Practical: Statistics with R

Introduction

1. R and RStudio.
2. R refresher: data types, iteration, plotting (plot, boxplot, heatmap, ...), documentation, loading packages, IO (`read.csv`, `load`), comments.
3. Writing R markdown vignettes. (At the end of your vignette, include a code chunk with a call to `sessionInfo()` to record the version of R and all loaded and attached packages.

Create a data set

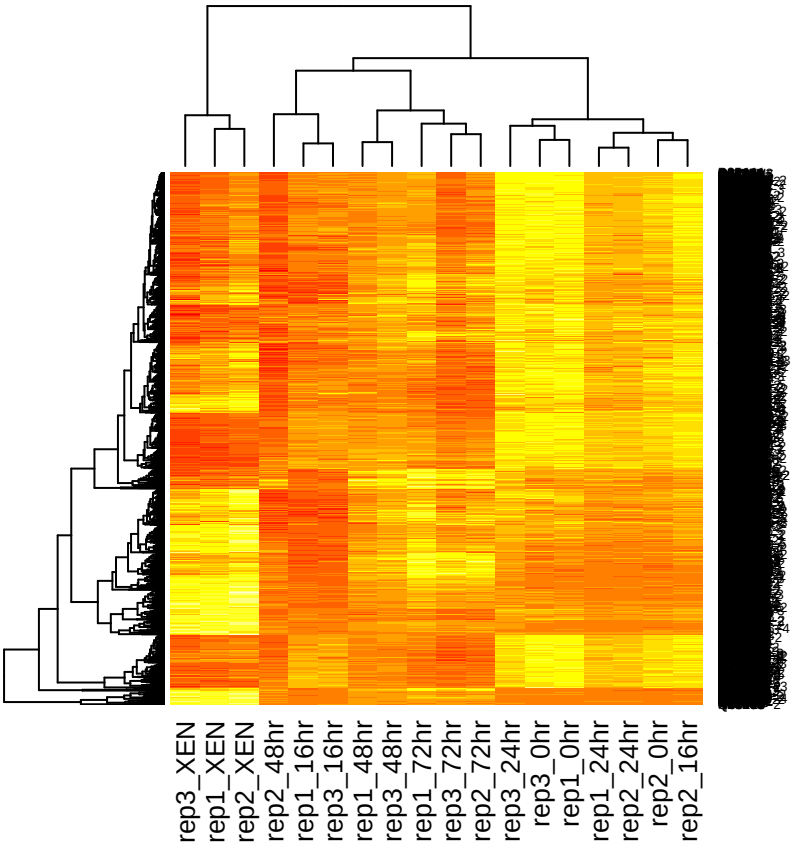
Create and visualise a dataset that matches the following experimental designs:

> We used a SILAC quantitative proteomics approach to measure the effect of drug A on human embryonic kidney HEK-293 cells in triplicates. We used a one sample t-test to analyse the fold-changes.

Assume that 1000 proteins where measured over all replicates.

Explore this design

The data illustrated in the heatmap below is available in the `timecourse.csv` file. In this experiment, Mulvey and colleagues, performed a time course experiment on mouse extra-embryonic endoderm (XEN) stem cells. Extra-embryonic endoderm differentiation can be modelled in vitro by induced expression of GATA transcription factors in mouse embryonic stem cells. They used this GATA-inducible system to quantitatively monitor the dynamics of global proteomic changes during the early stages of this differentiation event (at 0, 16, 24, 48 and 72 hours) and also investigate the fully differentiated phenotype, as represented by embryo-derived XEN cells.



Describe the experimental design in terms of units, replicates, experimental factors, controls. Formulate the hypothesis of the study in your own words. The experiments were performed using tandem mass tags (TMT), which allow to combine up to 6 experimental units per mass-spectrometry acquisition. Suggest and motivate a design.

Read the file into R. How many experimental units and features (proteins) are there?

Use the `heatmap` function to produce the heatmap shown above. Describe how and why this matches or contradicts the experimental design.

Calculate and compare the variability within and between two experimental groups of your choice.

Identify relevant groups

The data set in `dataonly.csv` is comprised of 2 groups, A and B, each with equal numbers of technical and biological replicates (balanced design), but the person who run the samples lost the sheet matching samples to groups. They can only remember that sample A belonged to group A.

Infer the likely groups from the data. Report the different types of variability's to motivate your answer.

Power calculation

The data set provided in the `cll.rda` file is a subset of the chronic lymphocytic leukemia (CLL) gene expression data from the CLL package. It contains 10 (out of 24) sample samples that were either classified as progressive or stable in regards to disease progression.

You are asked to perform a power analysis on this data. Estimate values for the effect size and standard deviation that are relevant for this kind of data.

Load the data and verify that you have 12625 rows and 10 columns.

To identify appropriate effect size and standard deviation for the power analysis calculate generate two figures.

First, calculate the standard deviations and means for all the genes and visualise them a scatter plot. Optional, to help you choose these values, calculate a regression of the means and standard deviations.

Then produce a so-called MA-plot (these plots were initially used to visualise micro-array data) by plotting the mean intensities against the log_2 fold-changes.

Estimate the number of samples required to obtain a power of 0.8 at a significance level of 0.01 for low, medium and high expression genes.

False discovery rate for single tests

Calculate the FDR for the following parameters. Optional: write an R function that takes these parameters as input and returns the FDR.

P(real)	Power	Singificance level
0.1	0.3	0.001
0.5	0.3	0.001
0.1	0.6	0.001
0.5	0.6	0.001
0.1	0.3	0.500
0.5	0.3	0.500
0.1	0.6	0.500
0.5	0.6	0.500

Hypothesis testing

It we toss a fair coin n times, we expect to observe heads roughly $\frac{n}{2}$ times. Such events, where one wants to quantify the number of p successes in n trials is modelled by the binomial distribution.

Quantify the probability to observe exactly 8 heads out of 12 trails using the binomial mass function $\binom{n}{k}p^k(1-p)^{n-k}$ where n is the size of the trial, k is the number of successes, p is the probability of a success and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. In R, $n!$ is calculated with `factorial(n)`. Alternatively, use `choose(n, k)` to calculate $\binom{n}{k}$.

In R, every distribution has a set of density (`d*`), distribution (`p*`), quantile (`q*`) and random generator (`r*`) functions. For the normal distribution, `*` is `norm`. For the binomial distribution, `*` is `binom`. See `?binom` for details.

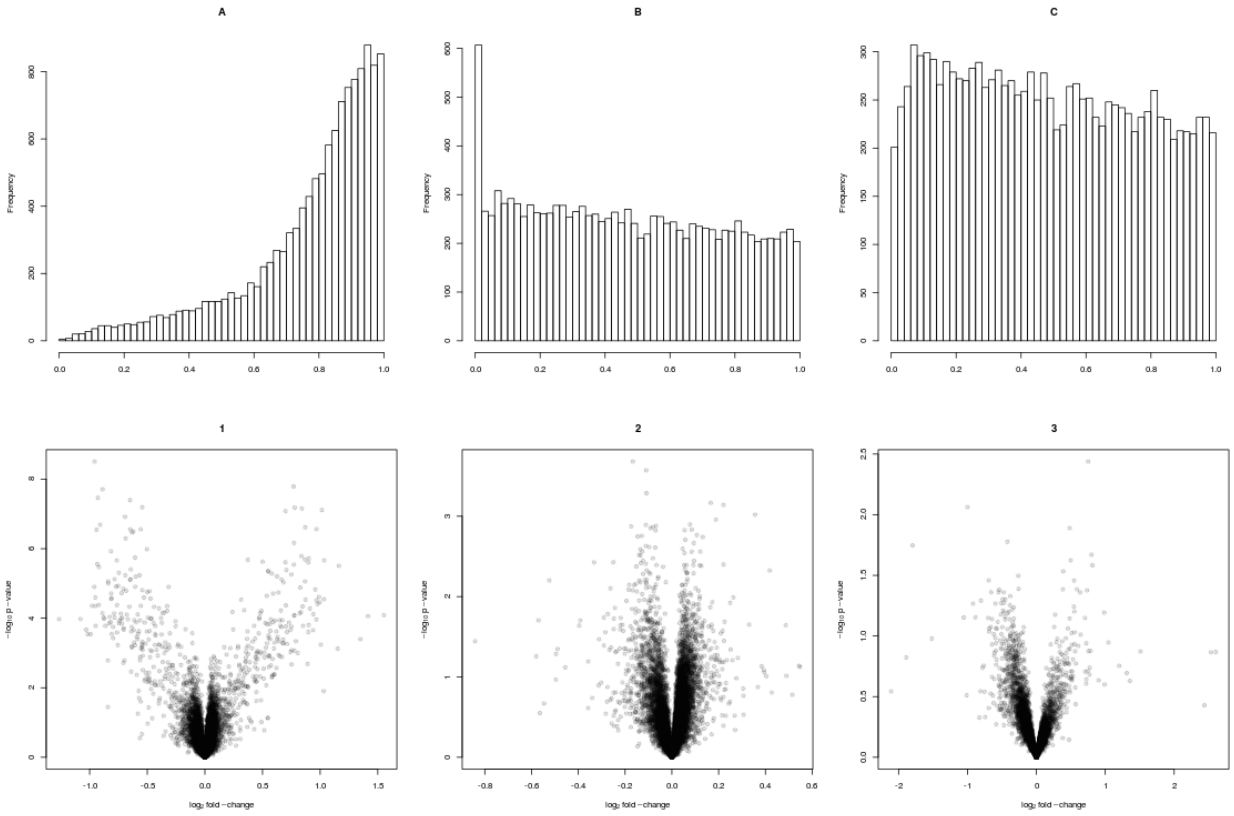
Generate a histogram that shows the respective number of heads for 12 trails.

Use the density function to verify your calculation above.

What is the probability to observe 10 or more heads $P(heads \geq 10)$.

P-value interpretation

Match the following p-value histograms and volcano plots.



A complete example

T-test

Perform a complete statistical analysis on the XEN and 48 hour groups of the time course experiment.

Start by reading the data into R and subset the relevant groups.

Calculate p-values for each gene using a Welch t-test (two sample t-test with unequal variances), plot a p-value histogram and consider if it is worth continuing the analysis.

Calculate log_2 fold-changes for all the genes and visualise then with a boxplot.

P-value adjustment

The Bonferroni family-wise error rate (FWER) correction works as follows: multiply each p-value by the total number of tests m. Tests with adjusted p-values smaller than your chosen level of significant are deemed significant.

The Benjamini-Hochberg false discovery rate (FDR) method enables to control the number of false discoveries due to multiple comparisons by applying the following procedure: multiply each p-value by its order and divide by the number of tests m. The adjustment of the smallest p-value (order 1) correspond to the Bonferroni correction; the second one is adjusted as $\frac{2 \times pv}{n}$, ... and the largest p-value remains unchanged.

> Perform the p-values adjustment for multiple comparisons applying the Benjamini-Hochberg and Bonferroni methods as implemented in the `multtest` Bioconductor package. The relevant function is `mt.rawp2adjp`.

Interpretation of the tests

> Draw and interpret the volcano plot using the Benjamini-Hochberg adjusted p-values.