More Excel Formulas and Procedures

 Spreadsheet Formulas for Statistics:
  
  TTEST(data set 1, data set 2, number of tails, type)
  type takes values 1, 2, or 3 and specifies
  1: a matched-pairs test (must have equal sample sizes)
    problems 11.41 and 11.43
  2: a pooled-variance test
    problems 11.8, 11.11
  3: an unequal variances test
    problems 11.17, 11.23

  FDIST(calculated F value, numerator df, denominator df)
  FDIST returns the p-value of the calculated F for the specified degrees of freedom
    problems 11.69, 11.73

  FINV(probability, numerator df, denominator df)
  generally, the probability here is the significance level of the test; FINV returns the
  critical t value for the specified significance level and degrees of freedom
    problems 11.69, 11.71

 Statistical Analysis Tools: All begin from Tools/Data Analysis

  t-Test Paired Two-Sample for Means
  This procedure assumes that the data are a paired sample and tests the hypothesis
  H\(_0\): \( \mu_d = 0 \), that is that the mean difference between pairs of observations is zero. The output
  gives the two sample means, the calculated value \( t = \frac{\bar{d}}{s_{\bar{d}}} \) and the one- and two-tailed
  critical values of \( t \) and p-values for the hypothesis test.
    formulas page 430
    problems 11.41 and 11.43.

  t-Test: Two Sample Assuming Equal Variances
  This procedure assumes that the two samples are drawn from populations with equal
  variances and uses the pooled standard deviation in calculating the standard error. The procedure tests the hypothesis
  H\(_0\): \( \mu_1 = \mu_2 \). The output gives the two sample means, two
  sample variances, the pooled variance, the calculated t statistic \( t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} \), and the
  one- and two-tailed critical values of \( t \) and p-values for the hypothesis test.
    formulas page 412
    problems 11.11 and 11.14.

  t-Test: Two Sample Assuming Unequal Variances
  This procedure and its output differs from the Equal Variances procedure only in that the standard error and degrees of freedom are calculated differently. The procedure assumes
  that the two samples are drawn from populations with different variances.
    formulas pages 418 – 419
    problems 11.23, 11.25

  F-Test Two-Sample for Variances
  This procedure tests the hypothesis H\(_0\): \( \sigma_1^2 = \sigma_2^2 \). To use a pooled-variance t-test for equality of population means (t-Test: Two Sample Assuming Equal Variances), we must be unable to reject this null hypothesis. The output gives the two samples’ respective variances, the ratio of the variances (the F statistic) and the one-tailed p-value for the calculated F ratio. Note that when using Excel, it is immaterial which variance is the numerator and which the denominator.
    formula page 443
    a problems 11.65, 11.66
Analysis of Variance (ANOVA)

ANOVA is used to test the null hypothesis $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$, that is to test the proposition that the means of three or more populations are equal. The alternate hypothesis is that at least one population mean is not equal to the others. In Excel ANOVA is a Data Analysis toolbox function. To reach it, from the task bar at the top of the screen, choose Tools, then choose Data Analysis. From the Data Analysis menu, choose Anova: Single Factor.

In the dialog box, for the Input Range, mark the range containing the data. Excel will interpret each column as a treatment group. It is wise to mark the column headings as well, but be sure to check the “Labels in first row” box. The default is to put the output on a newly created worksheet; to avoid this, click the radio button for “Output range,” and designate the upper left corner of where you want the output to appear. When you click OK the output will appear as below:

Anova: Single Factor (Output is for text problem 12.94, p. 537)

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp1</td>
<td>3</td>
<td>139</td>
<td>46.3333333</td>
<td>20.3333333</td>
</tr>
<tr>
<td>Comp2</td>
<td>4</td>
<td>201</td>
<td>50.25</td>
<td>14.9166667</td>
</tr>
<tr>
<td>Comp3</td>
<td>3</td>
<td>132</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Comp4</td>
<td>5</td>
<td>272</td>
<td>54.4</td>
<td>37.3</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>242.9833333</td>
<td>3</td>
<td>80.9944444</td>
<td>3.22084312</td>
<td>0.0651457</td>
<td>3.587434</td>
</tr>
<tr>
<td>Within Groups</td>
<td>276.6166667</td>
<td>11</td>
<td>25.1469697</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>519.6</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The P-value entry gives the p-value of the test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. If our test is at 5% significance level, we should fail to reject the null hypothesis and conclude that the various alloys produce the same hardness. Examining the column Average, we can see that there is indeed little difference in the mean hardness among the four samples of alloys.

- ANOVA formulas are on page 487
- Problems 12.23, 12.27; p-values are 0.07494 and 0.0429

- $\chi^2$ procedures and formulas

CHIDIST(x, degrees_freedom) gives the area to the right of the x value entered on a $\chi^2$ distribution with the specified number of degrees of freedom. CHIDIST(16.2, 11) returns 0.133867. This would be the p-value for a goodness of fit test or any upper one-tailed test.

- problems 13.15, 13.17, 13.19

CHIINV(probability, degrees_freedom) gives a $\chi^2$ value such that the proportion of the distribution which exceeds that value is the probability entered. CHIINV(0.05, 17) returns 25.58711. 5% of the $\chi^2$ distribution with 17 degrees of freedom lies beyond 25.58711. This would be the critical value for an upper one-tail test at 5% significance level.

- problems 13.7 and 13.9
Goodness of fit tests for uniform distribution:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Observed</td>
<td>Expected</td>
<td>Squares</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>15</td>
<td>=(A2 – B2)^2/B2</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>15</td>
<td>COPY</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>15</td>
<td>COPY</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>15</td>
<td>COPY</td>
</tr>
</tbody>
</table>

The $\chi^2$ statistic would be the sum of the values in column C, and chidist or chiinv formulas could be used to test hypotheses.

Problems 13.17, 13.19

Goodness of fit tests for the normal distribution: Use the normsdist function to establish the proportions of the sample expected to fall into various ranges. The simplest approach is to use ranges defined by the number of standard deviations from the mean. The class example used mean to – 1 st dev, – 1 to ~2 st dev and more than – 2 st dev, for example. But we need not use an even number of standard deviations. We could for example use mean to ± 0.5 st dev, – 0.5 st dev to – 1.5 st dev, and so on. The expression normsdist(-1) – normdist(-2) will give the area between – 1 and – 2 st dev. These proportions are then applied to the sample size to find the expected number in each range. The observed frequency – the number in each range in the sample – can most easily be found by using the Histogram tool in the Data Analysis tool box. Mark the data as the input range and mark the ranges you’ve established as the Bin range. You will now have observed and expected frequencies, and from here you proceed exactly as with goodness of fit for the uniform distribution.

Refer to the worked-out spreadsheet posted as “normaltest.xls”

Problems 13.25, 13.27

Further note on the Histogram tool. The designated bin range must contain only single numbers, which represent the top of each range. Do not try to enter something like 200 – 300 as a bin; enter only 300. The following is an example:

<table>
<thead>
<tr>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
</tbody>
</table>

If these four cells are designated as a bin range, the first will be interpreted as a label (be sure to check the “Label” box); Excel will then count the number of entries up to and including 40; the number over 40 up to and including 60; the number over 60 up to and including 80; and the number over 80. (No matter what your largest number, you’ll always get a count of the number that exceed it.)

$\chi^2$ contingency tables or tests for independence

The data are entered into a spreadsheet as below:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>Totals</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Jorge</td>
<td>23</td>
<td>81</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Sam</td>
<td>45</td>
<td>62</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bill</td>
<td>67</td>
<td>63</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To use the CHITEST formula we must calculate the expected number for each cell. Now, the expected number for cell B3 will be the total from B6 divided by the total from E6 (total sample size) multiplied by the total from E3, so =B3/B6*E3. The problem with that expression is that it will not copy properly; some of the cell addresses must be frozen. In particular, as we copy B3
should change to C3 then D3, while E3 should change to E4, then E5. In Excel we can freeze a cell address under the copy operation by inserting a $ in front of the row or column address we wish to freeze. So drop down a few rows and enter the following:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td></td>
<td>=B$3/SBS6*$D3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The expression in B9 may now be copied down and across to create the expected numbers for each cell. Finally enter CHITEST(B3:D5,B9:D11). The result is the p-value of the test $H_0$: these factors are independent, or Jorge, Sam and Bill are all the same with respect to characteristics A, B and C.

- Refer to the worked-out spreadsheet which appears as sheet2 in normaltest.xls
- Problems 35, 37, 39, 41