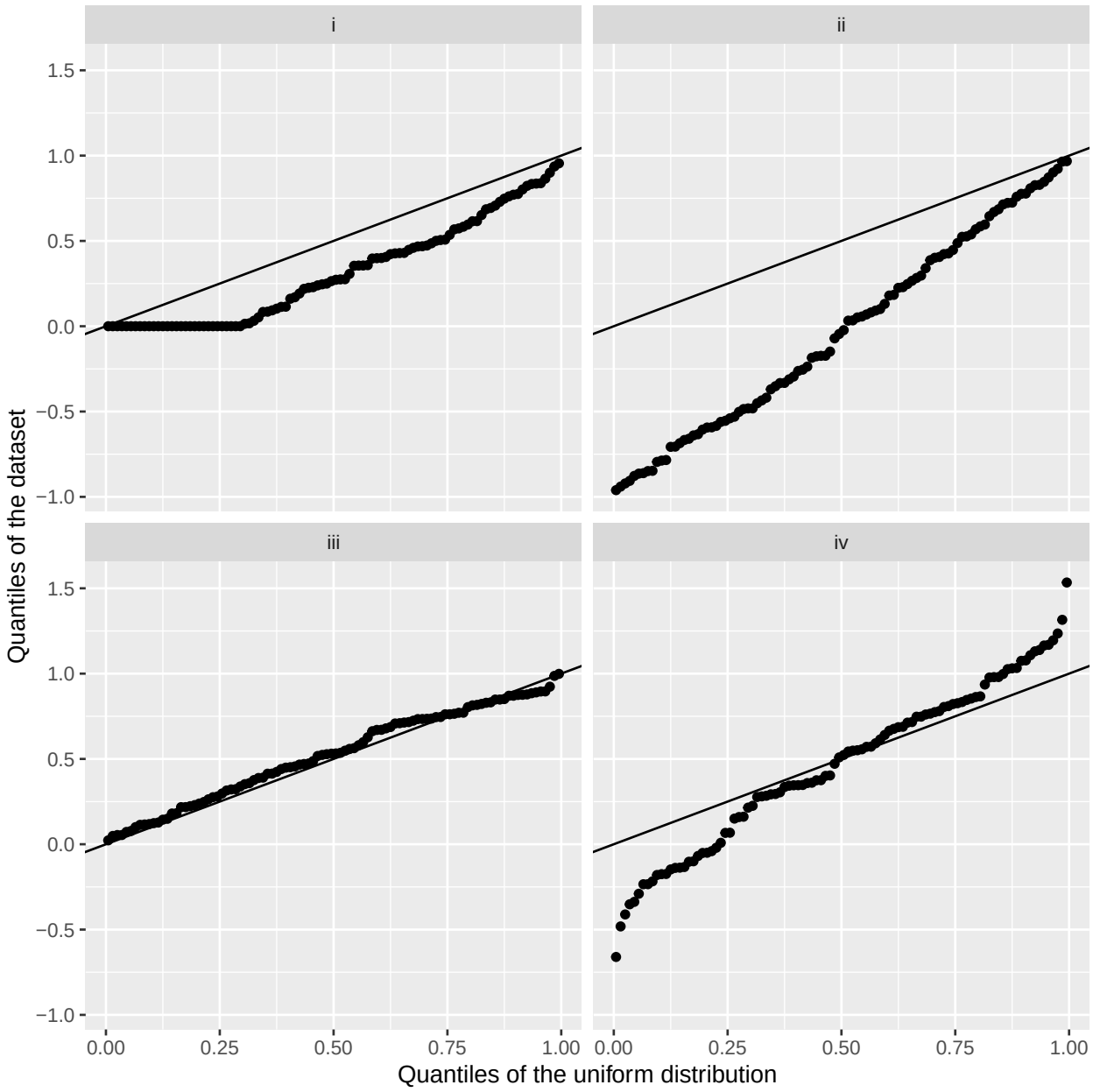
Data Analysis and Visualization in R (IN2339)

Exercise Session 9 - Statistical Assessments for Big Data

Jun Cheng, Christian Mertes, Vicente Yépez, Julien Gagneur

Quiz

1. Look at the following QQ-Plots. In each case the expected quantiles are taken from a uniform on the interval $[0, 1]$, whereas the observed quantiles were generated using the distributions specified in sentences (a)-(d). Match each plot with a sentence (each sentence matches exactly one plot):



a) The data follows a uniform distribution on the interval $[0, 1]$

b) The data follows a uniform distribution on the interval $[-1, 1]$

c) The data follows a normal distribution with $\mu = 0.5$ and $\sigma = 0.5$

d) The data follows a uniform distribution on the interval $[0, 1]$ with outliers at 0

2. Say we perform a Spearman correlation test for two variables. The test returns $P < 10^{-26}$ (i.e. values as or more extreme than the ones observed are astronomically unlikely under the null hypothesis). Indicate which, if any, of the following statements is **correct**:

a) The two variables necessarily show a strong positive correlation (i.e. $\rho > 0.5$)

b) The two variables necessarily show a strong negative correlation (i.e. $\rho < 0.5$)

c) The two variables necessarily are strongly correlated (i.e. $|\rho| > 0.5$)

d) Surely something very important has been discovered

3. Indicate which, if any, of the following statements is **correct**:

a) A simple yet effective strategy to control for multiple testing is to only reject the null hypothesis when $P = 0$, i.e. when we are *sure* of the association

b) Assume the null hypothesis is always *true*. If we do 200 tests, we will expect to have around 10 false positives when using $\alpha = 0.05$ as our threshold of significance

c) Assume the null hypothesis is always *false*. If we do 800 tests, we will expect to have around 80 false positives when using $\alpha = 0.1$ as our threshold of significance

d) Assume the null hypothesis is sometimes *false*. Then the P-values will follow a normal distribution

4. Assume we are doing 1000 tests. Indicate which, if any, of the following statements concerning Bonferroni and Benjamini-Hochberg are **correct**:

a) Assume we are using a permutation-based approach. If we use a Bonferroni correction, we will in general need to do more permutations to be able to reject the null than if we used a Benjamini-Hochberg correction.

b) If we let $\alpha = 0.01$ and use a Bonferroni correction, then the probability of one or more false positives (falsely rejecting the null) will be less than 1%

c) If we let $\alpha = 0.01$ and use a Benjaminin-Hochberg correction, then in expectation 1% of the tests we perform will reject the null

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork)
```

2. Load yeast data and required packages using the following code:

```
genotype <-fread("extdata/eqtl/genotype.txt")
growth_rate <-fread("extdata/eqtl/growth.txt")
marker <-fread("extdata/eqtl/marker.txt")

setnames(marker, "id", "marker")
genotype <- genotype %>%
  melt(id.vars = "strain", variable.name = "marker", value.name = "genotype")
```

Section 01 - Quantile-Quantile plots

We will simulate some data from different distributions and will compare their quantile-quantile plots.

1. We will use a standard normal ($\mu = 0$ and $\sigma^2 = 1$) distribution as a reference set. Please simulate 100 draws from a standard normal distribution. Add them as a column to a data table and plot a histogram of these values. Next, use `ggplot` to create a QQ-plot comparing the expected against the observed quantiles.

Use `geom_abline` to draw a line on the QQ-plot where the data should be if the observed and expected quantiles match exactly. (Do not mind warnings, if any).

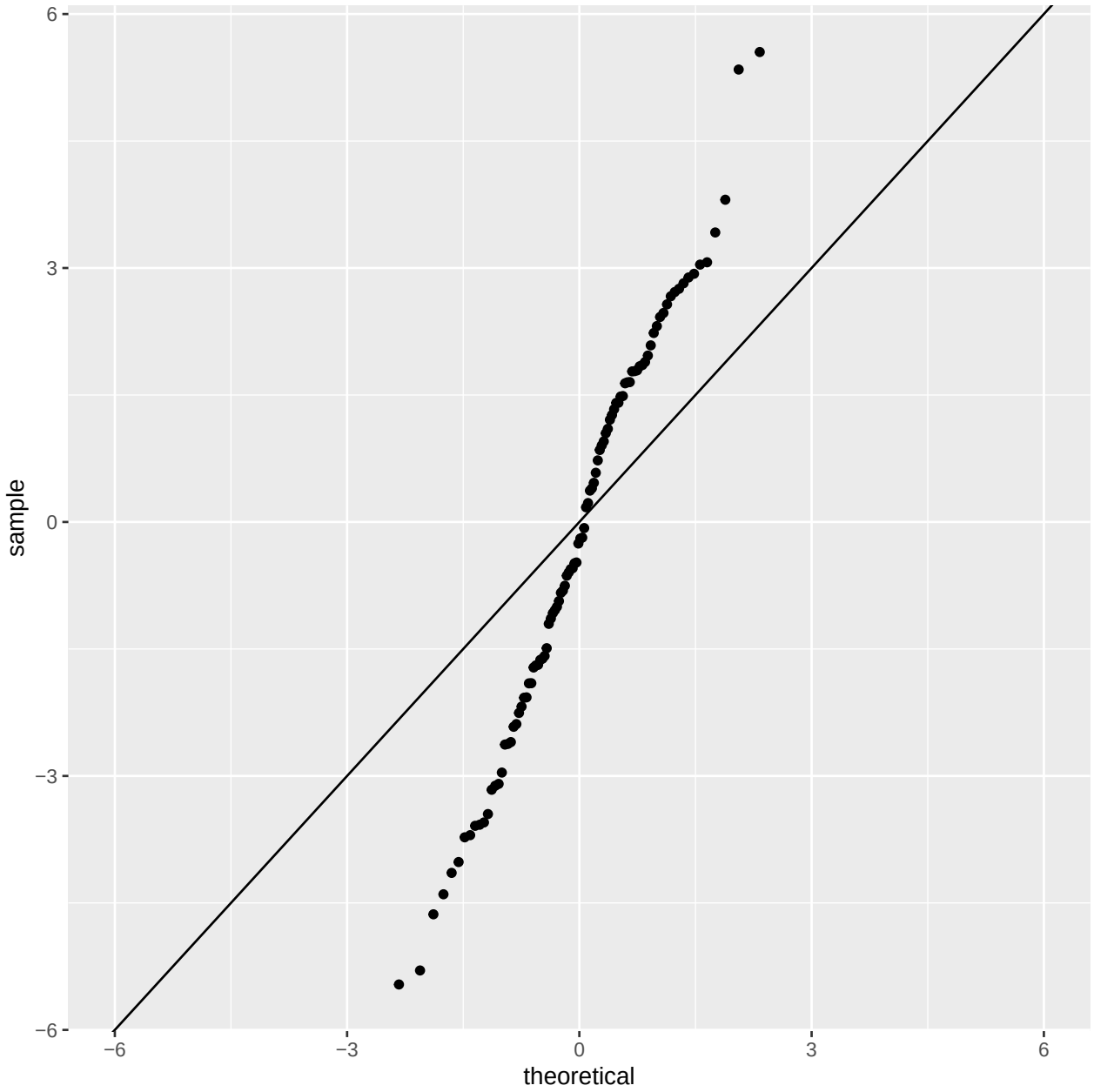
Hint 1: Set the x and y limits to `[-6,6]` where appropriate.

Hint 2: A small reminder of the R functions to simulate a normal distribution:

```
set.seed(10)
# rnorm allows you to draw random values from a normal distribution
rnorm(10) # 10 random draws
# pnorm gives the cumulative probability from a normal
# i.e. pnorm(x) = p(X < x) where X is the random variable
pnorm(0)
# qnorm returns the quantiles of a normal distribution
# it is the inverse of pnorm
# i.e. given a probability p,
# it finds x so that pnorm(x) = p
qnorm(0.5)
# qnorm can be used to find different types of quantiles
qnorm(seq(0.25,0.75,0.25)) # quartiles of the normal
qnorm(seq(0.1,0.9,0.1)) # deciles of the normal
```

2. Now add a normal distribution with $\mu = 4$ to your `data.table` and plot the Q-Q plot. How did it change?

3. How would you tweak the distribution so that you get the following Q-Q plot?



Section 02 - QTL mapping of growth

For the next questions, we will use the yeast dataset.

1. Test for markers associated with growth.

Report genetic markers significantly associating with growth rate in maltose. We would like the expected ratio of false discoveries among the reported significant markers to not exceed 5%.

To do so, run a Wilcoxon test for growth rate versus the genotype at each of the 1,000 markers. Remember that we tested this relationship for one specific marker last time. Plot a histogram and a Q-Q plot of the obtained P-values. Which ones would you consider significant and why? Do we need to correct for multiple testing?

Hint: plot P-values in `-log10` scale. Use `-log10(ppoints(pval))` to generate the expected quantiles. Note that this is similar to using `geom_qq` and then log-scaling the axes, except that we want to log-scale and also

reverse.

2. Plot the P-values against genomic position. Do you see positions that are associated with growth? The genomic position is defined by the chromosome the marker is on and the marker's position within that chromosome.

Hint: plot P-values in `-log10` scale. Use the `start` column from the marker table as position. Additionally, facet on the `chrom` column (because the `start` column indicates the genomic position of the marker within its particular chromosome).

3. How many markers significantly associate with growth before and after correcting for multiple testing?

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 03 - Let's do many tests

1. Consider the dataset `mtcars`. Find all pairs of quantitative variables that show significant association (at the 5% level) with each other. Do not make any assumptions of distribution.

Hint: for a solution that does not use for-loops, use `combn` and `apply`.

2. Now ensure that, on average, less than 5% of the significant associations you find are false positives.

3. Now ensure that the probability of having 1 or more false positives is less than 5%

Section 04 - P-values and FDR

Here we will use simulations to investigate the effect of sample size and of the proportion of true and false null hypotheses when performing multiple testing. We will do it for the problem of two-sample comparison with equal sizes.

We are interested in comparing the observations of two samples: $x_1, ..., x_n$ and $y_1, ..., y_n$. Specifically, we ask whether the expectations differ using a two-sample Student t-test.

1. Simulate data under the null hypothesis $H_0 : \mu_x = \mu_y = 0$.

We simulate $N = 10,000$ times two samples $x_1, ..., x_n$ and $y_1, ..., y_n$ where X and Y follow the standard normal distribution. We use sample size $n = 50$ for each group. For each simulated dataset, we compute the two-sided P-value of a t-test. We assume unequal variance as by default in the R function `t.test()`.

You can use the following functions for this exercise:

```
simulate_norm_groups <- function(sample_size=50, N_experiments=10000, mu_x=0, mu_y=0){
  sapply(seq(N_experiments), function(i){
    x <- rnorm(sample_size, mu_x)
    y <- rnorm(sample_size, mu_y)
    t.test(x, y, alternative="two.sided")$p.value
  })
}

plot_pval <- function(pvals, title="p-val distribution"){

  pval_dt <- data.table(pvals=pvals)
  histo <- ggplot(pval_dt, aes(pvals)) + geom_histogram(boundary = TRUE) +
    labs(title = title)

  qq <- ggplot(data = pval_dt, aes(sample = pvals)) +
```

```
    geom_qq(distribution =  stats::qunif, dparams = list(min = 0, max = 1)) +
    geom_abline(a=0,b=1) +
    ylim(c(0,1)) +
    labs(title = title)

  histo + qq

}
```

2. compute the quantiles and plot a histogram and a QQ-plot

If all tests are truly under the null hypothesis, the distribution of the P-values should be uniform by definition. Please plot the P-values for $sample\_size = 50$ with the provided function. Discuss.

3. Correct for multiple testing

Adjust P-values with the different methods seen in the class. Plot the results using the plot function. Do they behave as expected? Discuss.

Section 05 - sample size and power

1. We will now simulate data under the alternative hypothesis $H_1 : \mu_x \neq \mu_y$. Specifically, we simulate two samples $x_1, ..., x_n$ and $y_1, ..., y_n$ where X and Y follow the normal distribution with $\mu_x = 0$ and $\mu_y = 0.5$ respectively. Do $N = 1,000$ experiments and investigate the effect of different sample sizes n (10, 100, 1000) on the P-value plots. Discuss.

2. P-values for a mixture of null and alternative.

Provide the same plots as before when considering a dataset of $N_0 = 10000$ data points simulated under H_0 (true null) and $N_1 = 1000$ data points simulated under H_1 (false null). Discuss. You can also use a `-log10` transformation, as in the tutorial, to better visualize the lower end of P-values. The following function will allow us to plot the `-log10` transformed P-values in a QQ-plot:

```
plot_pval_log10 <- function(pvals, title="p val distribution"){
  n  <- length(pvals)

  dt <- data.table(
    observed = -log10(sort(pvals)),
    expected = -log10(ppoints(n))
  )
  ggplot(dt) +
    geom_point(aes(expected, observed)) +
    geom_abline(intercept = 0, slope = 1)

}
```

3. Mixture of H_0 and H_1 adjusted for multiple testing

Adjust the p-values with Benjamini-Hochberg (FDR) in the mixture from the previous question. Make a contingency table of true positives, true negatives, false positives and false negatives. Try this with different sample sizes for FDR = 0.05. Discuss.

Do the same thing for the bonferroni correction and compare the results

Hint: You can use the following function for this analysis:

```
error_analysis <- function(method='BH', sample_size=50, cut=0.05){
  pvals <- c(
    simulate_norm_groups(sample_size = sample_size, N_experiments = 10000),
```

```
    simulate_norm_groups(sample_size = sample_size, N_experiments = 1000, mu_y=0.5))

  names(pvals) <- rep(c("H0", "H1"), c(10000, 1000))

  pvals_adj <- p.adjust(pvals, method=method)
  table(ifelse(pvals_adj < cut, "significant", "not significant"), names(pvals))
}
```