control group of 30 people who had not had a heart attack. The units of cholesterol measurement are mg/dL of blood. Download the data cholestg.txt from course website.

P5. Inference on a population mean cholesterol level 2 days after the heart attack

A. Draw an appropriate plot to summarize the cholesterol level of patients 2 days after the heart attack. What are mean and standard deviation of the sample?
B. What is the 99% confidence interval for the mean cholesterol level of patients 2 days after the heart attack?
C. Perform a two-sided hypothesis test to see if the cholesterol level of patients 2 days after the heart attack is different from 200 mg/dL. What is the p value of this test? What can you conclude from it using significance level 0.01?
D. Perform a one-sided hypothesis test to see if the cholesterol level of patients 2 days after the heart attack is greater than 200 mg/dL at the significance level 0.01. What is the one-sided 99% confidence interval?
E. What are the conditions you need to check for problems B-D?

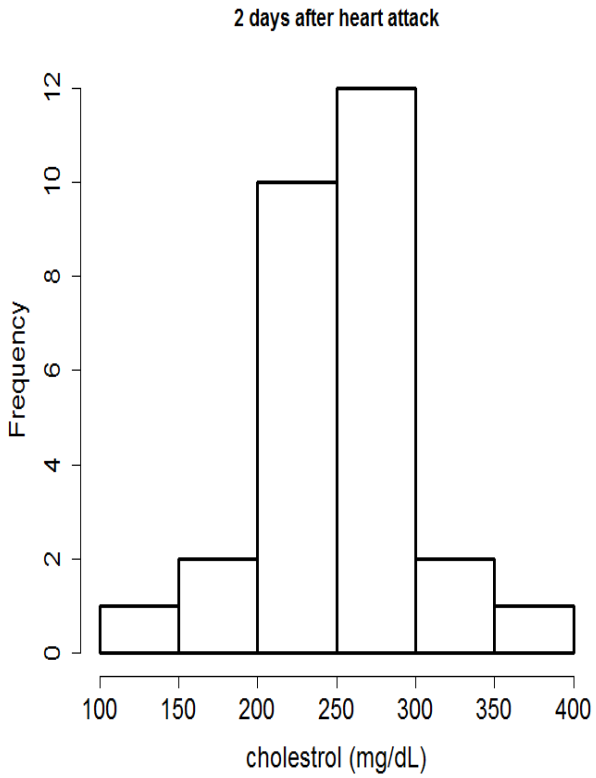
P5. Inference on a population mean cholesterol level 2 days after the heart attack

A. Draw an appropriate plot to summarize the cholesterol level of patients 2 days after the heart attack. What are mean and standard deviation of the sample?

```
> dt<-read.delim("cholestg.txt",header=T)
> x<-dt[dt$group==1 & dt$day==2,]$cholest
> hist(x,main="2 days after heart attack",
 xlab="cholestrol (mg/dL)")
```

```
> mean(x)
253.9286
> sd(x)
47.71049
```

Sample mean is 253.93 mg/dL and
sample standard deviation is 47.71 mg/dL.



P5. Inference on a population mean cholesterol level 2 days after the heart attack

B. What is the 99% confidence interval for the mean cholesterol level of patients 2 days after the heart attack?

We will use $(\bar{X} - t_{n-1,\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}})$ for the confidence interval for the population mean cholesterol level.

With sample size = 28 and two-sided 99% CI, the critical value to find from T-distribution is at 99.5% and degrees of freedom = 27.

99% CI= $\left(253.93 - t_{27,0.995}\frac{47.71}{\sqrt{28}}, 253.93 + t_{27,0.995}\frac{47.71}{\sqrt{28}}\right)$

= $\left(253.93 - 2.7707 \times \frac{47.71}{\sqrt{28}}, 253.93 + 2.7707 \times \frac{47.71}{\sqrt{28}}\right)$

= (228.95 mg/dL, 278.91 mg/dL)

> t.test(x, conf.level=.99)

```
> t.test(x, conf.level=.99)

        One Sample t-test

data:  x
t = 28.163, df = 27, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 228.9469 278.9103
sample estimates:
mean of x
 253.9286
```

P5. Inference on a population mean cholesterol level 2 days after the heart attack

C. Perform a two-sided hypothesis test to see if the cholesterol level of patients 2 days after the heart attack is different from 200 mg/dL. What is the p value of this test? What can you conclude from it using significance level 0.01?

We will use a one-sample two-sided t-test.

H_0: $\mu = 200$

H_A: $\mu \neq 200$

$T = \frac{253.93-200}{47.71/\sqrt{28}} = 5.98$ with degrees of freedom $= 27$

The critical value for T statistic of significance level 0.01 is $t_{27,\,\mathbf{0.995}} = 2.7707$

```
> qt(.995, df=27)
```

➔ $|T| = 5.98 > t_{27,\,0.995} = 2.7707$

P-value = $P(|T| > 5.98) < 0.002$ from the t-distribution table.

```
> 2*(1-pt(5.98, df=27))
2.230178e-06   ## P-value is 2×10^-6
```

Conclusion: With significance level 0.01, we reject the null hypothesis and conclude that the mean cholesterol level 2 days after heart attack is significantly different from 200 mg/dL.

R code to perform one-sample t-test (two-sided)

```
> x<-dt[dt$group==1 & dt$day==2,]$cholest
> t.test(x,mu=200,conf.level=.99)
```

```
        One Sample t-test

data:  x
t = 5.9811, df = 27, p-value = 2.223e-06
alternative hypothesis: true mean is not equal to 200
99 percent confidence interval:
 228.9469 278.9103
sample estimates:
mean of x
 253.9286
```

P5. Inference on a population mean cholesterol level 2 days after the heart attack

D. Perform a one-sided hypothesis test to see if the cholesterol level of patients 2 days after the heart attack is greater than 200 mg/dL at the significance level 0.01. What is the one-sided 99% confidence interval?

H_0: $\mu = 200$

H_A: $\mu > 200$

$T = \frac{253.93-200}{47.71/\sqrt{28}} = 5.98$ with degrees of freedom $= 27$

The critical value for T statistic of significance level 0.01 is $t_{27,\ \mathbf{0.99}} = 2.4727$

➔ $|T| = 5.98 > t_{27,\ 0.99} = 2.4727$

```
> 1-pt(5.98, df=27)
1.115089e-06   ## P-value is 1×10^-6
```

Conclusion: With significance level 0.01, we reject the null hypothesis and conclude that the mean cholesterol level 2 days after heart attack is significantly greater than 200 mg/dL.

P5. Inference on a population mean cholesterol level 2 days after the heart attack

D. Perform a one-sided hypothesis test to see if the cholesterol level of patients 2 days after the heart attack is greater than 200 mg/dL at the significance level 0.01. What is the one-sided 99% confidence interval?

99% one-sided confidence interval on population mean cholesterol level is

$\mu > 253.9286 - t_{27,0.99}\frac{47.71}{\sqrt{28}}$ ➔ $\mu >$

$253.9286 - 2.4727 \times \frac{47.71}{\sqrt{28}}$

➔ $\mu > 231.63\ mg/dL$

> t.test(x, alt = 'greater', mu=200, conf.level=.99)

```
        One Sample t-test

data:  x
t = 5.9811, df = 27, p-value = 1.112e-06
alternative hypothesis: true mean is greater than 200
99 percent confidence interval:
 231.634     Inf
sample estimates:
mean of x
 253.9286
```

P5. Inference on a population mean cholesterol level 2 days after the heart attack

E. What are the conditions you need to check for problems B-D?

In order to do a one-sample t test, we need to check

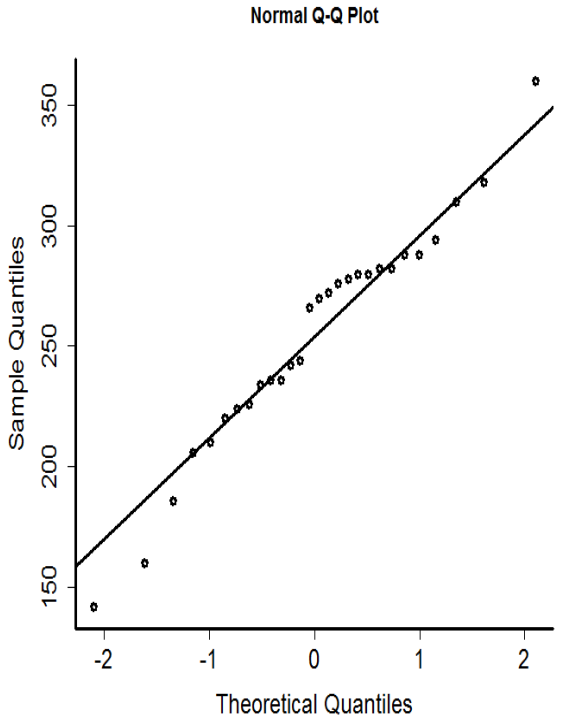
1. The samples are independent
2. The sample size is large (n>30) or cholesterol level is normally distributed.

To check the normality:

> hist(x)

> qqnorm(x)

> qqline(x)



- Distribution-free test: Wilcoxon signed rank sum test

> wilcox.test(x, mu=200)

```
> wilcox.test(x, mu=200)

        Wilcoxon signed rank test with continuity correction

data:  x
V = 380, p-value = 5.828e-05
alternative hypothesis: true location is not equal to 200

Warning message:
In wilcox.test.default(x, mu = 200) :
  cannot compute exact p-value with ties
```

> wilcox.test(x, mu=200, alt=‘;greater’)

```
> wilcox.test(x, mu=200, alt='greater')

        Wilcoxon signed rank test with continuity correction

data:  x
V = 380, p-value = 2.914e-05
alternative hypothesis: true location is greater than 200

Warning message:
In wilcox.test.default(x, mu = 200, alt = "greater") :
  cannot compute exact p-value with ties
```

P6. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

A. Draw an appropriate plot to compare the how the samples from two populations are different.

B. What is the 99% confidence interval for the mean difference in the cholesterol level?

C. If the null hypothesis of a hypothesis test is that the cholesterol level of patients 14 days after the heart attack is same as that of control (healthy patients), what is the p value of this test? What can you conclude from it using significance level 0.01?

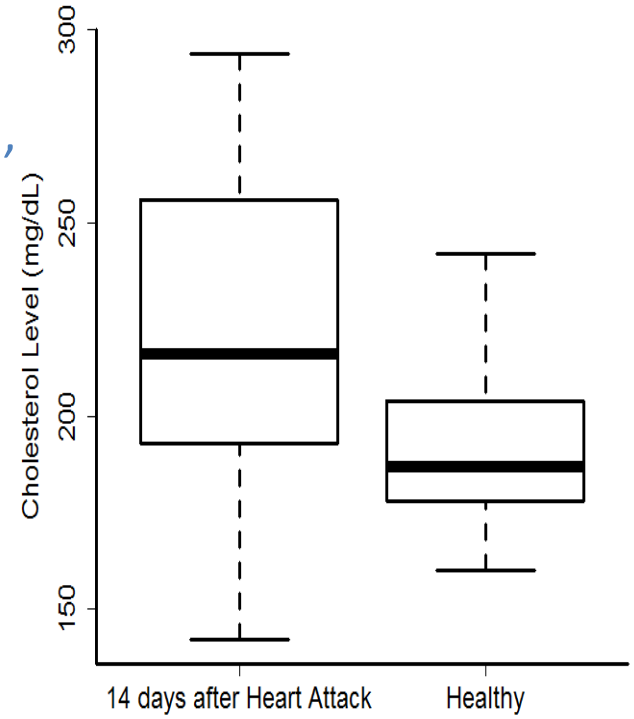
D. Check the conditions for problem B and C.

E. Assume that equal variances for cholesterol levels in both groups. Perform a two-sided test at the significance level 0.01 and estimate the 99% confidence interval for the mean difference in cholesterol level?

P6. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

A. Draw an appropriate plot to compare the how the samples from two populations are different.

```
> x<-subset(dt, group==1 & day==14)$cholest
> y<- subset(dt, group==2)$cholest
> boxplot(x,y,names=c("14 days after Heart Attack",
"Healthy"), ylab='Cholesterol Level (mg/dL)')
```




```
> mean(x, na.rm=T)
> sd(x, na.rm=T)
> length(x)
> sum(is.na(x))
```

P6. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

B. What is the 99% confidence interval for the mean difference in the cholesterol level?

$$(\bar{Y}_1 - \bar{Y}_2) - t_{v,\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s^2}{n_2}} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{v,\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s^2}{n_2}}$$

$$\bar{Y}_1 = 221.4737,\ \bar{Y}_2 = 193.133, n_1 = 19, n_2 = 30$$

$$s_1^2 = 1864.819, \quad s_2^2 = 497.292$$

v=24.16568

$t_{24,0.995}$ = 2.7969

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

$$-1.62mg/dL < \mu_1 - \mu_2 < 58.30mg/dL$$

> t.test(x, y, conf.level=.99)

```
        Welch Two Sample t-test

data:  x and y
t = 2.6459, df = 24.166, p-value = 0.01411
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -1.600077 58.280778
sample estimates:
mean of x mean of y
 221.4737  193.1333
```

P7. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

C. If the null hypothesis of a hypothesis test is that the cholesterol level of patients 14 days after the heart attack is same as that of control (healthy patients), what is the p value of this test? What can you conclude from it using significance level 0.01?

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

$$T = \frac{(\overline{Y_1} - \overline{Y_2}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 2.6459 < 2.7969$$

The p value is 0.01411. > 2(1-pt(2.6459, df=24.16568))*

The conclusion is that with significance level 0.01, we do not have enough evidence to reject the null hypothesis that the mean cholesterol levels are same between two groups.

Note that we would reject the null hypothesis if we increase the level of significance to 0.05(or decrease confidence level to 0.95).

> t.test(x, y, conf.level=.99)

```
        Welch Two Sample t-test

data:  x and y
t = 2.6459, df = 24.166, p-value = 0.01411
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -1.600077 58.280778
sample estimates:
mean of x mean of y
 221.4737  193.1333
```

P7. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

D. Check the conditions for problem B and C.

In order to do two-sample t test, we need to check

1. *Both samples are independent*
2. *All units in each sample are independent*
3. *Sample sizes in both groups are large (n>30) or both populations have normal distributions.*

P7. Now, we are interested in comparing the cholesterol level of patients 14 days after the heart attack with the control group (i.e. healthy people).

E. Assume that equal variances for cholesterol levels in both groups. Perform a two-sided test at the significance level 0.01?

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

$$T = \frac{(\overline{Y_1} - \overline{Y_2}) - 0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} = 3.025 > t_{47,.995} = 2.6846$$

> qt(0.995, df=47)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

The p value is 0.004. > 2*(1-pt(3.025, df=47))

The conclusion is that with significance level 0.01, we reject the null hypothesis and conclude that the mean cholesterol levels are significantly different between two groups.

```
> t.test(x,y, conf.level= .99, var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = 3.025, df = 47, p-value = 0.004022
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
  3.189578 53.491124
sample estimates:
mean of x mean of y
 221.4737  193.1333
```

Distribution-free test: Mann-Whitney rank sum test

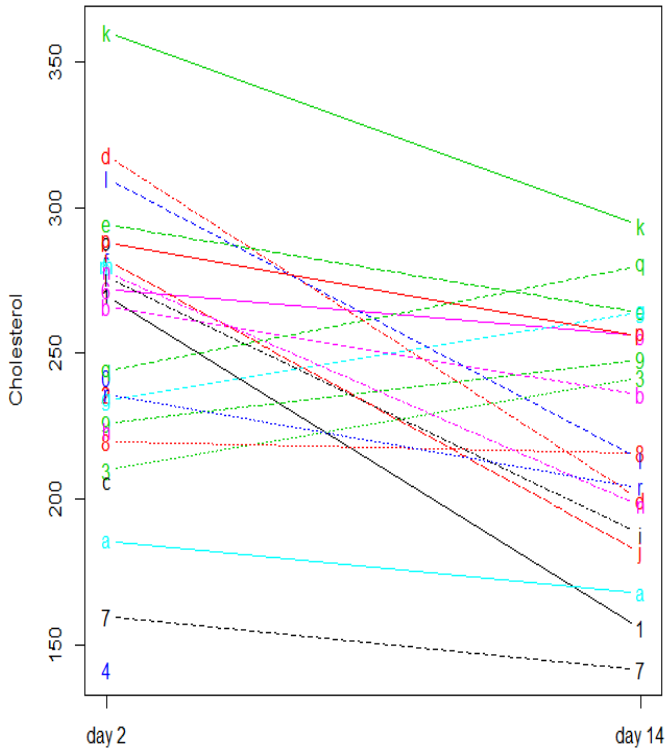
> wilcox.test(x,y)

```
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 409.5, p-value = 0.01083
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x, y) : cannot compute exact p-value with ties
```

P8. Now, we investigate the change of cholesterol level between 2 days and 14 days after heart attack.

A. What is the 99% confidence interval for the mean difference in the cholesterol level? Perform a two sided hypothesis test to see if you can reject the null hypothesis that the cholesterol level of patients 14 days after the heart attack is same as that after 2 days. What is the p value of this test? What can you conclude from it using significance level 0.01?
B. What are the conditions you need to check for problem A?

P8. Now, we investigate the change of cholesterol level between 2 days and 14 days after heart attack.




```
> x<- subset(dt, group==1 &
day==2)$cholest
> y<- subset(dt, group==1 &
dt$day==14)$cholest
> dt3 <- data.frame(x=x, y=y)
> matplot( t(dt3), type='b',  xaxt='n',
ylab='Cholesterol')
> axis(side=1, at=1:2, labels=c('day
2', 'day 14'))
```

P8. Now, we investigate the change of cholesterol level between 2 days and 14 days after heart attack.

A. What is the 99% confidence interval for the mean difference in the cholesterol level? Perform a two sided hypothesis test to see if you can reject the null hypothesis that the cholesterol level of patients 14 days after the heart attack is same as that after 2 days.

We will use a paired t-test, which is equivalent to one-sample t-test on the difference.

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

$\overline{y-x} = -38, sd(y-x) = 50.37195, n = 19$

$T = \frac{-38-0}{50.37195/\sqrt{19}} = -3.288$ with degrees of freedom $= 18$

The critical value for T statistic of significance level 0.01 is $t_{18,\ \mathbf{0.995}} = 2.8609$

➔ $|T| = 3.288 > t_{18,\ 0.995} = 2.8609$

P-value = P(|T| > 3.288) = 0.004

```
> 2*(1-pt(3.288, df=18))
```

Conclusion: *The p value is 0.004* and 99% confidence interval is (-4.74mg/dL, -71.26mg/dL).

With significance level at 0.01, we reject the null hypothesis and conclude that there is a statistically significant evidence that the mean cholesterol levels after 2 and 14 days are different.

```
> t.test(y, x, paired=TRUE, conf.level=.99)

        Paired t-test

data:  y and x
t = -3.2883, df = 18, p-value = 0.004085
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -71.263597  -4.736403
sample estimates:
mean of the differences
                    -38
```

```
> t.test(y-x, conf.level=.99)

        One Sample t-test

data:  y - x
t = -3.2883, df = 18, p-value = 0.004085
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 -71.263597  -4.736403
sample estimates:
mean of x
      -38
```

P8. Now, we investigate the change of cholesterol level between 2 days and 14 days after heart attack.

B. What are the conditions you need to check for problem A?

1. Paired samples
2. The pairs are independent.
3. The number of matched pairs is large ($n > 30$) or the within-pair differences are normally distributed.

Distribution-free tests: Wilcoxon signed rank test

- wilcox.test(y-x)

```
> wilcox.test(y-x)

        Wilcoxon signed rank test with continuity correction

data:  y - x
V = 34, p-value = 0.01481
alternative hypothesis: true location is not equal to 0

Warning message:
In wilcox.test.default(y - x) : cannot compute exact p-value with ties
```

- wilcox.test(x,y, paired=TRUE)