Chapter 13 - The Theory of Confidence Intervals

Laura Saba

July 14, 2017

What this chapter covers

- CI of a mean via the t distribution
- CI of a mean via resampling
- CI of a proportion via resampling
- CI of a proportion via binomial distribution

Probability theory vs. statistical analysis

- probability theory: starts with a population then computes probabilities of various samples
- statistical analysis: starts with data (i.e., sample) then computes likelihood of different populations

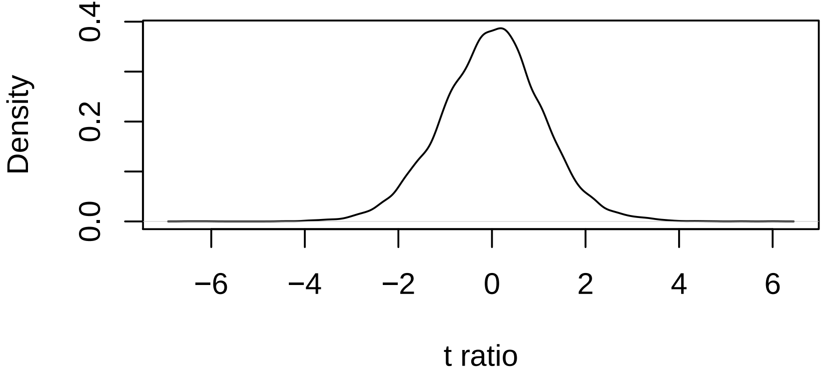
CI OF A MEAN VIA THE t DISTRIBUTION

- Assume that we know that the population of values follow a Gaussian distribution (mean $= \mu$ and sd $= \sigma$)
- Take a random sample of values, calculate the sample mean (m), the sample standard deviation (s), and the t ratio $t = \frac{m-\mu}{s/\sqrt{n}}$
- Repeat this random sampling many times

Calculating t ratios

```
#population parameters
mu=0
sigma=1
#number of samples drawn
n=12
#draw random sample
random_sample = rnorm(n,mean=mu,sd=sigma)
#calculate t ratio
t_ratio = (mean(random_sample) - mu)/
  (sd(random_sample)/sqrt(n))
#repeat many times
get_t = function(x) (mean(x) - mu)/(sd(x)/sqrt(n))
t_values=replicate(10000,
                   get_t(rnorm(n,mean=mu,sd=sigma)))
```

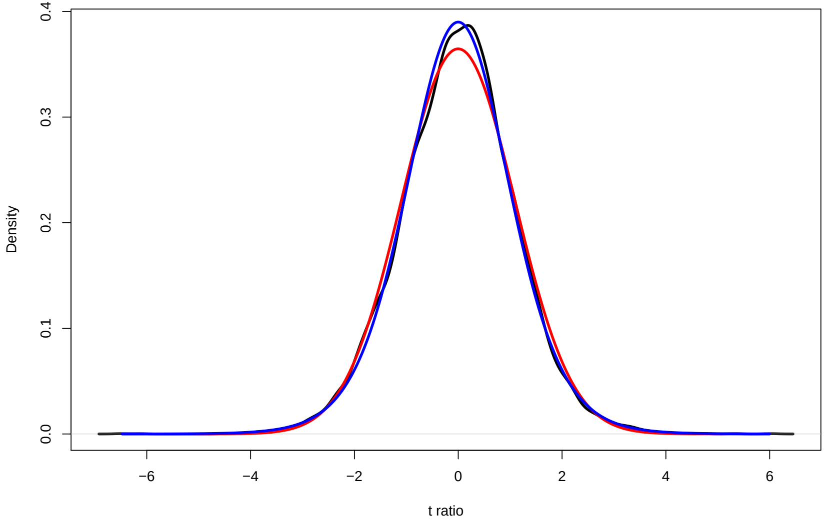
Distribution of t ratios




```
## Critical values of t ratio
cl=0.95
quantile(t_values,c(0.025,0.975))
```

```
##       2.5%     97.5%
## -2.191862  2.145175
```

Comparison of t-distribution and normal distribution



Critical value of t

Luckily we don't need to do simulation every time and can get an answer from R when we know the degrees of freedom (n-1).

```r
qt(0.025,df=11)
```

```
## [1] -2.200985
```

```r
qt(0.975,df=11)
```

```
## [1] 2.200985
```

The flip!

CI OF A MEAN VIA RESAMPLING

What if you cannot support the assumption of normality?

- Resampling (i.e., bootstrapping) is an alternative approach that doesn't assume normality.
  - Create many pseudosamples via resampling with replacement
  - For each pseudosample, calculate the mean
  - From the means of all pseudosamples, determine the 2.5th percentile and the 97.5th percentile (for 95% CI)
- The only assumption of resampling is that values are representative of the populations and that they vary independently.

CI OF A MEAN VIA RESAMPLING - in R

```
bodyTemp = c(37.0,36.0,37.1,37.1,36.2,37.3,
             36.8,37.0,36.3,36.9,36.7,36.8)
bs_means = replicate(10000,
                     mean(sample(bodyTemp,
                                 size=12,
                                 replace=TRUE)))
quantile(bs_means,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 36.54167 36.96667
```

```
t.test(bodyTemp)$conf.int
```

```
## [1] 36.51204 37.02130
## attr(,"conf.level")
## [1] 0.95
```

CI OF A PROPORTION VIA RESAMPLING

From Chapter 4, when polled 33 of 100 people said they would vote a certian way. What is the 95% confidence interval for this proportion?

```r
obs = c(rep("yes",33),rep("no",67))
bs_prop = replicate(10000,
  sum(sample(obs,size=100,replace=TRUE)=="yes")/100)
quantile(bs_prop,c(0.025,0.975))
```

```
##  2.5% 97.5%
##  0.24  0.42
```

With binomial data, there is no real advantage to the resampling approach over the distribution approach.

CI OF A PROPORTION VIA BINOMIAL DISTRIBUTION

```r
qbinom(p=0.975,size=100,prob=0.33)/100  #upper limit
```

```
## [1] 0.42
```

```r
qbinom(p=0.025,size=100,prob=0.33)/100  #lower limit
```

```
## [1] 0.24
```

CI of a proportion via R (normal approximation)

```r
prop.test(x=33,n=100)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  33 out of 100, null probability 0.5
## X-squared = 10.89, df = 1, p-value = 0.0009668
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2411558 0.4320901
## sample estimates:
##    p
## 0.33
```

What did we learn

- You can understand confidence intervals without understanding how they are computed.
- The math works by flipping around (solving) equations that predict samples from a known population in order to let you make inferences about the population from a single sample.
- An alternative approach is to use resampling (i.e., bootstrapping) methods.