Statistics

Peter Lamb

Week 5

In this chapter we will cover a few basic statistical analysis techniques. This section does not cover the theory behind the statistical method/design. A well thought out hypothesis and appropriate statistical analysis design is a very important step in the process. You should not use these tutorials as a basis for deciding on a statistical method, just on how to perform the analysis once you have decided on what to use.

Descriptive statistics

In this section we will cover a few of the simplest statistical functions. Import the data from the 'stats_example.dat' file in the **5-Statistics/data** folder as a matrix. Start by renaming the example data to save a bit of typing:

```
M = statsexample;
```

Min, Max and Mean

Calculate the minimum value in both columns of the **M** matrix.

```
>> min(M)

ans =

   -1.2839   -1.6690
```

Calculate the maximum value in *all* elements of the **M** matrix.

```
>> max(M(:))

ans =

   5.5282
```

Calculate the mean value in both columns of the **M** matrix.

```
>> mean(M)

ans =

   2.4155    0.4820
```

Range

Using the data in the **M** matrix from the previous examples calculate the range of the values in the two data columns.

```
>> range(M)

ans =

   6.8121    4.3462
```

Standard deviation

Calculate the standard deviation of the values in the two columns.

```
>> std(M)

ans =

   1.4712    1.0279
```

Variance

Calculate the variance of the values in the two columns.

```
>> var(M)

ans =

   2.1645    1.0566
```

Percentiles

Calculate the 25^{th}, 50^{th} and 75^{th} percentiles for column 1.

```
>> prctile(M(:,1),25)

ans =

   1.1803

>> prctile(M(:,1),50)

ans =

   2.5274

>> prctile(M(:,1),75)

ans =

   3.4970
```

Correlation coefficient

Calculate the correlation coefficient for the two data columns.

```
>> corrcoef(M(:,1), M(:,2))

ans =
```

```
    1.0000    0.7024
    0.7024    1.0000
```

Here, you can see four coefficients, which might seem strange. The coefficient in position (1,1) *(row 1, column 1)* is the correlation coefficient for `M(:,1)` with `M(:,1)`. Since they are identical, there is a perfect correlation. The coefficient in (2,1) is for `M(:,2)` and `M(:,1)`, (1,2) is for `M(:,1)` and `M(:,2)`, and (2,2) is for `M(:,2)` and `M(:,2)`. Long story short, you are interested in the non-trivial values in (2,1) and (1,2), your correlation coefficient is 0.7024.

Note that you can specify more output arguments to get the correlation coefficients (`R`), p-value (`p`), and the lower (`RL`) and upper (`RU`) bounds for a 95% confidence interval for the correlation coefficient. We'll come back to this later when reporting our statistics in a paper or thesis.

```
>> [R, p, RL, RU] = corrcoef(M(:,1), M(:,2));
```

Linear Models

Load the data in the file **'age_and_hr.mat'**. The variable called **D** contains two columns of participant information. Column 1 has the participants' age and column 2 has their maximum recorded heart rate during a VO_2 max test. Use the **regress** function in MATLAB to investigate the assumption that max HR is approximately equal to 220 beats/min minus your age in years.

Plot the two columns of data against each other, with age along the x-axis and maximum heart rate on the y-axis. Use green dots to mark the coordinates and add axis labels.

```
>> plot(D(:,1), D(:,2),'g*');
>> xlabel('Age (years)');
>> ylabel('Max recorded HR (beats/min)');
```

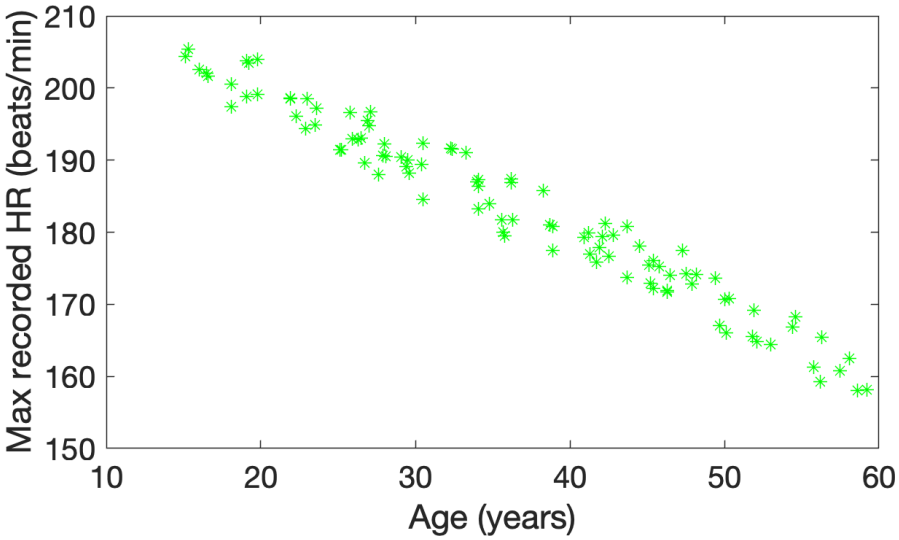

Figure 1: A scatter plot of max HR against age

Consult the help documentation for the **regress** function for information on how to use the function. We will make use of a constant term so need to add a column of ones to our age data (X) (see the note at the bottom of the description section in the documentation). We can use the **ones** function to create a column, with as many rows as the matrix: **D**.

Create the X values from the ages and then use the **regress** function to create the regression model and extract the statistical data.

```
>> age = [ones(size(D(:,1))) D(:,1)];
>> hr = D(:,2);
>> [m,~,~,~,stats] = regress(hr, age);
```

Side note: notice the use of **size** as input to **ones**; this is a common and important tactic when scripting an analysis. Look at the output of the **size** function, as used above:

```
size(D(:,1)) % size of column 1

ans =

   100     1
```

We could have entered **100** and **1** as input to **ones** to get a 100 x 1 column vector of ones; however, using the size of a variable in your workspace allows flexibility.

For example, imagine you wanted to add the data for one more participant, making your column vector grow to 101 x 1. Had you *hard-coded* the **100** and **1** as input, you would have to change every line of code that relied on there being 100 rows, using **size** requires no changes: it would now output 101 and 1. Back to the module...

The variable **m** has two elements. Since this is a linear regression these values correspond to the constants, m_0 (y-intercept) and m_1 (slope), in the standard equation of a line:

$$Y = m_0 + m_1 X$$

where

$$m_0 = 219.744$$

and

$$m_1 = -0.9997$$

It is clear that our result is very close to the estimation that:

$$Y = 220 + (-0.9997)X,$$

where Y is the max HR and X is age.

The **stats** variable has the R^2 and F statistic, the p-value for the full model, and an estimate of the error variance. Add the regression line to the plot for ages 0 to 70 years and then add the model and the R^2 value as a text label.

```
% 'hold' axes to add plots without creating new figure
>> hold on
>> plot((0:70), (m(1) + m(2)*(0:70)));
>> text(20, 210, 'y = 219.744 - 0.9997x (R^2 = 0.96)');
```

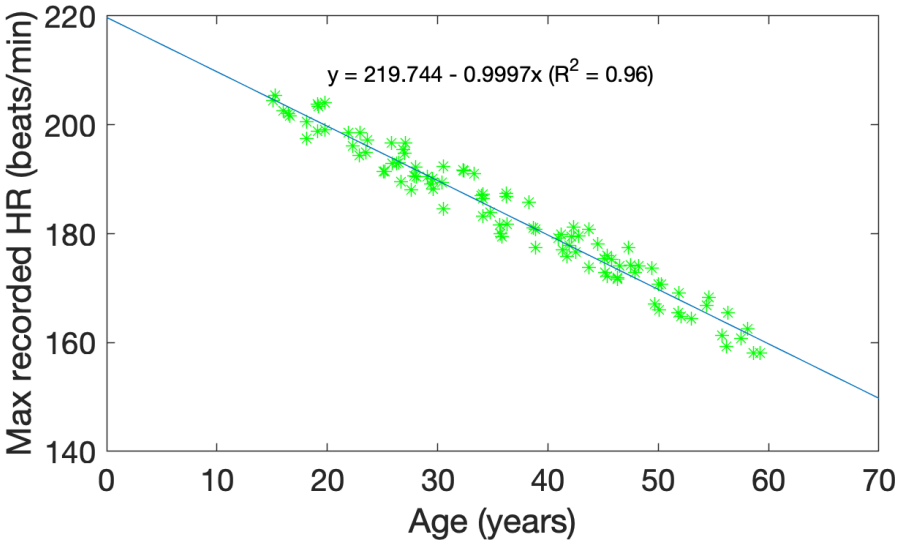

Figure 2: A linear regression of HR against Age

Again, whenever you see a new function used, get in the habit of looking up its help documentation, e.g. `>> doc text`. The **text** function allows us to add text to the plotting area.

t-tests

Import the data from the file **'ttest.csv'** into the workspace. Begin reading at line 2 to skip the header. This file contains pre and post-intervention data across various measures for 32 participants. The results are from a fictitious long-term fitness intervention study.

One-sample and paired-sample *t*-test

MATLAB's **ttest** function can be used to perform a one-sample or paired-sample *t*-test. We will begin by calculating the change in weight throughout the study (pre weight is in column 3 and post weight in column 4). This will give us a change score (one sample) for each participant. We can take the mean value of this result to see the average amount and direction of weight change.

```
>> tD = csvread('data/ttest.csv', 1)
>> chgWeight = tD(:,4) - tD(:,3);
```

```
>> mean(chgWeight)

ans =

   -4.9125
```

We can now use these change scores to perform a t-test of the null hypothesis that the changes are a random sample from a normal distribution with a given mean, **m**, and unknown variance, against the alternative that the mean is not **m**. We will return to the hypothesis outcome, **h** (where $h = 1$ indicates a rejection of the null hypothesis and $h = 0$ indicates a failure to reject the null hypothesis) and the p-value of the test. We will use the default values of mean, **m**, as 0 (no weight change), an alpha value of 0.05 and a two-tailed test.

```
>> m = 0;
>> [h, p] = ttest(chgWeight, m);
```

The result shows that the null-hypothesis was rejected, $h = 1$, since the p value ($p = 0.0377$) is less than the 0.05 threshold[1]. Confirm that reducing the significance level to an alpha value of 0.01, 1%, would result in a failure to reject the null hypothesis.

```
>> h = ttest(chgWeight, m, 0.01)

h =

   0
```

In the previous example we used single samples that were created as a change score pre and post intervention. It then follows that the null hypothesis would be that the mean was 0. We could have taken the raw values and used a paired-sample t-test to perform a similar statistical analysis. Verify that the result would, as expected, produce an identical outcome:

```
>> [h, p] = ttest(tD(:,4), tD(:,3));
```

We will now use the gender information, in column 2, to investigate the effect the intervention had on pre/post BMI scores within each sex. Extract the indices of males and females and perform a paired-sample t-test on each group (in column 2, 1 indicates male and 2 indicates female).

[1] add the outputs `ci` and `stats`, as in `[h,p,ci,stats] = ttest(chgWeight,m);` to output degrees of freedom, t-statistic and 95% confidence interval for reporting.

```
>> males = tD(:,2) == 1;
>> females = tD(:,2) == 2;
>> [hF, pF] = ttest(tD(females,6), tD(females,5));
>> [hM, pM] = ttest(tD(males,6), tD(males,5));
```

Verify that effect is significant in one gender group but not in the other.

Two-sample *t*-test

We will now perform an unpaired two-sample, or *'independent samples' t*-test on the same body weight data as before.

```
>> [h2, p2] = ttest2(tD(:,4),tD(:,3));
```

Compare this result to the paired-sample results from before (**h** and **p**), notice the difference. In this second case we are using the same set of data, but we are not treating the observations (rows) as paired, just as two separate samples. This type of test does not have the same level of power as a paired sample test and this shows in the change of the result. This highlights the importance of choosing the correct statistical model as you can get a very different result, even when using exactly the same data!

Analysis of Variance

One-way analysis of variance

Using the same data set as before we will look at running an analysis of variance, or ANOVA, on the weight change scores by gender. Next we will use MATLAB's **anova1** function to perform a one-way analysis of variance.

```
>> [pAOV,tbl,statsAOV] = anova1(chgWeight, tD(:,2));
```

The following figures should be displayed to show the results of the analysis.

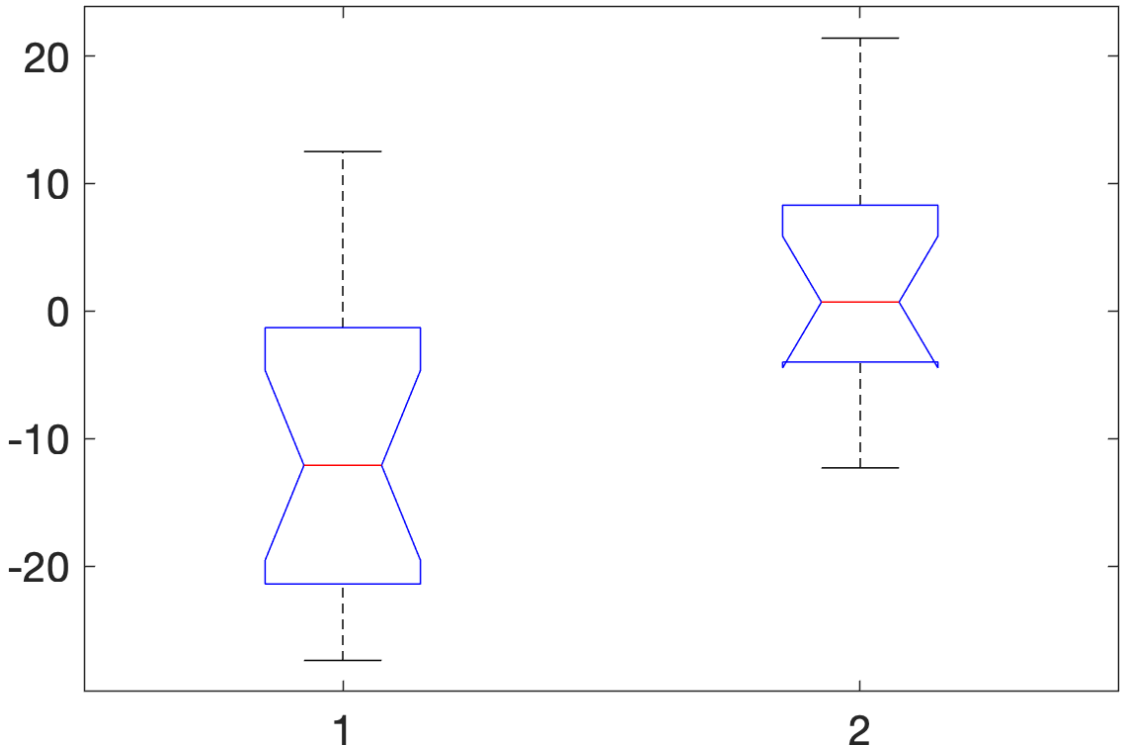

Figure 3: A box plot showing the change in weight for males and females

Source	SS	df	MS	F	Prob>F
Groups	1028.0	1	1028.0	7.6142	0.0098
Error	4050.3	30	135.0		
Total	5078.3	31			

Table 1: The table output from the ANOVA function (see `tbl` variable in the Workspace)

From the table and from the output variable, **p**, we can see that we should reject the null hypothesis that the two groups are drawn from a population with the same mean. This result indicates that there is a significant difference ($p = 0.0098$) in the average weight change between men and women during the intervention. In other words, if the samples were drawn from the same population we would only expect differences of this magnitude about 1% of the time in repeated

experiments. Additionally, the box plot clearly shows how these changes in men and women differ.

The **multcompare** function can be used to do *post hoc* tests following ANOVA. In this case we only have two groups, so it's obvious the the difference is between these two groups; however, we'll use the function to output the 95% confidence interval.

```
>> c = multcompare(statsAOV)

c =

    1.0000    2.0000  -19.8815  -11.4254   -2.9693
       0.0098
```

According to the documentation, the first two columns indicate the levels of the grouping variable (1 (males) was compared to 2 (females), if there was another level we would have two more rows, 1 vs 3 and 2 vs 3). The third column is the lower bound of the 95% confidence interval, the fourth is the difference in group means, the fifth is the upper bound of the 95% confidence interval and the last column is the p-value for this comparison.

In summary we can infer the following from the ANOVA and previous t-tests:

- There was a significant loss in body weight in the overall sample group, $t(31) = -2.17$, $p = .038$, 95% CI $[-9.53, -0.30]$.
- However, the ANOVA indicated that there was a significant difference in the nature of the change between the two gender groups $F(1, 30) = 7.61$, $p = .010$, 95% CI [-19.89, -2.97].
- The t-tests within the gender groups showed that there was a statistically significant loss in body weight in the male participants $t(17) = -2.22$, $p = 0.040$, 95% CI $[-11.44, -0.30]$, and only a slight non-significant increase in body weight for females $t(13) = -0.72$, $p = .483$, 95% CI $[-8.13, 4.06]$.

Take note in the above summary how statistics should be reported in APA format. In the first bullet the t indicates the type of test, (31) indicates the degrees of freedom followed by the t test statistic, the p-value is next (note the preference for giving the exact value rather than $p < .05$), followed by the 95% confidence interval. Some prefer the effect size to be reported in place of the 95% confidence interval, but we'll stay out of that debate for now.