EFAB StatChat
Issue No. 1 Februarv 2010

This Edition's Inquiry

Differences in Central Tendency Between Two Groups

Q: What measure of central tendency should I use to describe my data? How do I test for differences in this measure between two groups?

A: We typically see two measures of central tendency reported in the literature: mean and median. The mean is simply the arithmetic average of the data points (add the data values and divide by the number of observations).

The median is the center value when observations are listed in numerical order. The median is a robust measure of center, meaning that it does not change much in the presence of outlying observations or highly skewed data. This is because the order of the values is all that matters in finding the median; actual values are not used. In contrast, the mean is a nonrobust measure of center; outliers and skewed data can influence the value of the mean. For example, one substantially large value can shift the mean to be larger than most values in the data. When data have no outliers and are roughly symmetric, the mean and median will be approximately equal. Visual inspection of data using a histogram will usually aid in making the decision of which measure to report. The appropriate measure of spread when using the mean is the standard deviation. When reporting the median, interquartile range (the distance between then 75^{th} and 25^{th} percentiles) is the appropriate measure of spread. However, range is also frequently reported as the measure of spread when using the median.

Student's *t*-test is a two-sample location test of the null hypothesis that the means of two normally distributed samples are equal. This test assumes random sampling, statistical independence, and a normal population distribution. Student's *t*-test is a parametric test because of the normality assumption. However, because of the central limit theorem, we need not be too concerned with the normality assumption as long as n is large enough (typically when the total number of observations across groups is $\geq$30).

For the original Student's *t*-test, the variance of the two groups must be equal. If the variances are not equal in the two groups, then a fix up (Satterthwaite method) is applied to Student's *t*-test.

Nonparametric alternatives are most often used when sample sizes are limited, and thus, we cannot rely on central limit theorem to meet the normality assumptions of the parametric test or when the median is reported.

The Mann-Whitney U, or Wilcoxon rank sum, test evaluates the hypothesis that the probability distribution of the two groups is the same. This test assumes statistical independence, but does not require any assumption about the shape of the distribution of population values. If the data are normally distributed, the Mann-Whitney U test is less powerful than Student's *t*-test.

Extensions of Student's *t*-test and Mann-Whitney U test exist to handle situations in which two responses are measured on the same person (paired *t*-test, Wilcoxon signed rank test) and for comparing means from more than two groups (ANOVA, Kruskal-Wallis, analysis of ranks).

For more details:

http://www.indiana.edu/~statmath/stat/all/ttest/

http://www.statsoft.com/TEXTBOOK/stnonpar.html

THIS ISSUE

Programming Tips

The following illustrates the use of SAS® (SAS Institute, Inc, Cary, North Carolina) to carry out Student's *t*-test and Mann Whitney U test.

The data contain 14 randomly generated observations. The first seven observations (var2 = 0) come from a Poisson distribution with a mean of 5. The last seven (var2 = 1) come from a Poisson distribution with a mean of 2.

Student's *t*-test

Table 1 provides summary statistics of var1 by level of var2. Table 2 provides the test statistics and *P* values for the null hypothesis that the means are equal in the two groups. If the variances are equal, we use the pooled method; if unequal, we use the Satterthwaite method. Table 3 provides the test statistic and *P* value for the null hypothesis that the variances are equal in the two groups. This is used to inform us which test in Table 2 we should use.
Summary: We reject the null of equal variances (*P* = .0472) and use the Satterthwaite method.

On the basis of this method, we conclude not enough evidence exists to say that the means of the two groups are different (*P* =. 0642). The sample size is limited (n = 14), and we know that the population distribution is not normal (data were generated from Poisson distributions). Therefore, this is not the best test for our data.

Mann Whitney, or Wilcoxon rank sum, test

Table 4 provides the *P* value for testing if the two distributions are the same.
Summary: On the basis of this method, we would reject the null (*P* = .0449) and conclude the distributions are different. We use the *t* approximation because our sample size is limited.

More programming details and examples:
http://www.ats.ucla.edu/stat/sas/output/ttest.htm
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/ttest_toc.htm
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/npar1way_toc.htm

SAS Program

```
Data example;
input var1 var2;
cards;
4       0
2       0
3       0
10      0
3       0
2       0
4       0
1       1
3       1
1       1
3       1
2       1
1       1
0       1
run;
/*Student's t-test*/
proc ttest data = example;
class var2; *group variable;
var var1; * analysis variable;
run;
/*Mann-Whitney U/ Wilcoxon rank sum test*/
proc npar1way data = example Wilcoxon;
class var2; *group variable;
var var1; * analysis variable;
run;
```

Abbreviated SAS Output

Table 1

Variable	var2	N	Lower Mean	Mean	Upper Mean	Std
var1	0	7	1.4392	4	6.5608	2.7689
var1	1	7	0.5228	1.5714	2.6201	1.1339
var1	Diff (1–2)		-0.035	2.4286	4.8926	2.1157

Table 2

T-Tests

Variable	Method	Variances	DF	*t*	Pr>|t|
var1	Pooled	Equal	12	2.15	0.0529
var1	Satterthwaite	Unequal	7.96	2.15	0.0642

Table 3

Equality of Variances

Variable	Method	Num	Den	F Value	Pr>
var1	Folded F	6	6	5.96	0.0472

Table 4

Wilcoxon Two-Sample Test

t Approximation	
One-Sided Pr > Z	0.0225
Two-Sided Pr > |Z|	0.0449

Just for Fun

Statistics Word Search

F	V	C	N	N	A	G	B	E	E	F	N	D	W	D
G	U	I	N	N	E	S	S	C	T	S	V	H	E	M
G	E	X	O	L	Y	T	L	N	O	T	I	W	N	E
P	Q	Z	G	O	L	F	M	E	N	T	E	P	B	A
S	Y	F	M	U	S	N	T	R	N	K	K	S	S	N
U	T	A	R	T	H	H	N	E	S	I	A	H	T	C
A	N	U	S	L	A	Z	Y	F	S	M	D	M	O	C
N	A	C	D	I	E	E	N	F	W	S	A	M	L	I
N	A	I	D	E	M	A	M	I	L	R	O	V	J	E
G	U	I	R	R	N	B	X	D	G	J	G	G	I	W
C	L	W	O	C	E	T	A	O	C	E	N	T	E	R
X	D	E	F	E	J	J	T	T	S	I	Y	T	V	T
J	O	F	X	I	W	S	T	C	T	P	S	Y	E	L
X	F	Y	O	W	I	E	H	J	X	G	U	F	U	O
A	X	L	M	H	Y	L	J	M	X	X	X	R	E	W

Center
Difference
Gosset
Guinness
Histogram
Mann
Whitney
Mean
Median
Outlier
Oxford
Skewed
Student
Ttest

Stat Humor

Enjoy rap music? Check out Betsy's favorite rappers, StatzRappers.

http://video.google.com/videoplay?docid=5243677894327730537&ei=Nu7s7VWL57cqAKY07gE&q=statzrapper&safe=active#

Featured Statistician

Although he was never formerly employed as a statistician, William Sealey Gosset (1876–1937) is best known for his derivation of Student's *t*-distribution. After finishing a degree in chemistry from New College, Oxford, Gosset took a position with Guinness brewery. There, he applied statistics to help brew high-quality beer. Never heard of Gosset? After a previous employee had published papers containing trade secrets of the company, Guinness forbid its employees to publish papers. Gosset therefore published under the pen name of Student.

Find out more about Gosset:

http://www.shsu.edu/~icc_cmf/bio/gosset.html

Resources

The University of California at Los Angeles' statistical computing web-site (http://www.ats.ucla.edu/stat/default.htm) is a useful place to find help with SAS, SPSS® (SPSS, Inc., Chicago, Illinois), or STATA®(STATA Corporation, College Station, Texas). This site has >4,000 pages relating to statistical methods and techniques and receives approximately 800,000 visitors each month. It also has >60 online classes and seminars that are available 24/7 and cover introductory to cutting-edge topics in statistical computing.

Announcements

May 2010: Epidemiological Preparation for Statistical Consultation, Betsy Gunnels by MS Live Meeting. Details coming soon by e-mail.

Your Input Is Welcome!

If you have comments, suggestions, or contributions to the *EFAB StatChat*, send an e-mail to Betsy Gunnels at bic6@cdc.gov.

Solution to word-search puzzle

X X X X X X X X **E** X X X X **W** **D**
G **U** **I** **N** **N** **E** **S** **S** **C** **T** X X **H** **E** **M**
X X X X X X X X **N** X **T** **I** **W** X **E**
X X X X **O** X X X **E** X **T** **E** X X **A**
S X X **M** **U** X X **T** **R** **N** **K** X **S** X **N**
X **T** **A** X **T** X X X **E** **S** X X X **T** X
X **N** **U** X **L** X X **Y** **F** **S** X X **M** X X
N X X **D** **I** X X X **F** X **S** **A** X X X
N **A** **I** **D** **E** **M** X X **I** X **R** **O** X X X
X X X **R** **R** **N** X X **D** G X X **G** X X
X X X **O** X X **T** X **O** **C** **E** **N** **T** **E** **R**
X X X **F** X X X **T** X X X X X X X
X X X **X** X X **S** X X X X X X X X
X X X **O** X **I** X X X X X X X X X

EFAB StatChat is published by the Scientific Education and Professional Development Program Office (SEPDPO) (proposed), Centers for Disease Control and Prevention (CDC)
U.S. Department of Health and Human Services, Atlanta, GA 30333.

Suggested citation: Centers for Disease Control and Prevention. [Article title]. EFAB StatChat 2010;1:[inclusive page numbers].