STAT 40001/STAT 50001 Statistical Computing

Lecture 8

Department of Mathematics and Statistics

PURDUE
UNIVERSITY.
NORTHWEST

Confidence Interval

An interval estimator is a rule for calculating two numbers L and U that define an interval which you are fairly certain containing the parameter of interest.

The concept of fairly certain can be quantified using a statistical concept called the "Confidence coefficient" designated by $(1-\alpha)$ called the confidence level.

So, if θ be the parameter of interest based on the sample $X_1, X_2, ...X_n$ we are looking for L and U which are the functions of the sample such that we are $(1-\alpha)100\%$ confident that θ lies in the interval(L, U).

In summary, a 95% confidence interval can be interpreted to mean that if all possible samples of given size are taken from a population, 95% of the samples would produce intervals that captured the true population mean and 5% would not. So, the confidence interval is used to describe the method of construction rather than a particular interval.

for Loop

Often we need to repeat an action several times - sometimes over subjects in a data set.

```
> for(i in 1:n){
> the action to be repeated
}
```

for indicates that we're going to loop from a start index to an end index.
i is the index we're looping over
1 is our start index
n is the end index
{ opens the loop
} closes the loop.

Examples

```
> index<-NULL # (Initiates the variable;assigns NULL value)
> for(i in 1:4){
+ index<-c(index,factorial(i))
+ }
> index
[1]  1  2  6 24
```

Example

```
> for (i in 1:8) { print (i*i) }
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
[1] 49
[1] 64
```

Examples

```
> for(i in 1:10){
+ print(i+1)
+ }
#############################
x = 101:200
y = 1:100
z = rep(0,100) ### rep means repeat
for(i in 1:100){
z[i] = x[i] + y[i]
}
w = x + y
print(w-z)

for(i in 1:10){
for(j in 1:5){
print(i+j)
}
}
```

Examples- if statements

```
for(i in 1:10){
if(i==4)print(i)
}
#####################
for(i in 1:10){
if(i!=4)print(i)
}
##############################
for(i in 1:10){
if( i < 4)print(i)
}
#########################
for(i in 1:10){
if(i <=4)print(i)
}
#################################
for(i in 1:10){
if(i>=4) print(i)}
```

While function

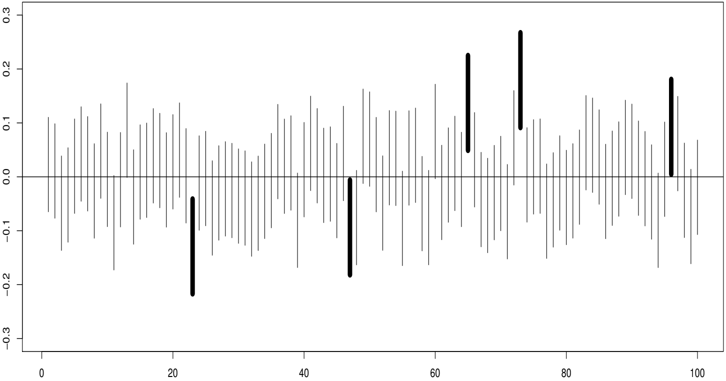
```
i=1
while(i<10){
print(i)
i=i+1
}
[1]  1
[1]  2
[1]  3
[1]  4
[1]  5
[1]  6
[1]  7
[1]  8
[1]  9
```

Confidence Interval

The function `interval.plot()` from **PASWR** library can be used to create the intervals.

```
set.seed(402)
m<-100 # Number of samples
n<-500 # Sample size
a<-array(0,m)
ll<-array(0,m)
ul<-array(0,m)
i<-0
while (i<m) {i<-i+1
a[i]<-mean(rnorm(n))
ll[i]<-a[i]+qnorm(0.025)*sqrt(1/n)
ul[i]<-a[i]+qnorm(0.975)*sqrt(1/n)}
interval.plot(ll,ul)
```

Confidence Interval



Central Limit Theorem

Generating from Uniform

```
plot(0,0,type="n",xlim=c(0,1),ylim=c(0,13.5),
xlab="Density", ylab="f(x)")
m=500;  a=0;  b=1
n=2
res<- mean(runif(n, a, b))
for ( i in 1:m) res[i]<- mean(runif(n, a, b))
lines( density(res), lwd=2)
```

Please repeat the above program by replacing n by 5, 10, 100

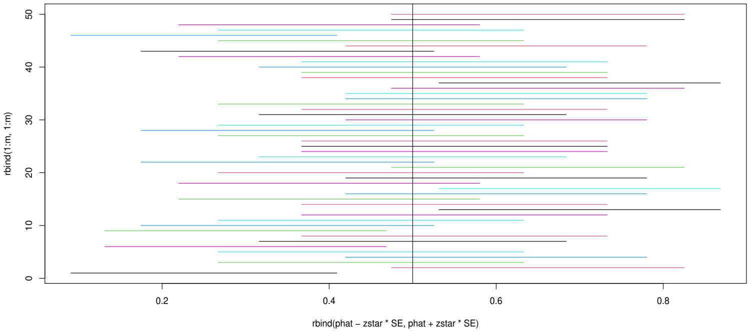
Generating from Gamma

```
plot(0,0, type="n",xlim= c(0,3), ylim=c(0,2),
xlab=" Density",ylab="f(x)")
m=500; a=1; b=1
n=2
res<- mean(rgamma(n, a, b))
for ( i in 1:m) res[i]<- mean(rgamma(n, a, b))
lines( density(res), lwd=2)
```

Please repeat the above program by replacing n by 5, 10, 100

Confidence Interval

Each CI is a Bernoulli trial: It either contains the true parameter value or not. After the CI is constructed, we talk about how “confident” we are that it contains the true value.



Confidence Interval

```
> m = 50; n=20; p = .5 # toss 20 coins 50 times
> phat = rbinom(m,n,p)/n # divide by n for proportions
> SE = sqrt(phat*(1-phat)/n) # compute SE
> alpha = 0.10; zstar = qnorm(1-alpha/2)
> matplot(rbind(phat - zstar*SE, phat + zstar*SE),
+ rbind(1:m,1:m),type="l",lty=1)
> abline(v=p) # draw vertical line at p=0.5
```

Confidence Interval

Note that $(1-\alpha)100\%$ confidence interval for μ is given by

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}. \qquad \text{If } \sigma \text{ is known}$$

$$\overline{X} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}. \qquad \text{If } \sigma \text{ is unknown}$$

Confidence Interval - Example

Using summary values

We can use zsum.test or tsum.test available in BSDA package: Basic Statistics and Data Analysis

Using raw data

- t.test available in base package
- z.test available in TeachingDemos package

Example: A student wants to estimate the mean amount of money spent by college students on books this semester. He surveys 50 randomly chosen students and calculates a sample mean of $\bar{X} = \$235$ and standard deviation of 75. Calculate a 95% confidence interval of the population mean.

```
> library(BSDA)
> zsum.test(mean.x =235, sigma.x =75, n.x = 50)$conf.int
[1] 214.2114 255.7886
attr(,"conf.level")
[1] 0.95
```

Confidence Interval - Example

Body weights of 9 girls selected at random from a highschool are as follows:
114, 100, 104, 94, 114, 105, 103, 105, 96
Calculate a 95% confidence for population mean μ

```
> weight=c(114, 100, 104, 94, 114, 105, 103, 105, 96)
> t.test(weight)$conf.int
[1]  98.57111 109.20667
attr(,"conf.level")
[1] 0.95
```

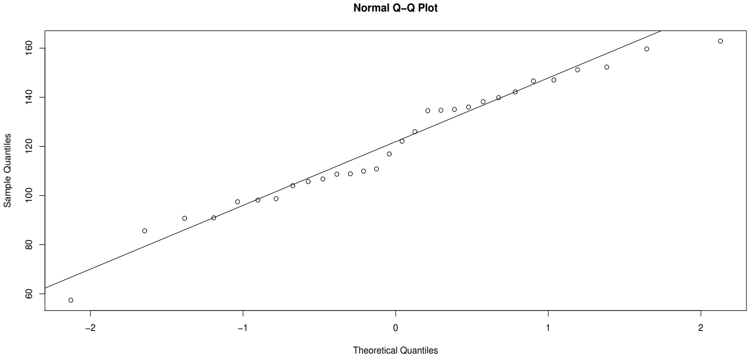
Confidence Interval- Example

The consumer expenditure survey, created by the U.S. Department of Labor, was administrated to 30 household. The amount of money each household spent per week is provided in the data frame `Grocery` in the PASWR package. Using the historical record of variance 900, Construct a 95% confidence interval for average expenditure.

```
library(PASWR)
attach(Grocery)
xbar <- mean(groceries)
Z <- qnorm(.975)
sigma <- 30
n <- 30
round(c(xbar-Z*sigma/sqrt(n),xbar+Z*sigma/sqrt(n)),2)
[1] 109.90 131.37
```

Assessing the Normality of the Data

```
library(PASWR)
attach(Grocery)
qqnorm(groceries)
qqline(groceries)
```



Confidence Interval- Example

The price for fourteen houses (three bedrooms/two-bathrooms) in Watauga County NC are provided in the data frame House in the PASWR package. Calculate a 95% confidence interval for the average price of a three bedroom/two bathroom house in this county.

```
library(PASWR)
attach(House)
MEAN<-mean(Price)
CT<-qt(.975,13)
ST<-sd(Price)
round(c(MEAN-CT*ST/sqrt(14), MEAN+CT*ST/sqrt(14)),2)
[1] 198.54 272.34
```

Assessing the Normality of the Data

```
library(PASWR)
attach(House)
qqnorm(House$Price)
qqline(House$Price)
```

Confidence Interval for the Difference in Population Means: Equal Variances

Suppose two normal populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, where $\sigma_X = \sigma_Y = \sigma$ is known. Take sample of sizes n_X and n_Y then $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X} - \overline{Y} \pm Z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

$$\left[(\overline{X} - \overline{Y}) - Z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, (\overline{X} - \overline{Y}) + Z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right]$$

Example

Example: Suppose independent random samples are taken from two normal distribution with the following summary

$$\begin{aligned} \bar{X} &= 4.00, \sigma_X = 3, n_X = 15 \\ \bar{Y} &= 4.41, \sigma_Y = 3, n_Y = 22 \end{aligned}$$

Calculate a 95% confidence interval for the difference in population means $(\mu_X - \mu_Y)$.

```
round(c((4-4.41)-qnorm(0.975)*3*sqrt(1/15+1/22),
(4-4.41)+qnorm(0.975)*3*sqrt(1/15+1/22)),2)
[1] -2.38  1.56
```

Example

The hardness of a piece of fruit is a good indicator of the fruits's ripeness. The data frame `Apple` in the **PASWR** package contains the hardness of 17 fresh apples and 17 warehouse stored apples. Assuming that both group have equal variances of 2.25, construct a 95% confidence interval for the difference in the mean hardness for fresh and warehoused apples.

Assessing Normality: Q-Q plot

```
library(PASWR)
attach(Apple)
fresh <- qqnorm(Fresh)
par(new=T)
old <-qqnorm(Warehouse)
plot(fresh,type="n",ylab="Sample Quantiles",
xlab="Theoretical Quantiles")
qqline(Fresh, col = 1)
qqline(Warehouse, col = 2)
points(fresh, col = 1, pch = 16, cex = 1.2)
points(old, col = 2, pch = 17,cex=1.2)
legend(-2, 9, c("Fresh", "Warehouse"),pch=c(16,17),
col = c(1, 2),title("Q-Q Normal Plots"))
```

Example- Apple hardness

```
> library(PASWR)
> attach(Apple)
> summary(Apple)
     Fresh           Warehouse
 Min.   :5.530   Min.   : 6.190
 1st Qu.:6.480   1st Qu.: 7.070
 Median :7.270   Median : 7.790
 Mean   :7.254   Mean   : 7.895
 3rd Qu.:8.090   3rd Qu.: 8.830
 Max.   :9.560   Max.   :10.500

mean.old<-mean(Warehouse)
mean.fresh<-mean(Fresh)
round(c(mean.fresh-mean.old - qnorm(0.975)*1.5*sqrt(2/17),
  mean.fresh - mean.old + qnorm(0.975)*1.5*sqrt(2/17)),2)
[1] -1.65  0.37
```

Confidence Interval for the Difference in Population Means: With Known and Unequal Variances

Suppose two normal populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, where $\sigma_X \neq \sigma_Y$ is known. Take sample of sizes n_X and n_Y then $(1-\alpha)100\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X} - \overline{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

$$\left[(\overline{X} - \overline{Y}) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, (\overline{X} - \overline{Y}) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right]$$

Example

Example: Suppose independent random samples are taken from two normal distribution with the following summary

$$\begin{aligned}\bar{X} &= 8.4, \sigma_X = 4.5, n_X = 50\\ \bar{Y} &= 8.8, \sigma_Y = 6, n_Y = 46\end{aligned}$$

Calculate a 95% confidence interval for the difference in population means $(\mu_X - \mu_Y)$.

```
round(c((8.4-8.8)-qnorm(0.975)*sqrt((4.5)^2/50+6^2/46),
(8.4-8.8)+qnorm(0.975)*3*sqrt((4.5)^2/50+6^2/46)),2)
```

```
[1] -2.54  6.01
```

Confidence Interval of means when the variances are unknown

Suppose two normal populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, where σ_X and $sigma_Y$ are unknown. The sampling distribution of $\bar{X} - \bar{Y}$ is students' t distribution with the standard statistic

$$t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{SE(\bar{X} - \bar{Y})}$$

Note that

$$SE(\bar{X} - \bar{Y}) = \begin{cases} \sqrt{s_p^2(1/n_X + 1/n_Y)} & \text{if} \sigma_X = \sigma_Y \\ \sqrt{s_X^2/n_X + s_Y^2/n_Y} & \text{if} \sigma_X \neq \sigma_Y \end{cases}$$

where, $s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$ is called the pooled variance. Note that the degrees of freedom is given by

$$\text{degrees of freedom} = \begin{cases} n_X + n_Y - 2 & \text{if} \sigma_X = \end{cases}$$

Confidence Interval

$(1-\alpha)100\%$ Confidence Interval for $\mu_X - \mu_y$ is
case 1: Unknown but equal variances so, $\sigma_X = \sigma_Y$

$$\left[(\bar{X}-\bar{Y}) - t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}, (\bar{X}-\bar{Y}) + t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\right]$$

where, $s_p^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}$ and $\nu = n_X + n_y - 2$.
case 2: Unknown and unequal variances so $\sigma_X \neq \sigma_Y$

$$\left[(\bar{X}-\bar{Y}) - t_{\alpha/2,\nu} \sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}, (\bar{X}-\bar{Y}) + t_{\alpha/2,\nu} \sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}\right]$$

where, $\nu = \left(\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}\right)^2 \times \left(\frac{(s_X^2/n_X)^2}{n_X-1}+\frac{(s_Y^2/n_Y)^2}{n_Y-1}\right)^{-1}$

Example

Suppose a random sample from a $N(\mu_X, \sigma)$ is taken where

$$n_X = 15, \sum_{i=1}^{15} x_i = 53, \text{and} \sum_{i=1}^{15} x_i^2 = 222.$$

Similarly another random sample is taken from a $N(\mu_Y, \sigma)$ population independent of the first sample such that

$$n_Y = 11, \sum_{i=1}^{11} y_i = 77, \text{and} \sum_{i=1}^{11} y_i^2 = 560$$

Obtain a 95% confidence interval for $\mu_X - \mu_Y$.
Solution: Note that

$$\bar{X} = 3.53, s_X^2 = 2.51, \bar{Y} = 7.00, s_Y^2 = 2.10$$

Now, using R we have

```
sp=sqrt((14*2.51+10*2.1)/24)
round(c((3.53-7)-qt(0.975,24)*sp*sqrt(1/15+1/11),
(3.53-7)+qt(0.975,24)*sp*sqrt(1/15+1/11)),2)
[1] -4.72 -2.22
```

Example

Suppose a random sample from a $N(\mu_X, \sigma_X)$ is taken where

$$n_X = 15, \sum_{i=1}^{15} x_i = 63, \text{and} \sum_{i=1}^{15} x_i^2 = 338.$$

Similarly another random sample is taken from a $N(\mu_Y, \sigma_Y)$ population independent of the first sample such that

$$n_Y = 11, \sum_{i=1}^{11} y_i = 66.4, \text{and} \sum_{i=1}^{11} y_i^2 = 486$$

Obtain a 95% confidence interval for $\mu_X - \mu_Y$.
Solution: Note that

$$\bar{X} = 4.2, s_X^2 = 5.24, \bar{Y} = 6.04, s_Y^2 = 8.47, \nu = \left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2 \times \left(\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}\right)^{-1} = 18.43 \approx 18$$

Now, using R we have

```
 round(c((4.2-6.04)-qt(0.975,18)*sqrt(5.24/15+8.47/11),
(4.2-6.04)+qt(0.975,18)*sqrt(5.24/15+8.47/11)),2)
[1] -4.06  0.38
```

Confidence Interval for Population Variance

Let $X_1, X_2, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$ and both are unknown. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

We want χ^2_L and χ^2_U such that

$$P(\chi^2_L \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_U) = 1 - \alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_U} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_L}\right) = 1 - \alpha$$

Hence, the $(1-\alpha)100\%$ CI for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_U}, \frac{(n-1)S^2}{\chi^2_L}\right).$$

But if $\chi^2_U = \chi^2_{\alpha/2}$ then $\chi^2_L = \chi^2_{1-\alpha/2}$. Therefore, $(1-\alpha)100\%$ CI for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right)$$

Example

Suppose 15 random sample is taken from a $N(\mu_X, \sigma_X)$. Suppose the sample variance is 5.24. Construct a 80% confidence interval for σ^2.

```
round(c(14*5.24/qchisq(0.9, 14), 14*5.24/qchisq(0.1,14)),2)
[1] 3.48 9.42
```

Confidence Interval for Population Proportion

We know that the maximum likelihood estimate of the population proportion p is the sample proportion $\hat{p}$. The asymptomatic properties of MLE allows that

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

as $n \to \infty$

Hence, $(1-\alpha)100\%$ confidence interval for p is

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

A more accurate confidence interval for p can be given as

$$\left[\frac{\hat{p} + \frac{Z^2_{\alpha/2}}{2n} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{Z^2_{\alpha/2}}{4n^2}}}{\left(1 + \frac{Z^2_{\alpha/2}}{n}\right)}, \frac{\hat{p} + \frac{Z^2_{\alpha/2}}{2n} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{Z^2_{\alpha/2}}{4n^2}}}{\left(1 + \frac{Z^2_{\alpha/2}}{n}\right)}\right]$$

Example

A researcher is interested in the the gender distribution of students in high school. He takes a random sample of 40 students and finds that only 26 are female. help him to construct 95% confidence intervals for the true proportion of female students.

```
> n<-40
> p=26/40
> z<-qnorm(0.975)
> round(c(p-z*sqrt(p*(1-p)/n), p+z*sqrt(p*(1-p)/n)),3)
[1] 0.502 0.798
```

The preferred method for smaller sample size yields

```
> n<-40
> p=26/40
> z<-qnorm(0.975)
> round(c((p+z^2/(2*n)-z*sqrt(p*(1-p)/n+z^2/(4*n^2)))/(1+z^2/n),
(p+z^2/(2*n)+z*sqrt(p*(1-p)/n+z^2/(4*n^2)))/(1+z^2/n)),3)
[1] 0.495 0.779
```