Inference for proportions

### Example: Coin tosses

Simulate 100 coin tosses.

heads <- rbinom(100, size = 1, prob = .5)  
heads

## [1] 1 0 1 0 1 0 0 1 1 1 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 0 0 1 1 1 1 1 0 0  
## [36] 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 0 1  
## [71] 0 1 0 1 1 1 1 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 0 1 1 0 1 1 1

x=sum(heads) #number of heads in the 100 coin tosses

There were x= 48 heads in 100 coin tosses.

## Calculating confidence intervals for

We generated the sequence of heads and tails with probability of head = 0.5. When estimating confidence intervals for , is 0.5 part of the interval? The following one sample proportion interval uses the normal approximation to the binomial distribution with or without the continuity correction.

prop.test(x, 100) # continuity correction TRUE by default

##   
## 1-sample proportions test with continuity correction  
##   
## data: x out of 100, null probability 0.5  
## X-squared = 0.09, df = 1, p-value = 0.7642  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3798722 0.5816817  
## sample estimates:  
## p   
## 0.48

prop.test(x, 100, correct = FALSE) # without continuity correction

##   
## 1-sample proportions test without continuity correction  
##   
## data: x out of 100, null probability 0.5  
## X-squared = 0.16, df = 1, p-value = 0.6892  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3846455 0.5768342  
## sample estimates:  
## p   
## 0.48

prop.test(x, 100, conf.level=0.95) # confidence level of 0.95 is the default

##   
## 1-sample proportions test with continuity correction  
##   
## data: x out of 100, null probability 0.5  
## X-squared = 0.09, df = 1, p-value = 0.7642  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3798722 0.5816817  
## sample estimates:  
## p   
## 0.48

We cannot guarantee that 95% of all confidence intervals calculated in this way contain the true probability . This is called "coverage." For small samples the Agresti-Coull interval has better coverage. Here we add 2 successes and 2 failures to each sample:

There are many others types of one sample proportion confidence intervals:

library(binom)  
binom.confint(x, 100)

## method x n mean lower upper  
## 1 agresti-coull 48 100 0.480000 0.3846428 0.5768370  
## 2 asymptotic 48 100 0.480000 0.3820802 0.5779198  
## 3 bayes 48 100 0.480198 0.3835587 0.5770551  
## 4 cloglog 48 100 0.480000 0.3794099 0.5735755  
## 5 exact 48 100 0.480000 0.3790055 0.5822102  
## 6 logit 48 100 0.480000 0.3840603 0.5774378  
## 7 probit 48 100 0.480000 0.3836490 0.5775398  
## 8 profile 48 100 0.480000 0.3835206 0.5774935  
## 9 lrt 48 100 0.480000 0.3835189 0.5774938  
## 10 prop.test 48 100 0.480000 0.3798722 0.5816817  
## 11 wilson 48 100 0.480000 0.3846455 0.5768342

## Confidence intervals and hypothesis tests for

Save the output in an object named temp.

temp<-prop.test(x, 100)

names(temp) shows us what output is generated and the names of the output objects.

Confidence interval for the proportion of heads

temp$conf.int

## [1] 0.3798722 0.5816817  
## attr(,"conf.level")  
## [1] 0.95

Estimated proportion

temp$estimate

## p   
## 0.48

P value for the hypothesis test versus

temp$p.value

## [1] 0.7641772

The exact test is based on binomial probabilities:

binom.test(sum(heads),100)

##   
## Exact binomial test  
##   
## data: sum(heads) and 100  
## number of successes = 48, number of trials = 100, p-value = 0.7644  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3790055 0.5822102  
## sample estimates:  
## probability of success   
## 0.48

## Coverage of confidence intervals

For a sample size 30 and true probability 0.5, how many one proportion z intervals with 95% confidence would contain the true p=0.5?

n=30  
p=0.5  
sample<-rbinom(n, 1, p) # simulate n trials of successes and failures  
x<-sum(sample) # number of success in n trials  
result <- prop.test(x, n, conf.level=0.95) # calculate a 95% confidence interval

Next check whether the true p lies within the confidence interval. If yes, return "1". If no, return "0"

r count=0 # 1= yes, the confidence interval contains p, 0 otherwise. lb <- result$conf.int[1] ub <- result$conf.int[2] if(p < ub & p > lb){ count=1 } else { count=0 }

Repeat this with 1000 samples. We can do this over and over again, of we write a function.

sim<-function(t=1000, n=100, p=0.3, clevel=0.95){  
count<-rep(NA, t)  
  
for(i in 1:t){  
sample<-rbinom(n, 1, p)  
x<-sum(sample)  
result <- prop.test(x, n, conf.level=clevel)  
 lb <- result$conf.int[1]  
 ub <- result$conf.int[2]  
 if(p < ub & p > lb){  
 count[i]=1  
 } else {  
 count[i]=0  
 }  
}  
  
return(100\*sum(count)/t)  
}

Let's check for 1000 intervals.

t=1000  
pct<-sim(t=1000, n=30, p=0.5, clevel=0.95)

Among the 1000 95% confidence intervals based on 1000 samples there were pct = 95.2 % confidence intervals that contained the true .

Is this what you would expect?

What about 80% confidence interval coverage?

sim(t=1000, n=30, p=0.5, clevel=0.8)

## [1] 89.2

## Inference for comparing two proportions

### Example Cat Bites:

Among adults with cat bites, is the proportion of women bitten higher among patients with depression compared to without depression? More often than women without depression?

*Reference: Hanauer DA, Ramakrishnan N, Seyfried LS (2013) Describing the Relationship between Cat Bites and Human Depression Using Data from an Electronic Health Record. PLoS ONE 8(8): e70585.* [*doi:10.1371/journal.pone.0070585*](doi:10.1371/journal.pone.0070585) *University of Michigan, Ann Arbor.*

In the dataset extracted from electronic health records during a 3 year period in southeastern Michigan there were 750 patients seeking medical care for cat bites. Among those 310 suffered from depression.

bites<-matrix(c(258, 306, 310-258, 440-306 ), 2,2)  
colnames(bites)<-c("female", "male")  
rownames(bites)<-c("depression", "no depression")  
bites

## female male  
## depression 258 52  
## no depression 306 134

2-sample test for equality of proportions with continuity correction sample 1 = depression, sample 2 = no depression

prop.test(bites,conf.level=0.95)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: bites  
## X-squared = 17.5245, df = 1, p-value = 2.836e-05  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.07422893 0.19937811  
## sample estimates:  
## prop 1 prop 2   
## 0.8322581 0.6954545

The results indicate that among adults with catbites, the proportion of women are higher in patients with depression than without depression.

### Extracting confidence intervals and results from hypothesis tests from the output

temp<-prop.test(bites,conf.level=0.95)  
# names(temp) shows us what output values are available

Confidence interval for the difference in the proportions for =proportion of women in sample 1 (depression), =proportion of women in sample 2 (no depression)

temp$conf.int

## [1] 0.07422893 0.19937811  
## attr(,"conf.level")  
## [1] 0.95

Sample proportions:

temp$estimate

## prop 1 prop 2   
## 0.8322581 0.6954545

P-value of the hypothesis test versus :

temp$p.value

## [1] 2.836344e-05