Oneway ANOVA: Analysis and Assumption Checking using Minitab

Review: Recall that Oneway ANOVA is a method for comparing the means of I different populations or groups by testing the ANOVA hypotheses

H_0: $\mu_1 = \mu_2 = \cdots = \mu_I$ vs.
H_1: *at least two means differ*

These hypotheses are tested using the F-ratio test statistic F = MSTr/MSE. The larger the F, the stronger the evidence against H_0 in favor of H_1. Therefore the p-value for testing H_0 vs H_1 is given by $P(F \geq F_{\text{actual}})$, computed assuming H_0 true, where F_{actual} is the F-ratio for the actual sample data.

Overview: In the following we detail how to analyze ANOVA data using Minitab's `One-Way` routine to compute the F-ratio test statistic and the ANOVA residuals. Then we state the ANOVA assumptions and describe how to use Minitab's `Stat -> ANOVA -> Test for Equal Variances` routine and the residuals to check them.

Using One-Way: We will use Minitab's ANOVA routine `One-Way`, not `One-Way (Unstacked)`, because it computes the residuals we need to check the ANOVA assumptions:

Stacking Data: In order to use `One-Way` we need the data to be in stacked form, that is, one column ("response") contains all the values and a second column ("factor") identifies the sample to which each value belongs. If the data is not stacked, you can stack it using

```
Data -> Stack -> Columns
```

(For details on stacking, use *Help* -> *Search Help* and enter "stack.")

Calling `One-Way`: To invoke `One-Way` do the following:

1. Click on `Stat -> ANOVA -> One-Way`
2. Select the `Response` column containing the values
3. Select the `Factor` column indentifying the samples
4. Check the `Store residuals` box then click OK.

Interpreting `One-Way` Output: Below is the output after stacking and analyzing the data in the Minitab data set `www.rose-hulman.edu/~inlow/bearing.MTW`:

```
One-way ANOVA: Vibration versus Brand

Source  DF     SS     MS     F      P
Brand    2  20.70  10.35  8.99  0.003
Error   15  17.28   1.15
Total   17  37.98
```

This data set - which consists of the vibration levels of three brands of motor bearings - was stacked into a response column called"Vibration" and a factor column called "Brand." The F test statistic (F-ratio) is 8.99 and the p-value is 0.003.

Anova Assumptions

For the p-value associated with the F-ratio to be accurate the following four assumptions must be met

- The I samples are IID
- The I samples are independent
- The I populations/processes have equal variance (homoscedasticity)
- The I populations/processes are normally distributed

The first two assumptions describe the data acquisition process; the last two describe the populations/processes under study.

Checking the Anova Assumptions

Sampling Assumptions (1 and 2): Determining if the data satisfies assumptions 1 and 2 requires knowledge of how the data was acquired/generated. A common violation of this assumption 2 is the case where the same subjects comprise all I samples, that is, the same subjects are measured multiple times. This results in what is called "Repeated Measures ANOVA," the I sample analog of paired data.

Population/Process Assumptions:

Equal Variances (3): To check this assumption, i.e., to verify that the populations or processes are *homoscedastic* (have "same scatter") you need the data to be in stacked form, that is, one column contains all the values and a second column identifies the sample to which each value belongs. Once the data is in stacked form, you can test for equal variances using

```
Stat -> ANOVA -> Test for Equal Variances
```

Select the `Response` and `Factor` columns which are same as those for `One-Way`. Using the Levene test p-value, reject H_0: *variances equal* if the p-value is less than or equal to 0.05. For the motor bearing vibration data above, the Levene p-value is .585, so we don't reject the equal variance assumption.

Normality (4): To check normality you need to do the following:

1. Check the `Store residuals` box on the `One-Way` dialog box. This tells `One-Way` to create a column called `RESI1` or similar.
2. Test the normality assumption by testing the residuals in `RESI1` using

   ```
   Stat -> Basic Statistics -> Normality Test...
   ```

   Reject H_0: *populations/processes normal* if the p-value is less than or equal to 0.05.

For the motor bearing vibration data, the p-value is 0.406 so we don't reject the normality assumption.

Exercise 1: In this example you will analyze data from an experiment to investigate the relationship between the strength of paper grocery bags and the hardwood concentration in the pulp. The goal was to determine how the strength of the bags increased as the hardwood concentration was increased. Was there, for example, a point of diminishing return? The engineering team decided to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. Six specimens from each concentration level were tested. The resulting specimens were tested on a lab tensile strength tester in **random order**.

i. Download the data from the following URL

`www.rose-hulman.edu/~inlow/hardwood.MTW`

and then put it into "stacked" form, i.e., the measurements in one column and the factor level (hardwood concentration) in a second column. You can either use Minitab's `Data -> Stack` routine or cut and paste the data.

ii. Are there differences in the mean strengths of the grocery bags due to differences in hardwood concentration? State the appropriate hypotheses to test and be sure to define all parameters.

iii. Test the hypotheses in part ii at significance level $\alpha = 0.05$. What is your test statistic and p-value? What do you conclude?

iv. Assess the ANOVA assumptions. Are they met?

v. What hardwood concentration would you recommend that they use and why?

Exercise 2: In this example you will analyze experimental data collected and analyzed for a senior ABBE project. The Rose senior and his advisor wanted to compare the average size of cells from three different populations. Duplicate their analysis by doing the following:

i. Download (open) the data set `cell size` provided in the week 10 handouts section on the course website.

ii. Put the data into "stacked" form, i.e., put the measurements (cell sizes) into one column and the factor level (population number) into a second column.

iii. Analyze the data and have Minitab compute the residuals.

iv. Assess the ANOVA assumptions. Are they met? Justify your conclusions and report all p-values.

v. Assuming the ANOVA assumptions are met, what do you conclude about possible differences in cell sizes across the three populations at $\alpha = 0.05$? What is your F-ratio and the corresponding p-value?