Week 8 Live Session

w203 Instructional Team

Feb 27, 2017

T-Distribution Functions in R

Suppose $X_1, X_2, ..., X_n$ is a random sample from a normal distribution with mean μ and unknown standard deviation.

Let $T = (\bar{X} - \mu)/(s/\sqrt{n})$ where $\bar{X}$ is the mean and s is the sample standard deviation (sd) of $X_1,, X_n$.

Then T is distributed as a t-distribution with df= n-1 degrees of freedom.

In R, dt gives the density, pt gives the distribution function, qt gives the quantile function, and rt generates random deviates.

Usage

dt(x, df, ncp, log = FALSE)

pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)

qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)

rt(n, df, ncp)

Arguments

x, q vector of quantiles.

p vector of probabilities.

n number of observations. If length(n) > 1, the length is taken to be the number required.

df degrees of freedom (> 0, maybe non-integer). df = Inf is allowed.

ncp non-centrality parameter delta; currently except for rt(), only for abs(ncp) <= 37.62. If omitted, use the central t distribution.

log, log.p logical; if TRUE, probabilities p are given as log(p).

lower.tail logical; if TRUE (default), probabilities are P(X $\leq$ x), otherwise, P(X > x).

Practice with T-Distributions

What is the 2.5th and 97.5th percentiles of the t distribution with 5 degrees of freedom?

```r
qt(c(.025, .975), df=5)
```

```
## [1] -2.570582  2.570582
```

What is the 2.5th and 97.5th percentiles of the t distribution with 1 to 10 degrees of freedom?

```
 qt(c(.025, .975), df=1:10)
```

```
##  [1] -12.706205   4.302653  -3.182446   2.776445  -2.570582   2.446912
##  [7]  -2.364624   2.306004  -2.262157   2.228139
```

P-Values and Confidence Intervals

The actual voltages of power packs labeled as 12 volts are as follows: 11.77, 11.90, 11.64, 11.84, 12.13, 11.99, and 11.77.

1) Find the mean and the standard deviation.

```
V<-c(11.77, 11.90, 11.64, 11.84, 12.13, 11.99,  11.77)
vbar <- mean(V)
vbar
```

```
## [1] 11.86286
```

```
s <- sd(V)
s
```

```
## [1] 0.1614223
```

2) What is the critical value for a 95% confidence interval for this sample?

```
c<-abs(qt(0.05/2, length(V)-1))
c
```

```
## [1] 2.446912
```

3) Find the 95% confidence interval for the sample.

```
sem <- s/sqrt(length(V))
c(vbar-c*sem,vbar+c*sem)
```

```
## [1] 11.71357 12.01215
```

4) Define a test statistic, i.e. a variable T, as a function of the values from (1) that tests whether the mean is 12.

```
T <- (vbar-12)/(s/sqrt(length(V)))
T
```

```
## [1] -2.247806
```

5) Calculate the p-value using the t statistic. Should you reject the null hypothesis? (Also, could you predict what would happen from your answer to (3)?)

```
  2*pt(-abs(T),df=length(V)-1)
```

```
## [1] 0.06563885
```

Since $p > .05$ we fail to reject the null hypothesis. Note that we could have also gotten the same result as follows:

```
t.test(V, mu = 12)
```

```
##
##  One Sample t-test
##
## data:  V
## t = -2.2478, df = 6, p-value = 0.06564
## alternative hypothesis: true mean is not equal to 12
## 95 percent confidence interval:
##  11.71357 12.01215
## sample estimates:
## mean of x 
##  11.86286
```

6) Suppose you were to use a normal distribution instead of a t-distribution to test your hypothesis. What would your p-value be for the z-test?

```
2*pnorm(-abs(T))
```

```
## [1] 0.02458857
```

This time, note that $p < .05$ so we would have rejected the null. The t-test is more conservative than the z-test. This example highlights the fact that an inscrupulous researcher might be tempted to use the z-test instead of the t-test in order to get a significant result. However, since we do not know the population standard deviation, the z-test is invalid.

Executing T-tests in R

The file athlet2.Rdata contains data on college football games. The data is provided by Wooldridge and was collected by Paul Anderson, an MSU economics major, for a term project. Football records and scores are from 1993 football season.

```
load("athlet2.RData")
```

We are especially interested in the variable, dscore, which represents the score differential, home team score - visiting team score. We would like to test whether a home team really has an advantage over the visiting team.

a. The instructor will assign you to one of two teams. Team 1 will argue that the t-test is appropriate to this scenario. Team 2 will argue that the t-test is invalid. Take a few minutes to examine the data, then formulate your best argument.

We check to see how many observations we have, as well as the shape of the population distribution of our variable.

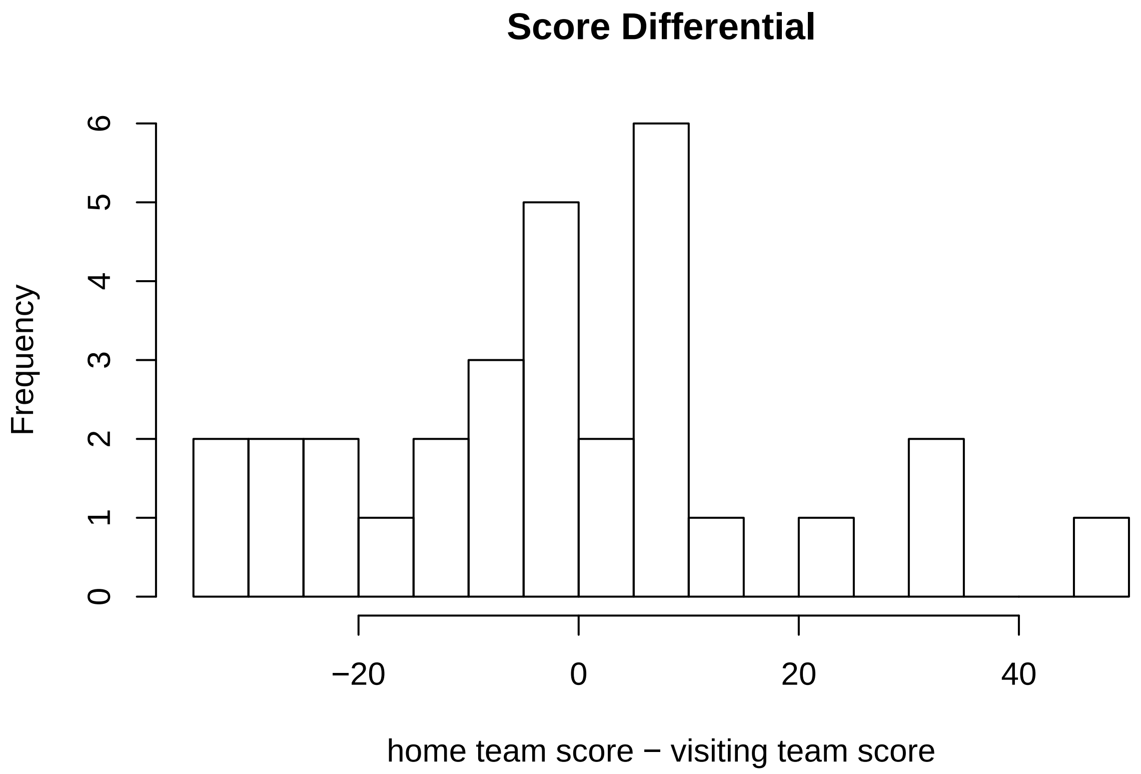
```
length(data$dscore)
```

```
## [1] 30
```

```
summary(data$dscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -33.00  -13.50   -3.00   -1.10    7.75   48.00
```

```r
hist(data$dscore, breaks=20,main="Score Differential", xlab="home team score - visiting team score")
```



Because $n = 30$ exactly, we just barely meet the rule-of-thumb requirement for the CLT. In this scenario, you could try to argue that 30 observations might not really be enough to achieve a normal sampling distribution. However, the histogram of dscore doesn't reveal any substantial skew, which is when we would normally worry about that we need more than 30 observations, so this argument isn't very strong.

For students arguing that the t-test is invalid, a better approach might be to ask whether the data points are truly independent. There are only so many college football teams, and it seems likely that each team appears in multiple rows of data. If a team achieves a high score in one game, it is probably likely to achieve a high score in other games. This suggests that we may be seeing less variation in our data set than actually exists in the population, and the denominator in our t-statistic will be too small, exaggerating our results.

b. Should you perform a one-tailed test or a two-tailed test? What is the strongest argument for your answer?

Even though you are specifically interested in a home team *advantage*, the better answer is to choose the two-tailed test. There are several reasons for this:

- Your reader may not share your theory that home teams have the advantage and may even believe the opposite.
- The one-tailed test makes it easier to reject the null, but there is no way to prove to your reader that you didn't start with a two-tailed test and then switch to the one-tail once you saw the direction of the effect. That would be cheating and your type-1 error rate would be greater than .05. As a result, readers may be suspicious of whether you really committed to the type of test ahead of time.
- Imagine that you select the one-tailed test, but then the data surprises you and you find that the visiting team has a strong advantage. Suppose the visiting team actually has a 10 point advantage over

the home team. By the rules of hypothesis testing, you can't reject the null, no matter how big the visiting team advantage is. But would you really pack up your study at this point and say you were unable to find evidence against the null? Will your reader believe that you'd really do this, and that you wouldn't cheat by switching to the two-tailed test?

c. Execute the t-test and interpret every component of the output.

```r
t.test(data$dscore, mu=0)
```

```
##
##  One Sample t-test
##
## data:  data$dscore
## t = -0.30781, df = 29, p-value = 0.7604
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -8.408919  6.208919
## sample estimates:
## mean of x
##      -1.1
```

Make sure sure understand every line of the output.

d. Based on your output, suggest a different hypothesis that would have led to a different test result. Try executing the test to confirm that you are correct.

The easiest thing to do is to look at the 95% confidence interval. Since 0 is inside the interval, we already know that the null cannot be rejected. If we had chosen a null hypothesis outside of this interval, say $mu = 7$, it would have been rejected.

Assumptions Behind the T-test

For each of the following scenarios, give your best argument for why a t-test may be *invalid*.

a. You have a sample of 50 CEO salaries, and you want to know whether the mean salary is greater than $1 million.

b. You have data on 1000 students that are sampled from 10 randomly chosen public universities, and you want to know whether public university students sleep less than 7 hours a night.

c. A nonprofit organization measures the percentage of students that pass an 8th grade reading test in 40 neighboring California counties. You are interested in whether the percentage of students that pass in California is over 80%

Your own T-Test Function

1. Using your understanding of the procedure for a one-sample t-test, write your own function to execute the test. You may use the following function header.

```r
my.t.test = function(values, alpha = 0.05, mu = 0)
{
  # Your code here
}
```

2. Autogenerate a sequence of 20 values and use it as the input of your function. Did you reject or accept the null hypothesis?
3. How can you test that the function we created works properly?