

STAT101: A Review of the Basics

by Judi Reine
Central Missouri State University

Testing for Normal Distribution:

The normal distribution is a symmetrical, bell-shaped distribution of values. Many statistical procedures are based on the assumption that the sample data was drawn from a normally distributed population. Violating this assumption will cause the results to give misleading conclusions. The UNIVARIATE procedure is used to test whether or not a distribution is normal with the Shapiro-Wilk statistic. The Shapiro-Wilk statistic tests the null hypothesis that the sample was drawn from a normally distributed population. The p-value given with the Shapiro-Wilk statistic ($P < W$) tells the probability of obtaining the given results if the data were drawn from a normal population. If the p-value is low, say less than .05, you should reject the null hypothesis of normality. Along with telling you if the data is normally distributed, the UNIVARIATE procedure will tell you something about the shape of your data with the Kurtosis and Skewness statistics. The Kurtosis statistic tests the peakedness (flatness) of the distribution. A positive kurtosis shows the distribution is relatively peaked--tall and skinny. A negative kurtosis shows a flat distribution. The skewness is used to describe the tails of the distribution. A positive skew indicates the longer tail is at the portion of the distribution with the higher values. A negative skew has the longer tail with the lower values. A normal distribution will have both a kurtosis and skewness of zero.

Comparing Two Sample Means:

There are three ways to compare the means of two groups -- the independent t-test, paired samples t-test (dependent t-test), and the Wilcoxon test. The independent t-test is used for unrelated groups. For example, to compare treatment and control groups or males and females. The dependent t-test compares related samples such as pre and post tests for the same person or before and after treatment results. The Wilcoxon test is used when the assumptions of an independent t-test are not met. Data that are not normally distributed is a common problem.

Examples: (Modified from problems in Cody and Smith)

Independent t-Test: 16 subjects with headaches were divided into two groups. One group was given aspirin and the other Tylenol. The length of time (measured in minutes) needed to feel relief from the headache was recorded as such (this is fictitious data):

Aspirin:	40	42	48	35	62	35	45	38
Tylenol:	35	37	42	22	38	29	39	32

The program and results of the t-test follow:

```
PROC TTEST;  
  CLASS GROUP;  
  VAR TIME;  
RUN;
```

t-TEST PROCEDURE

Variable: TIME

GROUP	N	Mean	Std Dev	Std Error	Minimum	Maximum
Aspirin	8	43.125	8.887	3.142	35.000	62.000
Tylenol	8	34.250	6.409	2.266	22.000	42.000

Variances	T	DF	Prob T
Unequal	2.291	12.7	0.0397
Equal	2.291	14.0	0.0380

For H_0 : Variances are equal, $F' = 1.92$ $DF = (7,7)$ $Prob F' = 0.4078$

First we test the null hypothesis that the variances of the two groups are equal. This is done with the F' statistics given at the bottom of the output which shows the probability that the variances are unequal due to chance alone. If the probability ($Prob F'$) is small, usually less than .05, then reject the hypothesis that the variances are equal and use the statistics labeled UNEQUAL. Otherwise, if the $Prob F'$ value is greater than .05, use the statistics labeled EQUAL. Since our value is greater than .05, we will use the equal statistics. If the $Prob |T|$ value for the correct column is less than .05 then reject the null hypothesis that the means are equal for the two groups and conclude that there is probably a difference in the means. For our example, we can conclude that there is a difference in response time between aspirin and Tylenol and that Tylenol has a quicker response time.

Dependent t-Test: In another study, eight subjects were given aspirin for one headache and Tylenol for another headache on a different day. Half of the group was given aspirin for the first headache and Tylenol for the second, and the other half was given Tylenol first then aspirin. Because the same person was given both aspirin and Tylenol, we can compare the differences in response time for the two treatments for each individual. We

can then test the differences to see if it is significantly different than zero; if it is, we can conclude that there is a difference between aspirin and Tylenol.

Subject:	1	2	3	4	5	6	7	8
Aspirin:	20	40	30	45	19	27	32	26
Tylenol:	18	36	30	46	15	22	29	25

```
DATA PAIRED;
  INPUT SUBJECT A_TIME T_TIME;
  DIFF = T_TIME - A_TIME;
PROC MEANS N MEAN STDERR T PRT;
  VAR DIFF;
RUN;
```

Analysis Variable: DIFF				
N	Mean	Std Error	T	Prob T
8	-2.250	0.750	-3.00	0.0199

Because the Prob |T| value is less than .05 we can reject the null hypothesis that the difference is equal to zero and state that the response time is shorter for Tylenol (because DIFF was computed as Tylenol time minus aspirin time).

Wilcoxon Test: A psychology experiment that measured the response to a stimulus had the following results:

Method A:	0	8	7	9	8	7	8	6	0	8	0	7
Method B:	4	5	5	4	6	3	4	4	5	4	5	5

The mean responses for method A is 5.667 and the mean for method B is 4.500. If you remove the zero responses from method A the mean increases to 7.556; this means that for the subjects that responded to method A the mean was 7.429. This is known as the threshold effect (some subjects don't respond at all to a stimulus). Data of this sort would inflate the standard deviation and make the t-test more conservative. To get around this we will use the nonparametric test, Wilcoxon, which does not assume a normal distribution. The Wilcoxon test first puts all the data in increasing order and ranks them, as follows:

S = Stimulus	M = Method												R = Rank											
S:	0	0	0	3	4	4	4	4	5	5	5	5	5	6	6	7	7	7	8	8	8	8	9	
M:	A	A	A	B	B	B	B	B	B	B	B	B	B	B	A	B	A	A	A	A	A	A	A	
R:	2	2	2	4	7	7	7	7	12	12	12	12	12	15.5	15.5	18	18	18	21.5	21.5	21.5	21.5	24	

Next, the sums of ranks for the A's and B's are computed:

A = 2 + 2 + 2 + 15.5 + 18 + 18 + 18 + 18 + 21.5 + 21.5 + 21.5 + 21.5 + 24 = 185.5
 B = 4 + 7 + 7 + 7 + 7 + 7 + 12 + 12 + 12 + 12 + 12 + 15.5 = 114.5

If the responses to the two methods were equal, we would expect the methods to be distributed equally among the ranks. If the response for method B was less than method A, we would expect the B's to be at the lower end of the rank ordering and therefore have a smaller sum of ranks than the A's. To test this we use PROC NPAR1WAY.

```
PROC NPAR1WAY WILCOXON;
  CLASS METHOD;
  VAR RESPONSE;
RUN;
```

NPAR1WAY PROCEDURE
 Wilcoxon Scores (Rank Sums) for Variable RESPONSE
 Classified by Variable METHOD

METHOD	N	Sum of Scores	Expected Under H ₀	Std Dev Under H ₀	Mean Score
A	12	185.5	150.0	17.097	15.458
B	12	114.5	150.0	17.097	9.542

Average Scores were used for ties
 Wilcoxon 2-sample Test (Normal Approximation)

S = 185.500 Z = 2.04715 Prob |Z| = 0.0406

Since our Prob |Z| statistic is small we will reject the null hypothesis that claims the sum of ranks are equal for the two methods. Therefore we can conclude that there is a difference in the two methods, with method A having a higher sum of ranks.

Comparing More than Two Sample Means

An Analysis of Variance (ANOVA) is used to compare the means of a number of groups to determine if there are any significant differences between them. The null hypothesis is that the means of all groups are equal. Assumptions are that the samples are independent, the populations are normally distributed, and the variance in all groups is equal (homoscedasticity).

One-way analysis of variance: In an example by Cody and Smith, 15 subjects are randomly assigned to three speed reading courses, X, Y, and Z. A reading test is given and the number of words per minute is recorded for each subject.

X	Y	Z
700	480	500
850	460	550
820	500	480
640	570	600
920 mean=786	580 mean=518	610 mean=548

Grand mean = 617.33

H_0 : mean(X) = mean(Y) = mean(Z)

To test the null hypothesis, we will use PROC ANOVA:

```
PROC ANOVA;
  CLASS GROUP;
  MODEL WORDS = GROUP;
  MEANS GROUP / DUNCAN;
  RUN;
```

Analysis of Variance Procedure

Class Level Information

Class	Levels	Values
Group	3	X Y Z

Number of observations in data set = 15

Dependent Variable: WORDS

Source	DF	Sum of Squares	Mean Square
Model	2	215613.333333	107806.66667
Error	12	77080.000000	6423.33333
Corrected total	14	292693.333333	
Model F =	16.78		PR>F = 0.0003

R-Square	C.V.	Std Dev	Words Mean
0.736656	12.98	80.1457	617.33333

Source	DF	ANOVA SS	F VALUE	PR>F
Group	2	215613.33	16.78	0.0003

Duncan's Multiple Range Test for Variable WORDS

Means with the same letter are not significantly different.

Alpha level = .05 DF=12 MS=6423.33

Grouping	Mean	N	Group
A	786.0	5	X
B	548.0	5	Z
B	518.0	5	Y

Because our p-value is low (0.0003) we reject the null hypothesis and conclude that the reading instruction methods were not all equivalent. To see which groups differ, we use the output from the Duncan option. The Duncan output shows that group X has the highest mean with 786 and that Z and Y are not significantly different (because they are both in grouping B). If we had not rejected the null hypothesis, we would have ignored the Duncan test all together.

Testing the Relationship Between Two Variables:

The Pearson correlation coefficient is used to show the strength of a relationship between two variables. The procedure assumes normally distributed populations. If one or both of the populations are skewed, you can use Spearman analysis. The Pearson correlation coefficient is a number that ranges from -1 to +1. A positive correlation means that as values on one variable increase, values on the second variable also increase (height and weight are positively correlated). A negative correlation means that as one variable increases, the other decreases (number of alcoholic drinks and score on driving test are negatively correlate).

```

DATA SAMPLE;
  INPUT HEIGHT WEIGHT @@;
  CARDS;
61 101 63 120 65 152 65 146 69 165 70 160 70 199 71 170 72 215
;
PROC CORR;
  VAR HEIGHT WEIGHT;
RUN;

```

Correlation Analysis
2 'VAR' Variables: HEIGHT WEIGHT
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	9	67.33333	3.90512	606	61.00000	72.00000
WEIGHT	9	158.66667	35.34827	1428	101.00000	215.00000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 9

	HEIGHT	WEIGHT
HEIGHT	1.00000 0.0000	0.90916 0.0007
WEIGHT	0.90916 0.0007	1.00000 0.0000

The results show that the correlation between height and weight is .90916 and the significance level is .0007. The small p-value indicates that it is unlikely to have obtained a correlation this large strictly by chance. It is important to remember that being significant is not the same as being strong or important. To test the strength we need to look at the correlation coefficient ($r=0.90916$) and square it ($r^2 = .82657$). We can now say that 83% of the variation in weight can be explained by variation in height. Another way to look at it would be that 17% ($1-.83$) of the variance of weight is due to factors other than height variation. The following, taken from Hatcher and Stepanski, is a guide for interpreting the strength of the relationship between two variables:

<u>Absolute value of the coefficient</u>	<u>Strength</u>
1.00	Perfect
0.80	Strong
0.50	Moderate
0.20	Weak
0.00	No correlation

Using this chart, we can see that the relationship between height and weight is strong. Even a weak correlation can be significant, so don't go by just the p-value.

If our data were not normally distributed or if one or more variables were ordinal, we should use the Spearman correlation. The only change to the program would be: PROC CORR SPEARMAN; The Spearman correlation is a distribution free test, that is, it makes no assumptions concerning the shape of the distribution. If you know the data are normally distributed, use the Pearson correlation, otherwise use Spearman. The interpretation of the Spearman coefficient is the same as the Pearson correlation.

Comparing Classification Variables:

Often times variables will be nominal or classification variables, that is, non-numeric values. Examples include gender, political parties, grouped age categories (21-25, 26-35 ...). To analysis variables such as these, use a Chi-square table.

Example: (From an example in Hatcher and Stepanski). A university administrator is preparing to purchase a large number of computers for three of the schools at the university. She is trying to decide if she should buy IBM compatibles or Macintosh computers. She sends out a two question survey: "Which school are you enrolled in?" and "Which type of computer do you prefer?". The program follows:

```

DATA COMPUTER;
  INPUT PREFER $ SCHOOL $ NUMBER;
  CARDS;
IBM      ARTS      30
IBM      BUS       100
IBM      ED        20
MAC      ARTS      60
MAC      BUS       40
MAC      ED       120
;
PROC FREQ;
  TABLES PREFER*SCHOOL / CHISQ;
  WEIGHT NUMBER;
  RUN;

```

The weight statement was used because we don't have the "raw" data. Instead we have final counts. If we had the raw data, the weight statement should be left off. We can see from the data that there are 140 students in the School of Arts and that 20 of them prefer IBM's and 120 prefer Macintoshes. Also, of the 140 students in the School of Business, 100 prefer IBM's. From the looks of it, it would appear that there is a significant difference. The results follow:

TABLE OF PREFER BY SCHOOL

PREFER	Frequency	SCHOOL			Total
	Percent	ARTS	BUS	ED	
	Row Pct				
	Col Pct				
IBM	30	100	20	150	
	8.11	27.03	5.41	40.54	
	20.00	66.67	13.33		
	33.33	71.43	14.29		
MAC	60	40	120	220	
	16.22	10.81	32.43	59.46	
	27.27	18.18	54.55		
	66.67	28.57	85.71		
Total	90	140	140	370	
	24.32	37.84	37.84	100.00	

STATISTICS FOR TABLE OF PREFER BY SCHOOL

Statistic	DF	Value	Prob
Chi-Square	2	97.385	0.000
Likelihood Ratio Chi-Square	2	102.685	0.000
Mantel-Haenszel Chi-Square	1	16.981	0.000
Phi Coefficient		0.513	
Contingency Coefficient		0.456	
Cramer's V		0.513	

Sample Size = 370

The computed chi-square value is 97.385 and its p-value is 0.000 (actually it isn't zero but very, very small). This means that there is less than 1 chance in 10,000 of obtaining a chi-square value this large if the variables were independent. Therefore, we can conclude that computer preferences is related to the school of enrollment.

Trademarks

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries.
 ® indicates USA registration.

References

Cody, Ronald and Smith, Jeffrey. Applied Statistics and the SAS® Programming Language. Prentice-Hall, Inc, Englewood Cliffs, NJ: 1991.

Hatcher, Larry and Stepanski, Edward. A Step-by-Step Approach to Using the SAS® System for Univariate and Multivariate Statistics. SAS Institute Inc., Cary NC: 1994.