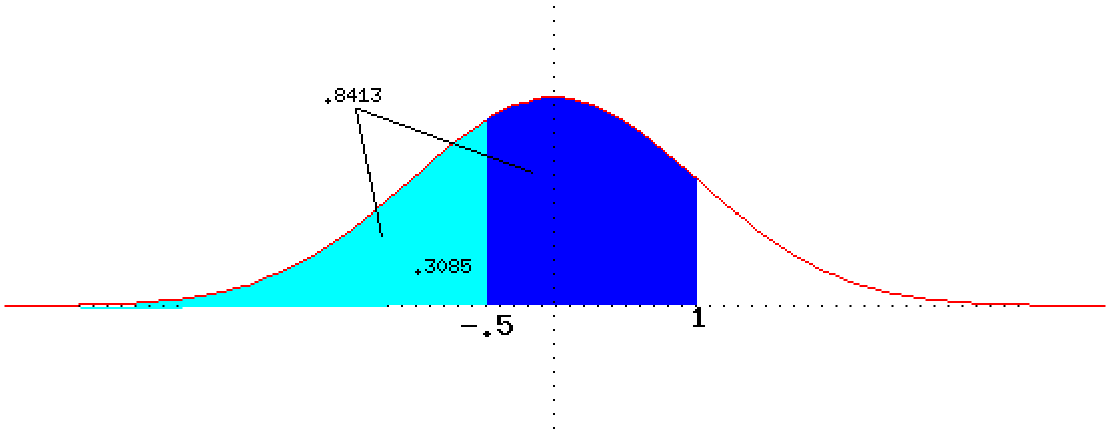
Confidence interval and the Student's t-test

Are you a blogger? Interested in participating in a paid blogging study?

By Joy Ying Zhang, joy+@cs.cmu.edu

Main Ideas

The story starts here: Let's take a look at the *Standard Normal Distribution*.
The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.



z	-3.0	-2.0	-1.0	0	1	2	3
proportion (-∞,z)	0.0013	0.023	0.159	0.5	0.841	0.977	0.9987

Or we can view this in another way:

range	proportion
(-1,+1)	0.6826
(-2,+2)	0.9544
(-3,+3)	0.9974

We can interpret the table above as: for 68.26% of times, z will fall into range (- 1,+ 1), for 95.44% z will be in the range (-2,+2) and for 99.74% z will have value in (-3,+3). Remember this, and we will be back.

==============================

Let's switch the topic to *confidence interval* for a moment:

If we want to find out the mean of a population, but

- the population is very large

- the population is not accessible

We can only take samples from the population. Assume this population has a true mean μ and true standard deviation σ, but of course we don't know their values. Suppose each sample is of size *n*. For each sample, we can calculate the sample mean $\overline{X}$. The mean of all the sample means: $\overline{X}_1$, $\overline{X}_2$,$\overline{X}_3$... is $\mu_{\overline{X}}$, and the standard deviation of the sampling distribution of the sample mean is $\sigma_{\overline{X}}$, also called the ***standard error*** of $\overline{X}$

BTW, There appear to be two different definitions of the ***standard error***. 1) The standard error of a sample of sample size *n* is the sample's standard deviation divided by $\sqrt{n}$. It therefore estimates the standard deviation of the sample mean based on the population mean (Press *et al.* 1992, p. 465). Note that while this definition makes no reference to a normal distribution, many uses of this quantity implicitly assume such a distribution. 2) The standard error of an estimate may also be defined as the square root of the *estimated* error variance $\hat{\sigma}^2$ of the quantity, $s_e \equiv \sqrt{\hat{\sigma}^2}$.

We have:

- $\mu_{\overline{X}} = \mu$, meaning that the sampling distribution of $\overline{X}$ is centered around the true population mean. And because $\mu_{\overline{X}} = \mu$, the sample mean is called an unbiased estimator of μ
- $\sigma_{\overline{X}}$ is the typical amount of error that will be incurred by estimating the population mean with the sample mean. Note that $\sigma_{\overline{X}}=\sigma/\sqrt{n}$ gets smaller as *n* gets larger. This means that $\overline{X}$ more accurately estimates the population mean for larger samples than smaller ones.

In conclusion, suppose a sample of size n is taken from a normal population with mean μ and standard deviation σ, the sampling distribution of $\overline{X}$ is also a normal distribution with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}}=\sigma/\sqrt{n}$. **The sampling distribution is normal if the original population is normal.**

Now, the standardized version of $\overline{X}$ is:

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$ ~ has a standard normal distribution

This means, whatever μ is, we have:

$$P(-2 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 2) = 0.9544$$

(see table)

Or, in other words,

$$P(\overline{X} - 2\sigma/\sqrt{n} \le \mu \le \overline{X} + 2\sigma/\sqrt{n}) = 0.9544$$

Or, that the **interval** $(\overline{X} - 2\sigma/\sqrt{n}, \overline{X} + 2\sigma/\sqrt{n})$ contains the population mean (μ) with 99.54% confidence. This is a 99.54% confidence interval for μ.

In conclusion:

We can measure the confidence intervals for the "real" mean μ if:

- Population is normal, or if the sample size is large
- σ is known.
- 100*(1-α)% confidence interval for the population mean is:

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

Here are some critical Z values. Z-values can be calculated and demonstrated here

α	**Confidence**	$Z_{\alpha/2}$
0.1	90%	1.64
0.05	95%	1.96
0.01	99%	2.58

0.001	99.9%	3.29

==========================

In the above section, **we assume that we know the standard deviation (σ) of the population** and transferred the $\overline{X}$ into ***Z*** which is standard normal distribution and use the z-value to estimate the confidence intervals for the population mean μ

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$ ~ standard normal distribution

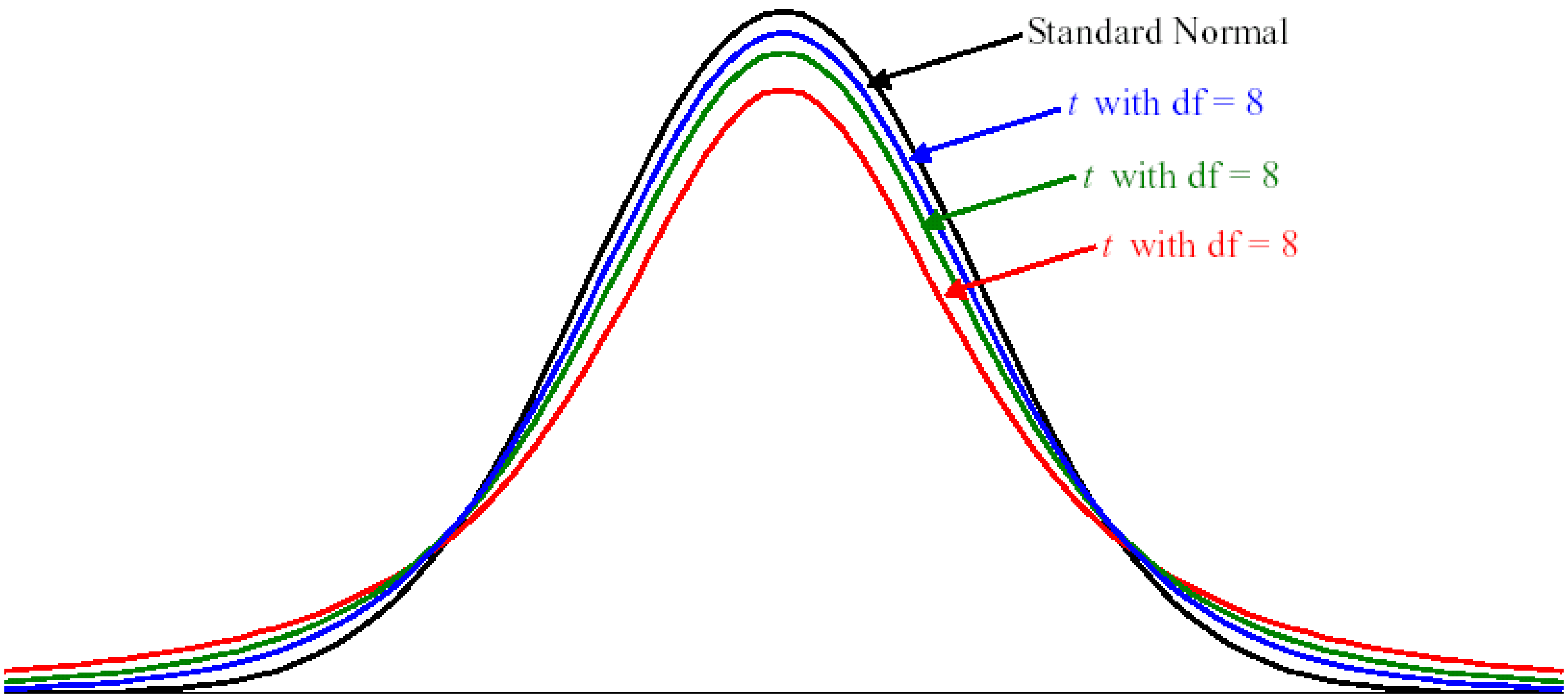
Yet, in the cases when **σ is unknown**, we can only estimate it with the sample standard deviation ***S*** and transfer the $\overline{X}$ into ***T*** which does not have a standard normal distribution. ***T*** follows what is called Student's t-distribution.

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$ ~ t-distribution

==========================

"Student" (real name: W. S. Gossett [1876-1937]) developed statistical methods to solve problems stemming from his employment in a brewery.

Student's *t*-distribution



The t-distribution has one parameter called the *degree of freedom (df)*, DF=n-1

The t-distribution is similar to the normal distribution:

- Symmetric about 0
- Bell shaped

The main differences between the t-distribution and the normal distribution is in tails (Play around with DF and see the difference of the tails):

- T-distribution has larger tails than the normal
- Larger DF means smaller tails, the larger the DF, the closer to the normal distribution
- Small DF means larger tails

T-test for one variable: calculating confidence interval for mean μ, σ unknown

- Suppose a sample of size n is taken from a population with mean μ and standard deviation σ
  Assumptions:
  - Population is normal, or the sample is large
  - σ is unknown
- A 100*(1-α)% confidence interval for μ is:

$$(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}})$$

α	0.1	0.05	0.02	0.01	0.002	0.001
Confidence	90%	95%	98%	99%	99.8%	99.9%
DF=1	6.314	12.71	31.82	63.66	318.3	636.6
DF=10	1.812	2.228	2.764	3.169	4.144	4.587
DF=20	1.725	2.086	2.528	2.845	3.552	3.850
DF=∞ **(same as Z distribution)**	1.645	1.960	2.326	2.576	3.091	3.291

Which T-test to use

(for more information of how to choose a statistical test)

Goal and Data	Type of T-test	Assumption	Comments	T	DF
Compare one group to a hypothetical value	one-sample t test	Subjects are randomly drawn from a population and the distribution of the **mean being tested is normal**	Usually used to compare the mean of a sample to a know number (often 0)	$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$	**n-1**
Compare two unpaired groups	unpaired t test	Two-sample assuming **equal variance (homoscedastic t-test)**	Two samples are referred to as **independent** if the observations in one sample are not in any way related to the observations in the other. This is also used in cases where one randomly assign subjects to two groups, give first group treatment A and the second group treatment B	$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	**n_1+n_2-2**

			and compare the two groups		
		Two-sample assuming **unequal variance (heteroscedastic t-test)**	The **variance** in the two groups are **extremely different**. e.g. the two samples are of very different sizes	$T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}$	$df' = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
Compare two paired groups	paired t test	The observed data are from the same subject or from a matched subject and are drawn from a **population with a normal distribution** **does not assume** that the variance of both populations are equal	used to compare means on the same or related subject over time or in differing circumstances; subjects are often tested in a before-after situation	$T = \dfrac{\overline{X} - \overline{Y}}{S_d / \sqrt{n}}$	**n-1**

Data set

Subject ID	Data

Males	Females	Males	Females
70	87	165.9	212.1
71	89	210.3	203.5
72	90	166.8	210.3
76	94	182.3	228.4
77	97	182.1	206.2
78	99	218	203.2
80	101	170.1	224.9
	102		202.6

t-Test: Two-Sample Assuming Equal Variances

To compute the two-sample t-test two major computations are needed before computing the t-test. First, you need to estimate the pooled standard deviation of the two samples. The pooled standard deviation gives an weighted average of the standard deviations of the two samples. The ***pooled standard deviation*** is going to be between the two standard deviations, with greater weight given to the standard deviation from a larger sample. The equation for the pooled standard deviation is:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

In all work with two-sample t-test the degrees of freedom or df is:

$$df = n_1 + n_2 - 2$$

The formula for the two sample t-test is:

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

For example, for this data set

t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	Variable 2
Mean	185.0714286	211.4
Variance	443.802381	101.0114286
Observations	7	8
Pooled Variance	259.2226374	
Hypothesized Mean Difference	0	
df	13	
t Stat	-3.159651739	
P(T<=t) one-tail	0.0037652	
t Critical one-tail	1.770931704	
P(T<=t) two-tail	0.0075304	
t Critical two-tail	2.16036824	

t-Test: Two-Sample Assuming Unequal Variances

Assumption:

1. The samples (n_1 and n_2) from two normal populations are independent
2. One or both sample sizes are less than 30

3. The appropriate sampling distribution of the test statistic is the t distribution
4. The unknown variances of the two populations are not equal

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}$$

Note in this case the Degree of Freedom is measured by

$$df' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

and round up to integer.

For example, for this data set

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	Variable 2
Mean	185.0714286	211.4
Variance	443.802381	101.0114286
Observations	7	8
Hypothesized Mean Difference	0	

df	8	
t Stat	-3.01956254	
P(T<=t) one-tail	0.008285256	
t Critical one-tail	1.85954832	
P(T<=t) two-tail	0.016570512	
t Critical two-tail	2.306005626	

Paired Student's t-test

For each pair of data, think of creating a new sequence of data: differences.

Data ID	Value X	Value Y (after treatment)	Difference
1	X_1	Y_1	X_1-Y_1
2	X_2	Y_2	X_2-Y_2
i	X_i	Y_i	X_i-Y_i
...	...	...	...
n	X_n	Y_n	X_n-Y_n

Hypothesis: $\overline{\text{Difference}}$ = μ, usually, if we just want to test if two systems are different μ=0

Apply the one-sample t-test on the *difference* sequence

$$T = \frac{\overline{X} - \overline{Y}}{S_d / \sqrt{n}}$$

Or,

Given two paired sets X_i and Y_i of *n* measured values, the paired *t*-test determines if they differ from each other in a significant way. Let

$$\hat{X}_i = (X_i - \bar{X})$$

$$\hat{Y}_i = (Y_i - \bar{Y}),$$

$$t = (\bar{X} - \bar{Y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{X}_i - \hat{Y}_i)^2}}.$$ with degree of freedom = n-1

For example, for the following data set:

ID	**X**	**Y**	***X-Y***
1	154.3	230.4	*76.1*
2	191	202.8	*11.8*
3	163.4	202.8	*39.4*
4	168.6	216.8	*48.2*
5	187	192.9	*5.9*
6	200.4	194.4	*-6*
7	162.5	211.7	*49.2*

t-Test: Paired Two Sample for Means			
	Variable 1	Variable 2	Variable1-Variable2
Mean	175.314	207.400	32.086

Variance	300.788	176.237	848.508
Observations	7.000	7.000	
Hypothesized Mean Difference	0.000		
df	6		
t Stat	-2.914		
P(T<=t) one-tail	0.013		
t Critical one-tail	1.943		
P(T<=t) two-tail	0.027		
t Critical two-tail	2.447		

$$t = \frac{175.314 - 207.4}{\sqrt{848.508}/\sqrt{7}} = \frac{-32.086}{29.13/2.646} = -2.914$$

Critical Values

You can find the t test critical values online

Or, you can use the perl library to calculate the values:

Statistics::Distributions - Perl module for calculating critical values and upper probabilities of common statistical distributions (download the package)

e.g.

$tprob=Statistics::Distributions::tprob (3,6.251); print "upper probability of the t distribution (3 degrees of " ."freedom, t = 6.251): Q = 1-G = $tprob\n";

Be Careful Here:

The returned p value stands for the proportion of the area under the curve between t and ∞, if one wants to measure the confidence C,

$C=1-2p$

Confidence and Precision

The *confidence level* of a confidence interval is an assessment of how *confident* we are that the true population mean is within the interval.
The *precision* of the interval is given by its width (the difference between the upper and lower endpoint). Wide intervals do not provide us with very precise information about the location of the true population mean. Short intervals provide us with very precise information about the location of the population mean.

If the sample size n remains the same:

- Increasing the confidence level of an interval decreases precision
- Decreasing the confidence level of an interval increases its precision

Generally confidence levels are chosen to be between about 90% and 99%. These confidence levels usually provide reasonable precision and confidence.

References:

- Student's *t*-Tests
- Student's t-test on wikipedia
- The Insignificance of Statistical Significance Testing: What is Statistical Hypothesis Testing?
- Resampling techniques for statistical modeling
- Resampling and the Bootstrap
- Confidence Intervals for One Population Mean, by Dr. Alan M. Polansky
- Z-value
- Standard deviation and variance
- Difference of Two-Means Test

- Statistical inference

00012457 since Dec. 1, 2006
View My Stats