Inference for means

Marianne Huebner, Michigan State University

2015-05-01

## Confidence Interval for

The formula for a confidence interval for the sample mean is

where is the critical value from the t-distribution and depends on the confidence level. For example, to calculate a 95% confidence interval for 19 observations (= 18 degrees of freedom), the critical value is

qt(.975,18)

## [1] 2.100922

Note that to get a t interval and t test the same function is used. Type

?t.test

to check what options are available

#### Example: Detergents

The production of a nationally marketed detergent results in certain workers receiving prolonged exposure to the bacillus subtilis enzyme. Nineteen workers were tested to determine the effects of these exposures on various respiratory functions. The airflow rate, FEV1, is the ratio of a person’s forced expiratory volume to the vital capacity, VC (max. volume of air a person can exhale after taking a deep breath). If the enzyme has an effect, it will be to reduce the FEV1/VC ratio. The norm is 0.80 in persons with no lung dysfunction.

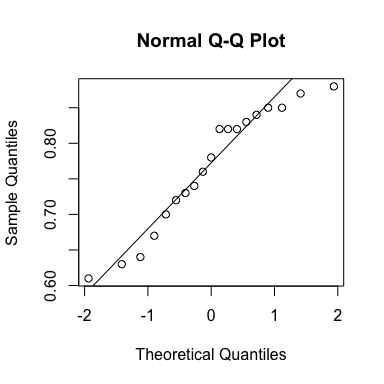
ratio<-c(0.61, 0.70, 0.63, 0.76, 0.67, 0.72, 0.64, 0.82, 0.88,   
0.82, 0.78, 0.84, 0.83, 0.82, 0.74, 0.85, 0.73, 0.85, 0.87)

There were measurements from 19 workers.

Assumptions for the t-interval are \* data are from a simple random sample \* independent \* identically distributed \* approximately normal distributed

The workers are likely unrelated and thus the independence assumption is satisfied. A Q-Q plot indicates the normal distribution of the data.

qqnorm(ratio)  
qqline(ratio)



A 90% confidence interval (t-interval) can be obtained with

temp<-t.test(ratio, mu=0.80, conf.level=0.90)  
temp

##   
## One Sample t-test  
##   
## data: ratio  
## t = -1.7091, df = 18, p-value = 0.1046  
## alternative hypothesis: true mean is not equal to 0.8  
## 90 percent confidence interval:  
## 0.7321399 0.8004917  
## sample estimates:  
## mean of x   
## 0.7663158

Specifically, the confidence interval is

temp$conf.int

## [1] 0.7321399 0.8004917  
## attr(,"conf.level")  
## [1] 0.9

The sample mean is

temp$estimate

## mean of x   
## 0.7663158

sd(ratio)

## [1] 0.08590761

#### Answer:

**A 90% confidence interval for the FEV1/VC ratio is (0.7321399, 0.8004917). This includes the true (normal) value of 0.8.**

A 95% confidence interval would be

t.test(ratio, mu=0.80, conf.level=0.95)$conf.int

## [1] 0.7249096 0.8077219  
## attr(,"conf.level")  
## [1] 0.95

## Hypothesis test for

For a hypothesis test on the population mean the null hypothesis is

with test statistic

The test statistic has t-distribution with degrees of freedom.

#### Detergent example continued

If the enzyme has an effect, it will be to reduce the FEV1/VC ratio. The norm is 0.80 in persons with no lung dysfunction. Hence the research question is whether the ratio is lower than normal. This is a one sided hypothesis test, namely the alternative is

temp.1sided<-t.test(ratio, mu=0.80, alternative="less")

#### Answer:

**The pvalue is 0.052307. The data provide evidence that exposure to B. subtilis may reduce the FEV1/VC ratio, but are inconclusive at the 5% significance level.**

## Coverage of confidence intervals

An aspirin manufacturer fills bottles by weight rather than by count. The weight per tablet should be 5 grains. Simulate a sample of 100 tablets taken from a very large lot. Suppose 20 tablets are sampled from a lot with mean weight 5 grains and standard deviation 0.3 grains. A 95% confidence interval is calculated. Does this confidence interval contain the true mean?

Take sims =100 such samples of 20 tablets:

sims = 100  
weights = replicate(sims,rnorm(20,mean=5,sd=0.3))

Calculate 95% confidence intervals for each sample:

tint<-matrix(NA, nrow=dim(weights)[2], ncol=2)  
for(i in 1:dim(weights)[2]){  
temp<-t.test(weights[,i], conf.level = 0.95)  
tint[i,]<-temp$conf.int  
}  
colnames(tint)<-c("lcl","ucl")

Check the first 10 confidence intervals from these samples:

tint[1:10,]

## lcl ucl  
## [1,] 4.957568 5.188872  
## [2,] 4.925218 5.250454  
## [3,] 4.969054 5.201837  
## [4,] 5.001427 5.237365  
## [5,] 4.867623 5.112377  
## [6,] 4.915731 5.141419  
## [7,] 4.839435 5.061189  
## [8,] 4.842963 5.119970  
## [9,] 4.874098 5.193608  
## [10,] 4.902349 5.106851

How many of these 100 confidence intervals contain the true mean?

tint = data.frame(tint)  
indx = (tint$lcl <=5) & (tint$ucl>=5)  
sum(indx)

## [1] 93

In this simulation 93 of the 100 confidence intervals (95% confidence level) contain the true mean. How many would you expect?

## Inference for comparing means

The confidence interval for comparing two means from independent samples is given by

The confidence interval for comparing two means from dependent samples is given by

where is from a t-distribution with degrees of freedom. The differences are treated as one sample, and is denoted as , is denoted as , and is denoted as . The index stands for "difference.""

### Example: Change of internet user per 100 people from 2001 to 2011

Internet users per 100 people downloaded from GAPMINDER (<http://www.gapminder.org/data/>) subset of 15 countries.

There are 15 countries in 2001 and 2011.

countries<-c("Argentina" ,"Australia","Belgium","Canada", "Egypt","France", "Germany",  
 "Greece", "India" , "Japan" , "Mexico" ,"Niger","Singapore", "United.Kingdom", "United.States")  
Y2011<-c(17.72058337, 63.02693628, 55.47689608, 71.59660113, 11.69839821, 41.39108117,  
 68.76941828, 24.17107186, 2.388075, 66.19826223, 17.21, 0.221341351, 61.00284566,  
 69.9749171, 68.26789985)  
Y2001<-c(9.780807285, 52.60563888, 31.05933918, 59.97791664, 0.838945611, 25.48284606,  
 31.66413443, 11.0173514, 0.660146377, 38.15162324, 7.038023117, 0.105185431,  
 40.08896306, 33.47495977, 49.18000685)

**Is there a sigificant change in internet users from 2001 to 2011 in these countries?**

If we consider the internet users per country as dependent samples the corresponding hypotheses are vs .

t.test(Y2001, Y2011, paired=TRUE)

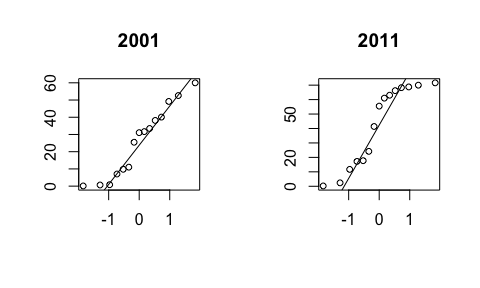
##   
## Paired t-test  
##   
## data: Y2001 and Y2011  
## t = -5.7256, df = 14, p-value = 5.244e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -22.72555 -10.33958  
## sample estimates:  
## mean of the differences   
## -16.53256

We could argue that a lot changes in 10 years and internet uses in 2001 are a different breed from internet users in 2011 and the number of internet users has increased, then we would be testing independent samples of no change versus the alternative hypothesis that in 2001 were fewer internet users with hypotheses vs .

t.test(Y2001, Y2011, alternative="less")

##   
## Welch Two Sample t-test  
##   
## data: Y2001 and Y2011  
## t = -1.8918, df = 25.757, p-value = 0.0349  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -1.622094  
## sample estimates:  
## mean of x mean of y   
## 26.07506 42.60762

To check the validity of this statistical methods for the data in question, we need to also check whether the data are approximately normally distributed.



There is a question whether the 2011 internet usage data follow a symmetric distribution for the selected countries.

Optional: Look at the data. Which countries changed the most?

plot(Y2001, Y2011, xlab="Internet users per 100 in 2001", ylab="Internet users per 100 in 2011", xlim=c(-1, 70),  
 ylim=c(0,80),col= "blue", pch = 19, cex = 1, lty = "solid", lwd = 2)  
text(Y2001, Y2011, labels=countries, cex= 0.7, pos=3)  
abline(a=0,b=1)

