1 Commands in JMP and Statcrunch

Below are a set of commands in JMP and Statcrunch which facilitate a basic statistical analysis. The first part concerns commands in JMP, the second part is for analysis which is easier in Statcrunch.

2 Basic Data handling

2.1 Downloading data from the web

- The data I post on my webpage will be either in a zipped directory containing a few files or just in one file containing data. Please learn how to unzip zipped documents (this is usually done by extracting the directory).
- Saving data on my website to your director:

  Right click on the link (the file or the zipped directory you want to save). Then click on `Save Link As...` By pressing on here it will give you the option of saving the data to your chosen directory.
- Once saved, if the file was zipped you will need to unzip it.

2.2 Opening a data file in JMP

- Once you have saved the data in your specified directory it is simple to upload it in JMP. Below are instructions on how to do this.
- Open JMP.
- For the files that I will be posting you need to change the preferences. To do this
  - Click on `Preferences`.
  - You will see a new box and on the LHS a column with many options. click on `Text Data Files`.
  - In import Setting tick the box `Space` and `Spaces`. This should mean that the data you load into JMP will give a new column for separations by spaces.
  - Save your preferences.
- Click on Open Data Table (or open which is in File on the top LH corner).
- Make sure that in the Open as option, the option Data, using Text Import Preferences, is marked.
- Go to the directory which contains your data.

- Make sure that you select Files of type: All files (*.*).
- You should see the name of the data file you want to load into JMP, double click on this file.
- You should see a spread sheet with all the data from this file (with several columns).

2.3 Saving JMP output

Using the export command saves the output as a `.png` file which can be easily included in any work document (as an image) or a latex file (if you are using latex you need to use the command `includegraphics`.

- Click on File $->$ Export $->$ Then click on Image. Press Next. Once you do this you can save in whatever folder that you wish.
- Alternatively, you can take a screen shot (in a mac, you make sure that you are in Preview, select Take Screen Shot and then a screen shot of what you like by going selecting From Selection).

3 Summarising statistics (such as mean, standard deviation) in JMP

- Click on the little box below Analyze (that if you put the curser over it is called summary).
- If you want the data grouped (such the 8 or 9 M&M grouping), then highlight the column containing the group code and then click on the button Group. The column name should be in this part of the table.
- The column of data that you want analysed should be highlighted in *section columns*.
- In the table containing *Statistics* scroll through what you want summarised, eg. sample mean, sample size, standard deviation (all for ech group you have selected the group option) etc.
- Click on okay and the data should be summarised in a new table.

3.1 Plotting Histograms in JMP

- I have posted on my page some data (in a zipped directory called Student651Data), this data is for you to play with. Student651Data contains height and course information data from a 651 class.

- You should see the data on a JMP spread sheet infront of you.
- The third column in this data are the heights of students (this is a numerical variable), the fourth column is a column of zeros and ones, one indicates a female and zero a male (this is a binary variable).
- We want to make a histogram of all the data and the male/female data.
- To do this go to `Analyze -> Distribution`.
- To make a relative frequency histogram of all the heights (regardless of gender), highlight the third column (which in this case are the heights), and press Y. columns. This should put the third column here. Then press `Okay`. You should get a histogram of all heights on a separate page (to change orientation of the plot or other settings left click on the Y-variable title (in our case column 3) and select the desired option).
- The same thing can be done by going to `Graph -> Chart` double clicking on Column 3 and pressing okay.
- **Changing the range in the axis**

  Move the cursor over the x axis until the 'hand' appears and double click, then change the x axis range. We wil need to do this for every histogram.

3.2 Making a boxplot in JMP (single sample)

- Input data in JMP. JMP does not know the name boxplot, but it does give automatically give a boxplot when you make a histogram using using the distribution option (detailed above).
- You can get various types of Boxplots (Quantile and Outlier) and other options by manipulating the output. Right click on the Y-variable title and choose the option that you want.

3.3 Comparing mutiple samples using boxplots and histograms

- Boxplots and Histograms are excellent for comparing multiple samples.
- Do this is JMP with the height data, comparing male and female heights.
- **How the data should appear in the table** To compare samples you need to make sure that there are (at least) two columns in your spreadsheet. One column should contain your Y-value such as height of a randomly selected person and the other column should contain the factor (often coded as a number) for example 0 if the

person corresponding to the height is female and 1 if the person corresponding to the height is male.

(i) Go to `Analyse -> Distributions`. You should see a box pop up. Put the height column into the `Y, Columns` Box (highlight the height column and then the y, Columns) and put the factors (gender classification) into the `By` Box. Then press okay.

- You should get both boxplots and histograms.
- The default boxplot is Outlier, you can get a quantile boxplot by going to the output (much of the stuff in JMP can be done by manipulating the output) and right clicking on the title of the Y-axis (in this example it is called Column 3), and selecting Quantile Box Plot. I prefer the quantile boxplot.
- To compare several boxplots it is important to aign them and make the scale uniform.

  To align multiple box plots, right click on Distributions. For example, in the case of the height data right click on both `Distributions Column 4=0` and `Distributions Column 4=1`. Select or equivalently tick `Stack` and select or tick `Uniform Scaling`.

  This should stack the boxplots and also align they scales - in order to make a fair comparison.
- If you only want to compare the boxplots and not the histograms, then get rid of the histograms by right clicking on the column name (in the case of the height example,it would be Column 3) and unselecting or equivalently unticking the `Histogram` in `Histogram Options`.

3.4 Making QQplots in JMP

- Load data in JMP
- Go to `Analyse -> Distributions`. You should see a box pop up. Put the height column into the `Y, Columns` Box (highlight the height column and then the y, Columns) and put the factors (gender classification) into the `By` Box. Then press okay.
- By default you will get a histogram and boxplot.
- You can get a QQplot by manipulating the output. Right clicking on the name of the Y-axis (eg. Column 3) and select Normal Quantile Plot. You should now see a QQplot next to the boxplot. Compare all three plots.

3.5 Making one sample confidence intervals

- Load data into JMP.
- Go to `Analyse` → `Distribution`.
- Put the column you are interested in, into the box [Y, columns] and press `okay`.
- In the output rotate the plot round (to make it look better), by right clicking on the name of the column, `Histogram options` → `vertical`.
- To make a CI right click on the column name, and go to Confidence Interval, and select the confidence level you want.
- In the out put you should get the CI.
- The first column should contain the parameter, the second the estimator of that parameter (for example, the sample mean)

3.6 Doing a one sample t-test/z-test

- Here you want to do a two-sided or one-sided z or t-test in JMP (ie. test $H_0 : \mu = 5$ against $H_A : \mu \neq 5$ or $H_0 : \mu > 5$ against $H_A : \mu \leq 5$).
- Load data into JMP.
- Go to `Analyse` → `Distribution`.
- Put the column you are interested in, into the box [Y, columns] and press `okay`.
- Right click on the name of the column and click on `Test mean`.
- Here you will have two boxes, specify mean and Enter True standard deviation.
- In Specify mean, give the mean you want to test (the null mean).
- In Enter true standard deviation, you have the option of leaving this blank.
  - If you enter a standard deviation it will do the test as if the standard deviation (population variance) were known. And it will use the normal distribution.
  - If you leave this box blank, it will use the estimated standard deviation s from the data. In this case it will use the t-distribution instead of the normal distribution.
  - Experiment with this. Leave this column blank and do the test. Also put the standard deviation you get from the output and do the test.

    You should see that for large sample sizes there is very little difference. But for smaller sample sizes, the p-value when using the t-distribution (the column is kept blank), should be a little larger than when you input the standard deviation from the output.

- The results of the test should be given in the box Test Mean. It does both the one sided and two sided tests.
- You will get the distribution under the null and the sample mean.

4 Comparing the means of two populations from two independent samples

4.1 Comparing the means of two populations using the independent sample t-test

- Here you have two (completely independent) samples from two populations. The samples can be of different sizes.

  Our aim is to infer something about the differences in the population means based on the sample means of both samples.

  We are assuming that the sample sizes are relatively large and/or there are not many outliers (small sample sizes and many outlier can effect the results in the test). If this is true, then we use an independent sample t-test (see Lecture 21 for details).

(a) You input this data as two columns.

One column is the response variable (this would be the heights of people etc.)

The other column is the factor, this is an indicator variable, which discriminates between groups. For example, it would give 0 for males and 1 for females.

(b) Before you do the analysis make sure that the response variable is labelled as a continuous random variable and the factor is labelled as a nominal random variable.

You do this by going to the variables which should be listed in column on the left of your data. Left click on the variable which is your response variable (for example, this will Column 3 for the 651 height data) and make sure that continuous is ticked. Left click on the factor (which is Column 4 of the 651 data) and make sure that Nominal is ticked.

(c) Go to `analyze` → `Fit X by Y`. You should see a new box pop up. Highlight the response variable (in Select columns - for example Column 3 for the 651 heights) and put it into the Y,response box (by clicking on Y, response button). Similiarly, highlight the factor (in Select columns - for example Column 4 for the 651 example) and put it in the X. factor box.

(d) Press `Okay`.

(e) You should see two colums both with dots on them indicating the values of the response variable for both factors.

Observe the distribution of the points and whether they ‘look’ they come from the same population or not.

(f) To do the test and obtain CIs. We need to right click on the One Analysis Bar and select `t-test`. If you want to do the test under the assumption that the standard deviation (and equivalently the variances) of both populations are the same, then select `Means/Anova/Pooled t`.

(g) You will get a lot of output and a distribution.

In the output you will see the difference of the sample means, the standard error $\sqrt{\frac{s^2}{37}+\frac{s^2}{27}}$, the total number of observations in both samples, a 95% CI for the mean difference $\mu_X - \mu_Y$, and results of the tests $H_0 : \mu_X - \mu_Y = 0$ against, $H_A : \mu_X - \mu_Y \neq 0$, $H_0 : \mu_X - \mu_Y \geq 0$ against, $H_A : \mu_X - \mu_Y < 0$, and $H_0 : \mu_X - \mu_Y \leq 0$ against, $H_A : \mu_X - \mu_Y > 0$.

You should understand the output and be able to construct 99% CI (90% etc) and do test for other alternatives.

4.2 Comparing the means of two populations using the Wilcoxon sum rank test

- In the case that the sample sizes (of either sample) is small (say less than 30) and/or there are several outliers (can be checked by making box plots and QQplots), the assumptions on normality will not be strictly true and we use instead a nonparametric test.
- Here we will use a Wilcoxon sum rank test (as described in Lecture 22).
- To do the test in JMP follow steps (a-e).
- To do the test. Right click on the One Analysis Bar and select `nonparametric` and click on `Wilcoxon test`.
- The most important table to look at is the first table, where the sums of the orders from both samples is given . See `two_sample_independent_wilcoxon_JMP.pdf` for the details.

5 Comparing population means when there is dependence between the pairs (this is for paired data)

5.1 Matched data: The paired t-test

- Here you have two samples from two populations. The samples are both of the same size and paired. This means there is a dependence between the two pairs. For example, runners running at both high and low altitude. There is a dependence because each pair involves the same runner.

Our aim is to infer something about the differences in the population means based on the sample means of both samples. For example, how much harder is it on average to run at a high altitude than a low one.

We are assuming that the sample sizes are relatively large and/or there are not many outliers (small sample sizes and many outlier can effect the results in the test). If this is true, then we use an independent sample t-test (see Lecture 23 for details).

- If you are not sure there is a dependence between the pairs (or it is not clear from the way the data was collected that there is), then make a plot of the the pairs against each other.

  To make the plot:

  (i) Go to `analyze`. Then go to `Fit X by Y`. Highlight one column and press `Y, response`. Then highlight the other column and press `X, factor`.

  (ii) You should get a plot of one column against the other. Each pair is one x-y value.

  (iii) If you see 'random dots', there appears to be no dependence.

  If on the other hand, you see some sort of line, then there does appear to dependence.

(a) Input the data into JMP. There should be two columns, the rows of the table which contain the pairs. Ie. the first column contains the run times of several runner at a high altitide and the second column contains the run time of the same runners at a low altitude.

(b) Go to `analyse`. Press in `matched pairs`.

(c) Highlight Column 1 and press `Y, Paired Response`. Then highlight Column 2 and press `Y, Paired Response`. Both should now be in the right hand box. It does not matter which order you do it. The only difference is the way JMP will do the subtraction.

The way I describe, JMP will calculate sample mean in column 2 minus sample mean in column 1. If you do it the other way around, JMP will evaluate it the other way around. This matters when you do a one-sample test and need to keep track of which way round you do the hypothesis.

(c) Press `okay`.

(d) You will get a whole load of output. Look at `paired_t_test_runner_JMP.pdf` to understand the output.

5.2 Paired data: Wilcoxon sign rank test

- We do this test for paired data (like the paired t-test), but when the number of observations is small and/or there are many outliers in both samples.

(a) Do the same was in (a-c) of the above (paired t-test).

(b) Now right click on `Matched pairs` and click on `Wilcoxon sign rank`.

(c) You will get a whole load of output. Look at `wilcoxon_sign_rank_runner_JMP.pdf` to understand the output.

6 Comparing the means of multiple populations - one way ANOVA

- Suppose we have five samples taken from fve populations (in general this can be k samples (of any size) from n populations).

- We want to see whether the mean in all three populations are the same against the alternative that at least one is different.

  Ie. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against H_A : at least one mean is different.

(i) The data needs to be inputed into JMP as two columns, on containing all the data and the other indicating which group (or population) it comes from.

(ii) Input the data into JMP, and make sure that the observations (samples, such as the speed of light data is continuous random variable) and the groups is an ordinal random variable. You do this by going to the variables on the left hand side of the data and left clicking on markers next to them.

(iii) Go to `analyze` click on `Fit Y by X`.

(iv) Put the observations into the `Y, response` box. For example, the speed of light column goes here.

Put the factors (the column which indicates the group that the observations belong to) into `X, factor`.

(v) Press `Okay`.

(vi) you will get a chart which plots the n groups side by side. From here you can see by eye whether the sample mean looks similar or not.

(vii) To get the Analysis of Variance table. Right click on `Oneway analysis...` and select `Means/Anova`.

- In the output you will get an ANOVA table. Please learn to inteprete this.
- One-way ANOVA is based on normality of the data. Though it is fairly robust to oultier and skew, if the sample sizes are quite small it makes sense to check for normality.

(i) To check for normality do not make a QQplot of the orginal data. You need to make a QQplot of the residuals.

(ii) JMP will give you the residuals as follows. Go to the title `Oneway analysis...` and click on `Save` and `Save residuals` (not Save standardised or the other options it gives).

- This will give you a new data column in your data spreadsheet.

(iv) Using the instruction for QQplots, you can make a QQplot of the residuals (a QQplot of this new data).

(iv) It is easy to show that that sample mean of residuals is zero.

Therefore never do an ANOVA on the residuals. It is pointless and you will never reject the null.

6.1 Linear regression

(i) Suppose we want to fit the model $Y = \beta_0 + \beta_1 x + \varepsilon$ to bivariate data. For example we want to fit Height $= \beta_0 + \beta_1$foot size $+\varepsilon_t$ to the data. Then upload the data into JMP.

(ii) Click on `analyze` and `Fit Y by X`.

(iii) Put into `Y, response` the response variable, for example height. Put into `X, factor` the regressor (explanatory variable), for example foot size.

(iv) Press `okay`. You should see a scatter plot of Y against X.

(v) To fit a line of best fit, right click in `Bivariate fit of....` and select `Fit line`. You will now see the line of best fit on the scatter plot together with a lot of output.

Note if you select `Fit mean` you will get a horizontal line which is the sample mean.

(vi) To make diagnostic plots to check model assumptions. Right click on the small red triangle just below the scatter plot and adjacent to a line with Linear Fit by the size. When you do this you should see the option for `Plot Residuals` click on this. You should see four plots.

(a) The first plot shows a plot of the residuals against the prediction equation $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$. No pattern here suchs that there isn't any 'dependence' between residual and mean and that the linear model is appropriate. A similar effect can be achieved by plotting the residuals against x_i.

(b) The second plot shows the predicted $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$ against the actual y_i. If the plots lie on the line, the prediction matches the observations.

(vii) Alternatively, you can directly analyse the residuals yourself. To do this you need to save the residuals. Right click on the small red triangle just below the scatter plot and adjacent to a line with Linear Fit by the size. When you do this you should see the option for `Save Residuals` click on this. A new column should appear in your output containing the residuals (note the sample mean of residuals in zero).

(1) To check whether the linear model is an appropriate model plot the residuals againist the regressors (explanatory variables).

To make this plot go to `Analyse` and `Fit Y by X`. Put the residuals into `Y, response` and regressor (such as shoes size) into `X, factor`. Then press `okay`, you should get a scatterplot.

(2) To check normality of the

7 Statcrunch commands

Statcrunch is best used with Firefox/Mozilla (especially if you are using a Mac). To upload data go to Data − > Load Data − > From File and choose select the location from where it comes from. Remember you will need to tell Statcrunch whether the file is separated by spaces or commas etc (using the Delimiter option in the pop-up window).

7.1 Saving plot and tables

- To save a plot is straightforward. Click on Options − > Save − > to my computer. Then save it in the desired folder. The image will be saved as a `.png` file.
- To save a table (such as output) is a little harder.

  Either:

  - Save as a screenshot (see directions for JMP screenshot saves).
  - Or click on Options − > Print − > Click on PDF and choose `Save as PDF`. This will save it as a PDF file. From here you may want to crop to size (in linux use the `pdfcrop` command).

  Both the above methods will save Statcrunch tables.

7.2 Calculating probabilities

To calculate probabilities (eg. Binomial, Normal, t):

- Go to Stat $->$ Calculators $->$ and select the distribution.

  In class we will only be using Binomial, normal, t, chi and F.

- The plots will help you see what areas have been calculated.

7.3 Categorical Data

I found that the analysis of categorical data was easiest in Statcrunch, especially if you only have the summaries.

7.4 Proportions

Obtaining Binomial probabilities:

- Go to Stat $->$ Calculators and select Binomial.
- Place in n the sample size and p the probability (for example the probability under the null in the one-sample case).
- In Prob (X <= or < *etc* BLANK) (you place the number out of 5 you want to calculate - for example if you want to calculate the probability that getting 3 heads out of 5 on a coin, you type n=5, choose = and BLANK = 3). The probability will be calculated and given after the equal sign.

One Proportions:

- Example Ernie gets 600 votes out of 1000 and you want to test $H_0 : p \leq p$ against $H_A : p > 0.5$.
- Go to Stat $->$ Proportions $->$ One-sample $->$ with summary.
- Place Number of successes = 600 and Number of observations = 1000.
- Click Next, and either choose the test (and the p, in this case 0.5) or the confidence interval (with the level).
- Finally press calculate and you should get the appropriate table

Two Proportions:

- Go to Stat $->$ Proportions $->$ Two-sample $->$ with summary.
- Continue the same as before but inputting the information from two samples.

Feetsize	Hairy	Not Hairy	Subtotals
Big feet	70	10	80
Not big feet	70	100	170
Subtotals	140	110	Total=250

7.5 Continency Tables and Chi-squared Test for independence

- Example:
- Select Stat − > Tables − > Contingency − > With Summary.
- In Select Columns for table select *Hair* and *Not Hairy.*
- In Row labels in Select *Feetsize.*
- Press Next and select the desired options (including the chi-square (test for independence)).