

Confident in confidence intervals?

Peter Ralph

6 October 2020 – Advanced Biological Statistics

1

t distribution reminder

2.1

Recall our AirBnB example:

```
airbnb <- read.csv("../Datasets/portland-airbnb-listings.csv")
airbnb$price <- as.numeric(gsub("$", "", airbnb$price, fixed=TRUE))
airbnb$instant_bookable <- (airbnb$instant_bookable == "t")
instant <- airbnb$price[airbnb$instant_bookable]
not_instant <- airbnb$price[!airbnb$instant_bookable]
(tt <- t.test(instant, not_instant))
```

```
##
## Welch Two Sample t-test
##
## data: instant and not_instant
## t = 3.6482, df = 5039.8, p-value = 0.0002667
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.475555 14.872518
## sample estimates:
## mean of x mean of y
## 124.6409 114.9668
```

2.2

How's the t test work?

The *central limit theorem*.

2 . 3

In words:

The number of *standard errors* that the *sample mean* is away from the *true mean* has a t distribution.

2 . 4

In words:

The number of *standard errors* that the *sample mean* is away from the *true mean* has a t distribution.

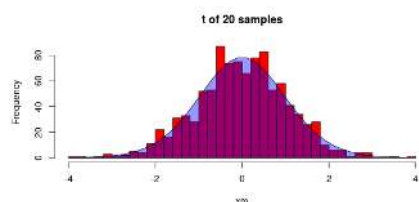
- ... with $n - 2$ degrees of freedom.
- "standard error" = s / \sqrt{n} = SD of the sample mean

2 . 4

In words:

The number of *standard errors* that the *sample mean* is away from the *true mean* has a *t* distribution.

- ... with $n - 2$ degrees of freedom.
- “standard error” = s/\sqrt{n} = SD of the sample mean



2 - 4

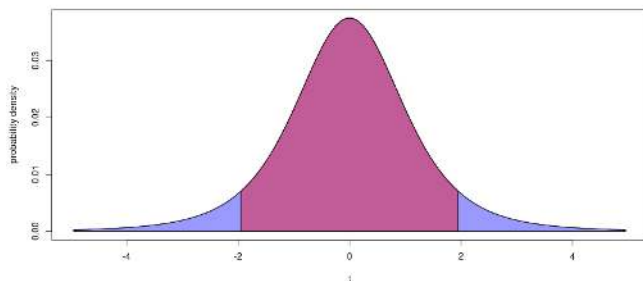
For instance, the probability that the sample mean is within 2 standard errors of the true mean is approximately

$$\int_{-2}^2 \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{(n-2)\pi}\Gamma\left(\frac{n-2}{2}\right)} \left(1 + \frac{x^2}{n-2}\right)^{-\frac{n-1}{2}} dx.$$

2 - 5

For instance, the probability that the sample mean is within 2 standard errors of the true mean is approximately

$$\int_{-2}^2 \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{(n-2)\pi}\Gamma\left(\frac{n-2}{2}\right)} \left(1 + \frac{x^2}{n-2}\right)^{-\frac{n-1}{2}} dx.$$



2 - 5

Intuition

1. Simulate a dataset of 20 random draws from a Normal distribution with mean 0, and do a t test of the hypothesis that $\mu = 0$.

2 / 6

Intuition

1. Simulate a dataset of 20 random draws from a Normal distribution with mean 0, and do a t test of the hypothesis that $\mu = 0$.
2. Do that 1,000 times, and make a histogram of the resulting p -values. What proportion are less than 0.05?

2 / 6

Intuition

1. Simulate a dataset of 20 random draws from a Normal distribution with mean 0, and do a t test of the hypothesis that $\mu = 0$.
2. Do that 1,000 times, and make a histogram of the resulting p -values. What proportion are less than 0.05?
3. Change mean of the simulated values to 1, and do the same.

2 / 6

Confidence intervals

3 . 1

A *95% confidence interval* for an estimate is constructed so that no matter what the true values, 95% of the the confidence intervals you construct will overlap the truth.

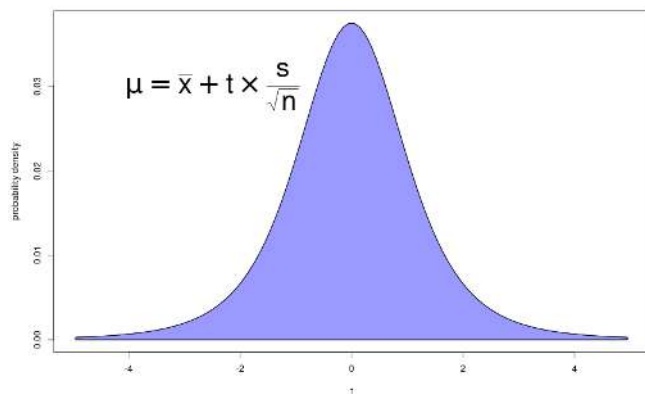
3 . 2

A *95% confidence interval* for an estimate is constructed so that no matter what the true values, 95% of the the confidence intervals you construct will overlap the truth.

In other words, if we collect 100 independent samples from a population with true mean μ , and 95% construct confidence intervals for the mean from each, then about 95 of these should overlap μ .

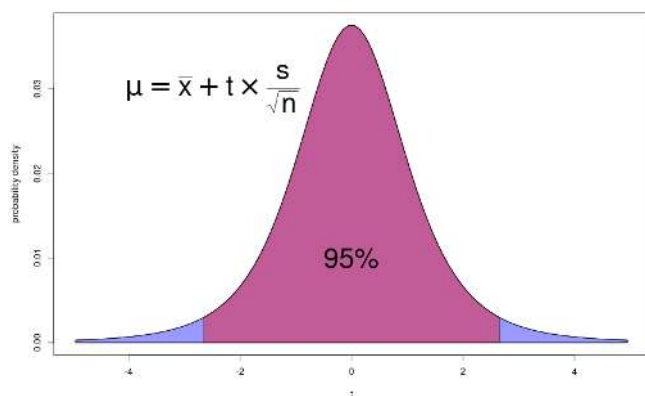
3 . 2

How's that work?



3 . 3

How's that work?



3 . 4

Check this.

if we collect 100 independent samples from a population with true mean μ , and construct 95% confidence intervals from each, then about 95 of these should overlap μ .

Let's take independent samples of size $n = 20$ from a Normal distribution with $\mu = 0$. Example:

```
n <- 20; mu <- 0  
t.test(rnorm(n, mean=mu))$conf.int
```

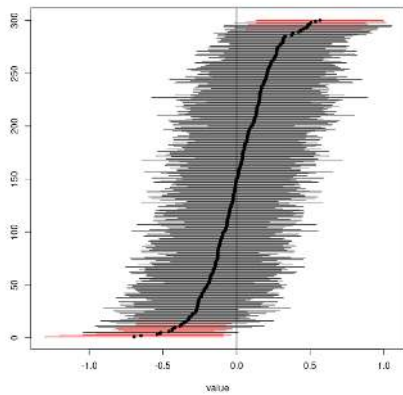
```
## [1] -0.4019083  0.5862080  
## attr(,"conf.level")  
## [1] 0.95
```

3 . 5

```
tci <- replicate(300, t.test(rnorm(n, mean=mu))$conf.int)
mean(tci[,1] > 0 | tci[,2] < 0)
```

```
## [1] 0.05
```

3.6



3.7

What's that 95% mean?

Suppose we survey 100 random UO students and find that 10 had been to a party recently and so get a 95% confidence interval of 4%-16% for the percentage of UO students who have been to a party recently.

3.8

What's that 95% mean?

Suppose we survey 100 random UO students and find that 10 had been to a party recently and so get a 95% confidence interval of 4%-16% for the percentage of UO students who have been to a party recently.

There is a 95% chance that the true proportion of UO students who have been to a party recently is between 4% and 16%.

3 - 8

What's that 95% mean?

Suppose we survey 100 random UO students and find that 10 had been to a party recently and so get a 95% confidence interval of 4%-16% for the percentage of UO students who have been to a party recently.

There is a 95% chance that the true proportion of UO students who have been to a party recently is between 4% and 16%.

Not so good: the true proportion is a *fixed* number, so it doesn't make sense to talk about a *probability* here.

3 - 8

Power analysis

4 - 1

Statistical power is how good our statistics can find things out.

4 / 2

Statistical power is how good our statistics can find things out.

Formally: the probability of identifying a true effect.

4 / 2

Statistical power is how good our statistics can find things out.

Formally: the probability of identifying a true effect.

Example: Suppose two snail species' speeds differ by 3cm/h. What's the chance our experiment will identify the difference?

4 / 2

A prospective study

Suppose that we're going to do a survey of room prices of an AirBnB competitor. How do our power and accuracy depend on sample size? Supposing that prices roughly match AirBnB's: mean $\mu = \$120$ and SD $\sigma = \$98$, estimate:

1. The size of the difference between the mean price of a random sample of size n and the (true) mean price.
2. The probability that a sample of size n rooms has a sample mean within \$10 of the (true) mean price.

4 . 3

Group exercise

Answer those questions *empirically*: by taking random samples from the price column of the airbnb data, make two plots:

1. Expected difference between the mean price of a random sample of n Portland AirBnB rooms and the (true) mean price of *all* rooms, as a function of n .
2. Probability that a sample of size n of Portland AirBnB rooms has a sample mean within \$10 of the (true) mean price of *all* rooms, as a function of n .

4 . 4

In class: part 1

```
true_mean <- mean(airbnb$price, na.rm=TRUE)
# do it once
n <- 20
sample_mean <- mean(sample(airbnb$price, n))
# do it a lot of times
nvals <- 10 * 10:100
nreps <- 100
sample_means <- matrix(NA, nrow=nreps, ncol=length(nvals))
for (n in 1:length(nvals)) {
  n <- nvals[j]
  sample_means[,j] <- replicate(nreps, mean(sample(airbnb$price, n), na.rm=TRUE))
}
```

```
## Error: object 'j' not found
```

```
plot(nvals, colMeans(abs(sample_means - true_mean)),
     main="distance from the true mean, by sample size",
     xlab="sample size",
     ylab="mean absolute error",
     pch=20)
```

```
## Warning in min(x): no non-missing arguments to min; returning Inf
```

```
## Warning in max(x): no non-missing arguments to max; returning -Inf
```

```
## Error in plot.window(...): need finite 'ylim' values
```

4 . 5

In class: part 2

```
plot(nvals, colMeans(abs(sample_means - true_mean) < 10),  
     xlab='sample size',  
     ylab='Prob(error < 10)',  
     main='probability sample mean is within $10 of true mean',  
     pch=20)
```

```
## Warning in min(x): no non-missing arguments to min; returning Inf
```

```
## Warning in max(x): no non-missing arguments to max; returning -Inf
```

```
## Error in plot.window(...): need finite 'ylim' values
```

4 . 6

// reveal.js plugins