DATA SCIENCE DISCOVERY

#37: t Tests

December 2, 2019

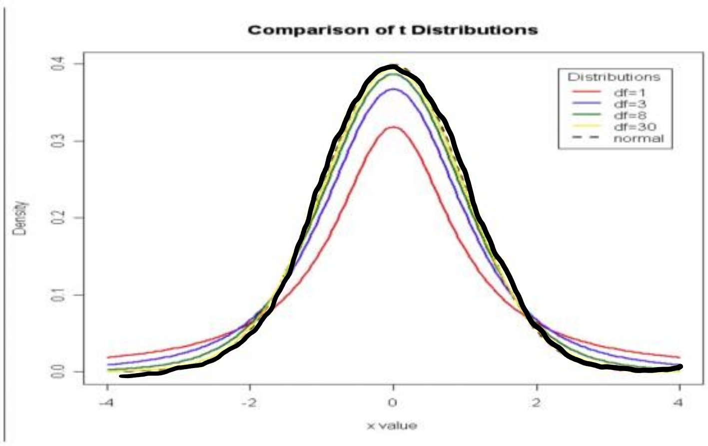
Karle Flanagan and Wade Fagen-Ulmschneider

1 Sample t Test

T-tests are very similar to z-tests. They test if a difference we observe is due to chance. We use t-tests and the Student's Curve (*t*-Distribution) when *ALL 3 of these conditions are met:*

1. The sample is **SMALL**, for our purposes less than or equal to 25. (If the sample size is greater than 25, we'll use *z*.)
2. The histogram for the population is close to the **NORMAL** curve.
3. The SD of the population is **UNKNOWN**. (If the SD is known, you can use *z*.)

*****When the sample size is small, using the sample SD to estimate the SD of the population is not very accurate. It is likely to be too low, so we use SD^+ instead. And instead of using the normal curve we use the Student's curves and the *t*-table.**



So, if you have a Small sample drawn from a Normal population an Unknown SD use t.

Here's how the t-curves compare to the Normal Curve (k represents the degrees of freedom).

Note how the t curves get closer and closer to the Normal Curve as the degrees of freedom increase.

How to compute the SD^+:

from sample

$$SD^+ = SD * \sqrt{\frac{n}{n-1}}$$

Note: SD+ is ALWAYS greater than SD, but the difference becomes negligible as n gets large.

if n > 26 ⇒ z

What's the difference between SD and SD+ when n=26?

SD = 1

$SD^+ = 1 \times \sqrt{\frac{26}{25}} = 1.019$

There's a t-table just like we have the normal table. There is a different curve for each number of *degrees of freedom* where the $\boxed{\text{degrees of freedom} = n - 1.}$ The curves are fatter in the tails than the normal curve. This means you need stronger evidence to reject the null.

T Table: http://courses.las.illinois.edu/stat/stat100/t_table.pdf

Example 1: Suppose the Keurig coffee maker claims to brew an 8 oz. cup of coffee in 60 seconds, but I think it actually takes more time than that. To test the coffee maker's claim, I randomly sampled 16 new coffee makers and find the average brewing time to be 64 seconds with an SD of 2 seconds. *Assume the brewing time is approximately normally distributed.*

n

sample statistics

We want to test whether this difference is real or whether it is simply due to chance. First of all, what test should we use? Perform the appropriate test.

Small Sample ✓
Normal data ✓
Unknown pop. SD ✓

$\Rightarrow$ t-test (SD^+)

① H_0 (Null): $\mu = 60$, Keurig is doing as claimed.

H_a (Alt): $\mu > 60$, True brewing time is more than 60 sec.

②

Small sample of 16

obs value

sample mean $= 64$

exp. value $= 60$

$SD^+ = \sqrt{\frac{n}{n-1}} \times$ SD of sample

$= \sqrt{\frac{16}{15}} \times 2 = 2.275$

df $= n - 1 = 16 - 1 = 15$

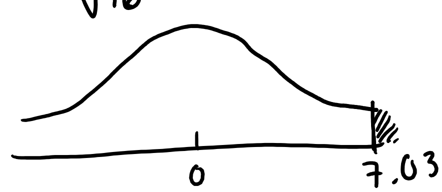
③ test statistic

$t = \frac{\text{val} - EV}{SE} = \frac{64 - 60}{0.569} = 7.03$

$SE = \frac{SD^+}{\sqrt{n}}$

$= \frac{2.275}{\sqrt{16}} = 0.569$

④ p-value $= 1 -$ t.cdf$(7.03, 15) = 0.000002$



⑤ $p < 5\% \Rightarrow$ Reject H_0

There is evidence of the alternative.

ttest_1samp(df, 31)

Example 2: A car company claims that their *Super Spiffy Sedan* averages 31 mpg. You randomly select 8 Super Spiffies from local car dealerships and test their gas mileage under similar conditions. Assume the gas mileage follows the normal curve. You get the MPG scores and store them in a csv file called "t_test.csv." Do a t-test in Python to figure out if the mileage is statistically different from 31.

DataSet: https://docs.google.com/spreadsheets/d/1TAKz3WM3WYmPUP5FrLu4jYniMQhpWH1UnrJCoGJjlak/edit?usp=sharing

Python 1-Sample T-Test:
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ttest_1samp.html

scipy.stats.ttest_1samp(*a*, *popmean*, *axis=0*) * from scipy.stats import ttest_1samp

Calculates the T-test for the mean of ONE group of scores.

This is a two-sided test for the null hypothesis that the expected value (mean) of a sample of independent observations *a* is equal to the given population mean, *popmean*.

Parameters:	**a** : *array_like* (data frame) sample observation **popmean** : *float or array_like* expected value in null hypothesis, if array_like than it must have the same shape as *a* excluding the axis dimension **axis** : *int, optional, (default axis=0)* Axis can equal None (ravel array first), or an integer (the axis over which to operate on a).
Returns:	**t** : *float or array* t-statistic **prob** : *float or array* two-tailed p-value

H_0: MPG = 31

H_A: MPG != 31

TWO SAMPLE t-TESTs

Previously, we used the 2 sample Z-test to compare the means of *two* populations using;

where $SE_{difference}$ =

But if our sample sizes are small (n<25) drawn from roughly normal populations with unknown SD's then we should use the 2 sample t-test instead:

where $SE^{+}_{difference}$ =

The 2-sample t-stat doesn't exactly follow any t-curve so deciding how many degrees of freedom to use is problematic. Since we're estimating the SD's of 2 samples, the simplest and most conservative approach is to use the df of the smaller sample.

df= n_1 -1 where n_1 is the smaller sample size.

Example 1: A randomized double-blind test was done to test the effectiveness of a drug designed to improve memory in Alzheimer's patients. 20 patients took the drug and 16 patients took the placebo. Patients were given a pre and post-memory test (a list of 15 words to recall) and an improvement score (post-test score – pre-test score) was recorded for each patient.

	n	Mean	SD
Drug	20	3	4
Placebo	16	1	3

For a, b, c fill in each blank with either =, >, or < .

a) $H_0 : \mu_{drug} - \mu_{placebo}$ ___ 0

b) Suppose you have prior evidence that the drug works then $H_a : \mu_{drug} - \mu_{placebo}$ ___ 0

c) Suppose you have no idea if the drug works at the start of the study but after looking at the results you believe it does, then $H_a : \mu_{drug} - \mu_{placebo}$ ___ 0

d) Compute the t-statistic. How many df?

e) Find p-value for both the 1-sided H_a and 2-sided H_a

f) If you used the Z test would your p-values be bigger or smaller?