Statistics 2: Comparing Mean Values, Graphs, Regression

For A Single Numeric Variable—The MEANS Command

The MEANS command is used when the variable of interest is numeric and measured on a continuous scale. A continuous variable can have decimal values (real numbers like 44.645) or integer values (44). In some ways, AGE can be considered either categorical (with 1-year categories) or continuous, but to use the MEANS command the Mean or Average value of the values must be of interest. Use the Read command to open the project ***'Sample.prj'*** from the. downloaded ***Epi info Projects*** folder. Examine the output from a request for MEANS AGE in the **Oswego** dataset.

T-test, ANOVA

The MEANS command can compare Mean values of a variable between groups of records. The numeric variable of interest, AGE, for example, is processed as for a single variable, but another categorical or "group" variable (such as ill/not ill) is used to divide the records into groups for comparison. MEANS of AGE cross-tabulated by ILL, for example, compares the ages of the ill and well persons and provides statistics to evaluate whether there is really a difference.

If there are only two groups, an independent t-test is performed. If there are more than two groups, then a one-way analysis of variance (ANOVA) is computed. "One way" means that there is only one grouping variable (in the above example: ill/not ill). If there were two grouping variables, such as ill/not ill and sex, then that would be a two-way ANOVA, which Epi Info does *not* perform. The one-way ANOVA can be thought of as an extension of the independent t-test to more than two groups. Because the ANOVA test requires some assumptions about the data and the underlying population, another test (Kruskal-Wallis, also known as the Mann Whitney/Wilcoxon test if there are only two groups) is also provided. This is a non-parametric test, meaning that it does not require assumptions about the underlying population. Non-parametric tests are more conservative in detecting a statistically significant difference, so a result that is "significant" in the non-parametric test will also be so in the ANOVA test.

The output is provided in different sections:

1. A table of descriptive statistics of the continuous variable by each group: number of observations, mean, variance, and standard deviation; minimum and maximum values; the 25th, 50th (median), and 75th percentiles; and the mode.
2. If there are only two groups, a T-Test table showing two different methods of performing the T-test (Pooled and Satterthwaite) and the corresponding t statistics and p-values.
3. An Analysis of Variance (ANOVA) table and a p-value for whether the means are equal.
4. A test to determine whether the variances in each group are similar (Bartlett's test for homogeneity of variance).
5. A non-parametric equivalent to the independent t-test and one-way ANOVA.

Most attention is usually paid to the p-value. If the p-value is <0.05 (or some other value used to define statistical significance), then there is a statistically significant difference between two or more of the means.

The ratio of the "Between MS" (Mean Square) and "Within MS" is what makes the *F statistic*, in this example, 1.560. If the p-value is ≥ 0.05, then there is no statistically significant difference between the means. In the ANOVA, the *F statistic* is always provided. The t-value is the square root

of the F statistic, so whichever is used (F or t), the p-value is the same.

The t-test for independent samples with unequal variances (Satterthwaite) is a procedure with two groups that can be used when the variances are not equal.

To compute confidence intervals around the mean values, use the following formula:

Mean ± (t-value) * (Within MS/n_i)∧0.5

Instead of the standard error, there is (Within MS/n_i)∧0.5, that is, the square root of the within mean square divided by the number of individuals in the group. The use of a standard error based on the within mean square is more accurate than if the variance for each group was used separately. In the example, the variance in age for those who are ill is 477 and for those not ill, 424; note that the within mean square is 457. Also note that the Bartlett's test for homogeneity of variance in the above example found that the variances can be assumed to be the same.

One of the assumptions of the ANOVA is that the variances in each of the groups are similar. One way to test this assumption is through use of Bartlett's test for homogeneity of variance, a fancy way of asking, "Are the variances about the same?" The variances are compared and a chi square value, the degrees of freedom, and p-value are presented. A note is also provided on the screen about how to interpret the results. In the example, because the p-value is >0.05, the variances can be assumed to be equal. If the variances are not similar, then: 1) the non-parametric results discussed next should be used; and 2) the data could be transformed.

The MEANS command provides a non-parametric equivalent to the ANOVA. The non-parametric results should be used when: 1) the means of each group is not a good measure of centrality (i.e., the data are skewed or have some other non-normal distribution); 2) the data are rankings or ordinal data rather than precise numeric values (e.g., using Likert scales where 1=strongly agree and 7=strongly disagree with intermediate levels of agreement is an example of ordinal data; compare this to cholesterol values, which are precise quantitative values); 3) there is a small sample size in each group with numeric data (which Rosner suggests should be less than 10 in each group); and 4) the variances are significantly different between the groups. (NOTE: The t-test for independent samples with unequal variances is a procedure with two groups that can be used when the variances are not equal).

The non-parametric equivalent to the t-test where two groups are compared is the Mann-Whitney U test, the Wilcoxon rank-sum test (called the Wilcoxon two-sample test in the output from ***Analysis***). If there are more than two groups being compared, the Kruskal-Wallis test is the non-parametric equivalent to ANOVA.

In the example, the p-value is 0.28, which would indicate that there is little difference in the rankings of age between those who were ill and those who were not ill, leading to the same conclusion as the p-value from the ANOVA table.

Transforming Data

As an example of transforming data, read in Course.prj (located in the Course folder in the downloaded *Epi info Projects* folder) and open **LEADPRAC**.

1. Examine the MEANS of baseline lead values (LEAD0) cross-tabulated by Location (LOCATION).
   - Check normality: Are the medians close to the means?
   - Are the ranges and quartiles equidistant from the medians?

- Bartlett's test indicates that the variances are nonhomogeneous.

2. Define a new variable called TransLead. Assign the value of this variable equal to LN(LEAD0) to obtain a logarithmic transformation of the lead values.
3. Now perform the MEANS command again on TransLead and note that ANOVA can be used appropriately because the group variances are homogeneous.

Paired t-test

In past versions of Epi Info this was possible, but now it is not, because the current version does not report whether mean values differ from zero. However, one can easily perform a paired t-test in Microsoft Excel. An Epi Info table can be exported (WRITE) to an Excel spreadsheet format.

Graphs

Using the form **LEADPRAC** create the following graphs with the Graph command under Statistics:

1. Under Graph Type, select **Column**
2. Under the Independent Axis section, select LOCATION from the **Main Variable(s)**
3. Under the Dependent Axis section, select ***Show Value of***...Count
4. Under the Series section, select ***Bar of Each Value of***...Sex
5. Click **OK.** The Column chart appears in the Output window
6. Click on the small spanner icon that appears beside the chart to open the chart configuration dialog. You can set the Chart and Legend titles, and the X and Y axes labels. You can save or print the image as required

Unlike the GRAPH command in earlier versions of Epi Info, that of this version of Epi Info 7 appears to offer only limited options for customization. Most of the functionality of the GRAPH command is now in the Visual Dashboard.

1. From the Epi Info main menu select Tools > Analyze Data > Visual Dashboard or click the Visual Dashboard button and select the LEADPRAC dataset
2. Right click anywhere on the canvas and select Add Analysis gadget > Charts > Column chart.
3. Select LOCATION from the **Main variable** dropdown list. Click on the Advanced options dropdown arrow to expand it. Select ***Stratify by***...SEX
4. Click on the Display options to expand it. Select **Show annotations** under ***Color and Styles***. Select **Show legend**.
5. Click **Run**

Try creating the following charts:

- Pie chart of EyePencil
- Scatterplot of AGE, LEAD0
- Pareto chart of LEAD0

Linear Regression

The Linear Regression command can be used for simple linear regression (only one independent variable), for multiple linear regression (more than one independent variable), and for quantifying the relationship between two continuous variables (correlation). Regression is used when the primary interest is to predict one dependent variable (**y**) from one or more independent variables ($\mathbf{x_1, \ldots x_k}$).

The correlation coefficient or ***r*** (sometimes referred to as the Pearson correlation coefficient) is a useful measure of how two continuous variables are related. If the correlation is greater than 0, the variables are **positively correlated**; as x increases, y also increases. If the correlation is less than 0, the variables are **negatively correlated**; as x increases, y decreases. If the correlation is exactly 0, then the variables are **uncorrelated**. The correlation coefficient can vary between +1 and −1. For positive correlations (r > 0), the closer to +1, the stronger the correlation; for negative correlations (r < 0), the closer to −1 the stronger the correlation.

If the data are ordinal or far from normal, significance tests based on the Pearson correlation coefficient are not valid and a non-parametric equivalent to Pearson's should be used.

Simple Linear Regression

Examine a simple linear regression on the relation between Estriol and Birthweight (BW) in the EstriolAndBirthweight dataset of the Sample.prj project. Select **Advanced Statistics > Linear Regression**. Select Birthweight as the Outcome Variable and Estriol in the list of Other Variables.

- **Coefficient** The slope of the line, sometimes referred to as the "regression coefficient." In this example, 0.608 can be interpreted as: For every 1 unit increase in estriol (1 mg/24 hr), there is a 0.608 increase in each birthweight unit (g/100). The standard error of the slope, 0.147, is also provided and can be used to calculate confidence intervals. 95% Confidence intervals of the coefficient can be calculated as ± 1.96*SE. P value indicates if the coefficient is significantly different from zero.
- ***Y*-Intercept** This is where the line intercepts the *y* line. In this example, the line would intercept the (birthweight) line (y) at 21.5. The general form of the simple linear regression line is:

  $$y = a + bx$$

  where *y* is the dependent variable, *a* is the intercept, and *x* is the independent variable. In the above example, the regression line is:

  BIRTHWEIGHT = a + b (estriol)

  BIRTHWEIGHT = 21.5 + 0.61 (estriol)
- For any given value of estriol, a BW value can be predicted. P value indicates whether the *y*-intercept is significantly different from zero.
- **r^2** Sometimes represented as r^2 or R^2 (i.e. R squared). The R^2 value = Regression Sum of Squares / Total Sums of Squares (in the above example, 250.57/674 = 0.37). The R^2 can be thought of as the proportion of variance of y (in this example, birthweight) that can be explained by x (in this example, estriol). In this example, 37% of the variance in birthweight can be explained by the women's estriol levels. If $R^2 = 1$, then all of the variability is explained, which would mean that all data points fall on the regression line. If $R^2 = 0$, then no variance is explained.
- **Correlation coefficient** The Pearson correlation coefficient, or "r". In this example, the correlation is the square root of $R^2 = 0.61$, indicating a relatively strong positive correlation

between estriol and birthweight.

- **F-Statistic** The F-statistic is the Regression Mean Square / Residual Mean Square (in the above example, 250.57/14.6 = 17.1). The F-statistic is calculated to determine if the slope of the regression line is significantly different from 0. Epi Info does not provide the p-value corresponding to this value of F.

Multiple Linear Regression

Using the **BabyBloodPressure** dataset, the dependent variable is systolic blood pressure, and the independent variables are birthweight in ounces and age in days. The independent variables should be selected as Other Variables.

Variable	Coefficient	Std Error	F-test	P-Value
AgeInDays	5.888	0.68	74.9229	0.000001
Birthweight	0.126	0.034	13.377	0.002896
CONSTANT	53.45	4.532	139.1042	0.000000

Correlation Coefficient: r∧2= 0.88

Source	df	Sum of Squares	Mean Square	F-statistic
Regression	2	591.036	295.518	48.081
Residuals	13	79.902	6.146	
Total	15	670.938		

An adjusted r^2 value takes into account the model's associated degrees of freedom. The adjusted r^2 value = 1 – [Total df/Residual df] * [Residual SS / Total SS] (in the example, 1 – [15/13] * [79.9/670.9] = 0.86). The adjusted r^2 can be thought of as an r^2 value that adjusts for the number of independent variables. Unlike the r^2, which will always increase as the number of independent variables increase, the adjusted r^2 can become smaller.

The F-statistic can be used to determine the p-value by use of a table. Find the p-value of F (numerator df), (denominator df), (F-statistic), which in this case is F,2,13,48.08, in which the p-value is < 0.001, thus concluding that the two variables together are significant predictors of SBP.

The regression line is:

SBP = 53.45 + 0.126*BWT + 5.888*AGE

The Linear Regression command also allows dummy variables to automatically be created from dichotomous variables (assigning the values of 0 and 1).

Open the project COURSE and select the table Absorption (make sure Tables is selected) to Read. Assign Abs (the calcium absorption in rickets) as the outcome variable and XRTOTAL and BEADING as the independent variables. Because BEADING is dichotomous, select it in the list and click on the button MAKE DUMMY. Note in the results that the coefficient for BEADING indicates that the absorption fraction is about 9% higher when rib beading is present (true) compared with when it is not.