HDS 1. Multiple testing problem

Matti Pirinen, University of Helsinki

29.10.2019

Suppose that we have a large number p of variables of which some could be explaining outcome y. How do we know which ones are important? What are the statistical measures we can use?

Example 1.1

1. **Genome-wide association study.** We have n individuals measured on $p \sim 10^6$ positions on the genome where each x_{ij} has value of 0,1 or 2 denoting how many copies of the reference DNA letter the individual i carries at position j of the genome. In addition, each individual has been measured for cholesterol levels (outcome y). Which genomic positions affect cholesterol levels? We can do linear regression of y on each predictor x_j separately which leads to the regression summary statistics $(\widehat{\beta}_j, \mathrm{SE}_j, P_j)$ for each x_j. The task is to infer which of the 10^6 predictors are truly altering cholesterol levels.
2. **Tumour - normal comparison.** We compare the levels of p (in thousands) proteins between the tumour sample and a control sample from a healthy tissue of the same patient. When we have hundreds of patients we can compute t-statistics from a paired t-tests for each protein to see whether it has different levels in tumour than in healthy tissue across the patients. Again we end up with p estimates for differences $(\widehat{\beta})$, their SEs and P-values, one for each of the p proteins. Which proteins are statistically clearly differentiated between tumour and normal samples?
3. **Brain images.** Imaging data are very high dimensional. For example, we could define thousands of regions from the brain that could be compared between certain groups of individuals, (e.g. groups stratified by age, sex or a psychiatric condition). How do we determine which regions show differential activity between the groups?

P-value

In the simplest modeling approach we start by computing some summary statistics of association, such as effect size estimate $\widehat{\beta}$, its standard error SE and a P-value P, and proceed to do some inference from them. The purpose for using P-value is to see whether the observed data seem inconsistent with the null hypothesis. Typically the null hypothesis states that the variable is not important, or technically, that its effect size is 0. We have one null hypothesis per each variable.

P-value is a probability of getting something "at least as extreme" as what has been observed, if the null hypothesis was true.

Therefore, small P-value is taken as evidence that the null hypothesis may not be true. Logic goes that if P-value is very small then it would be very unlikely to observe the data at hand under the null hypothesis – and therefore either null hypothesis is not true or we have encountered an unlikely event. But note that P-value is NOT a probability that the null hypothesis is true: P-value is probability of data given a hypothesis NOT probability of hypothesis given data. P-value is not at all the whole story in statistical inference but we start from it.

A more formal definition of P-value

Suppose that we have

- observed data y,

- defined a null hypothesis (NULL) that determines a way to generate data sets that are similarly structured as y,
- defined a test statistic $t = t(Y)$ whose value can be computed from the data in such a way that larger values of the test statistic (in our opinion) imply higher discrepancy between data and the null hypothesis.

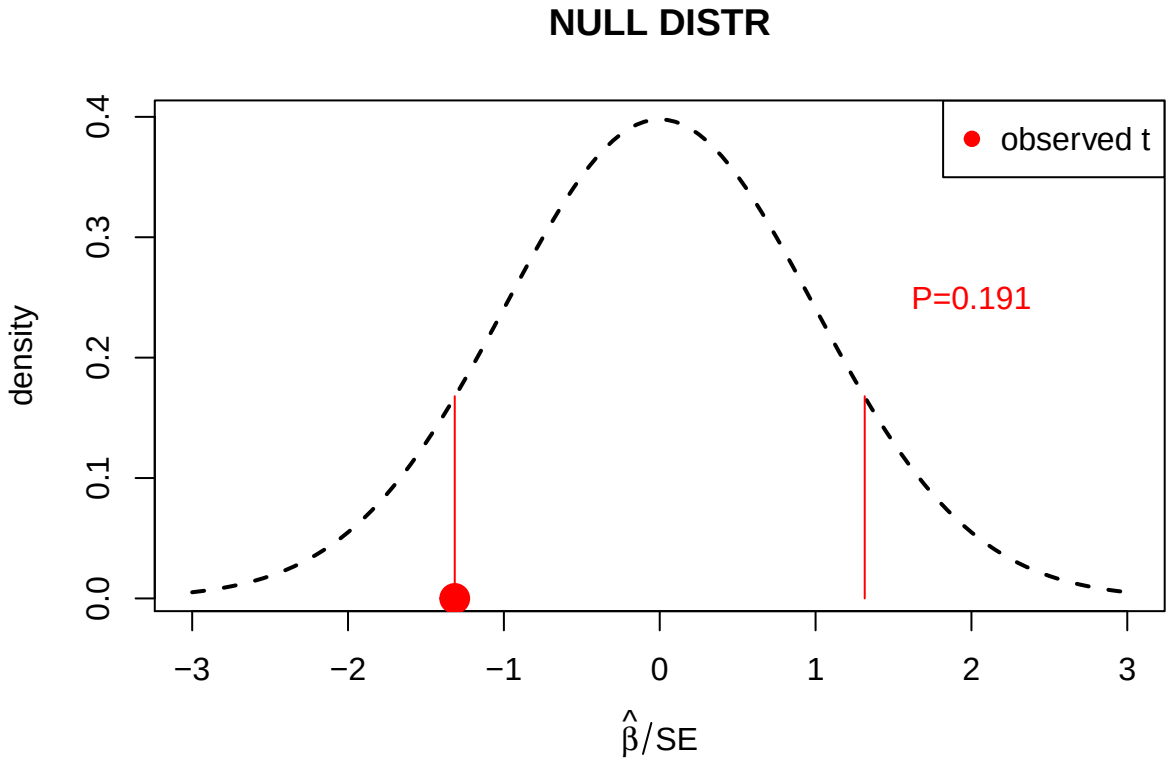
P-value (of data y) is a probability that if additional data Z were generated according to the null hypothesis, then the corresponding test statistic $t(Z)$ computed from data Z would be at least as large as our observed $t(y)$. That is,

$$\text{P-value of data } y \text{ is } \Pr(t(Z) \geq t(y) \,|\, Z \sim \text{NULL}).$$

Example 1.2

Let's do one linear regression with $p = 1$ and put its P-value in its place in the null distribution of t-statistic. The goal is to test whether the slope (β_1) of the model $y = \beta_0 + x\beta_1 + \varepsilon$ is zero. The null hypothesis is $H_0 : \beta_1 = 0$. By fitting the linear model with `lm()` we get the estimate $\widehat{\beta}_1$ of slope and its P-value. Here P-value tells that if the true slope $\beta_1 = 0$, what is the probability that we observe a data set from which the computed slope is at least as large (in absolute value) as the observed $\widehat{\beta}_1$. Most often we don't look at the null distribution of $\widehat{\beta}_1$, which depends on sample size and variances of x and ε, but instead we look at the null distribution of the t-statistics $t = \widehat{\beta}_1/\text{SE}_1$ which has distribution $t(n-p-1)$, i.e., t-distribution with $n-p-1$ degrees of freedom. (When $n-p-1 > 50$, $t(n-p-1)$ is very accurately the same as $\mathcal{N}(0,1)$ and hence doesn't noticably depend on the sample size.)

```
set.seed(29)
n = 100
x = rnorm(n) # predictor
y = rnorm(n) # outcome, independent of x
lm.fit = lm(y ~ x)
x.grid = seq(-3, 3, 0.05) # to define the plotting region
plot(x.grid, dt(x.grid, df = n-2), lty = 2, lwd = 2, t = "l",
     xlab = expression(hat(beta)/SE), ylab = "density", main = "NULL DISTR") #null distribution of t-st
t.stat = summary(lm.fit)$coeff[2,3] # observed t-statistic: Estimate/SE
points(t.stat, 0, pch = 19, cex = 2, col = "red")
segments(t.stat*c(1,-1), c(0,0), t.stat*c(1,-1), rep(dnorm(t.stat,0,1),2), col="red")
text(2, 0.25, paste0("P=",signif(summary(lm.fit)$coeff[2,4],3)), col="red")
legend("topright",pch=19,col="red",leg="observed t")
```



P-value is the probability mass outside the red segments, i.e., the sum of the two tail probabilities. It tells how probable, under the null, it is to get at least as extreme (from 0) observation as we have got here.

GAME: Guess a P-value

What is (approximately) the P-value

1. That in 10 fair coin tosses we get 9 Heads and 1 Tails? Is it 10^{-2} or 10^{-7} or 10^{-17}?
2. That in 100 fair coin tosses we get 90 Heads and 10 Tails? Is it 10^{-2} or 10^{-7} or 10^{-17}?
3. That if we (say, 20 people) would make a random sitting order, you would keep your current place? Is it 0.05 or 10^{-3} or 10^{-20}?
4. That if we (say, 20 people) would make a random sitting order, no-one would change places? Is it 0.05 or 10^{-3} or 10^{-20}?
5. That physicists used as significance level to claim Higg's Boson found in 2012? Is it 0.05 or 10^{-4} or 10^{-7}?

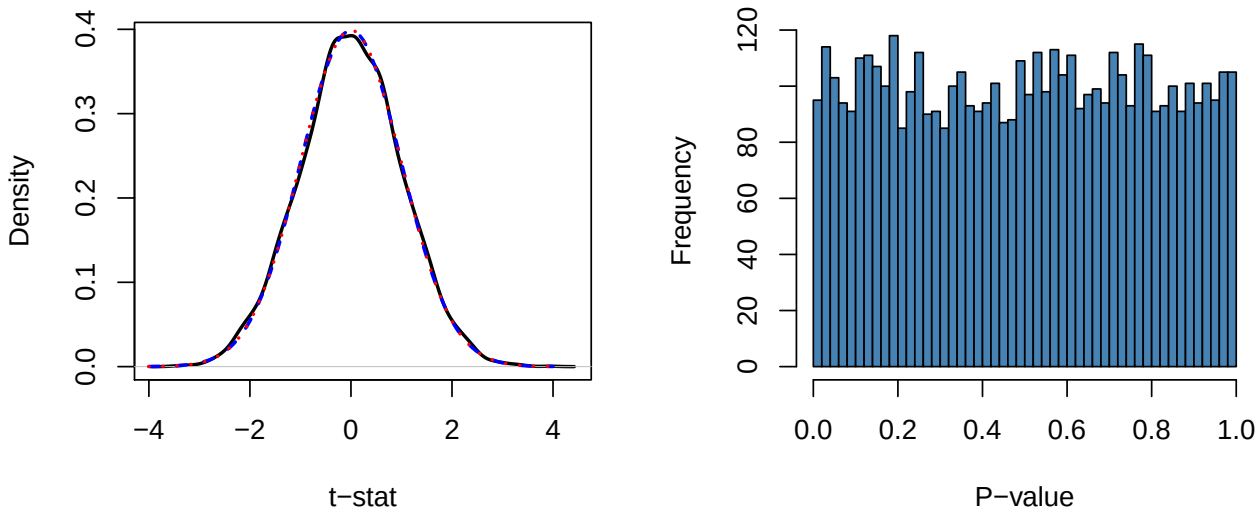
Recently, there has been much discussion about how traditional use of P-values ("significance testing at P < 0.05 threshold") has led to poor replicability of the reported scientific findings and hence poor science, in particularly with high dimensional data sets with "multiple testing issues". We will have a look at this problem next. For this discussion, see also

- R. Nuzzo: Scientific method - statistical errors from 2014.
- American Statistical Association's Statement on Statistical Significance and P-Values from 2016.

Distribution of summary statistics

Let's generate some data, first without any real effects. Our purpose is to see how a large set of P-values behave.

```r
set.seed(6102017)
n = 1000 #individuals
p = 5000 #variables measured on each individual
X = matrix( rnorm(n*p), n, p) #just random variables
y = rnorm(n) #outcome variable that is not associated with any of x

#by mean-centering y and each x, we can ignore intercept terms (since they are 0, see Lecture 0)
X = as.matrix( scale(X, scale = F) ) #mean-centers columns of X to have mean 0
y = as.vector( scale(y, scale = F) )

#apply lm to each column of X separately and without intercept (see Lecture 0.)
lm.res = apply(X, 2 , function(x) summary(lm(y ~ -1 + x))$coeff[1,])
# lm.res has 4 rows: beta, SE, t-stat and P-value
pval = lm.res[4,] #pick P-values

par(mfrow = c(1,2))
plot(density(lm.res[3,]), sub = "", xlab = "t-stat", main = "", lwd = 2) #should be t with n-2 df
curve(dt(x, df = n-2), from = -4, to = 4, add = T, col = "blue", lwd = 2, lty = 2) #t distr in blue
curve(dnorm(x, 0, 1), from = -4, to = 4, add = T, col = "red", lwd = 2, lty = 3)#normal distr in red
hist(pval, breaks = 50, xlab = "P-value", main = "", col = "steelblue")
```



On left we see that the empirical distribution of t-statistic (black) accurately follows its theoretical $t(n-2)$ distribution (blue), and that since n is large enough, this distributions is indistinguishable from the normal distribution $\mathcal{N}(0,1)$ (red).

Histogram on right shows that the P-values seem distributed uniformly between 0 and 1. This is indeed their distribution when the data follows the null hypothesis, as we will establish later. (To quantitatively assess whether the histogram truly looks "uniform", we can determine that under the unifrom distribution we would expect each bin to have $p/50 = 5000/50 = 100$ P-values and that with $\geq 95\%$ probability, in any one bin, the value would be within interval

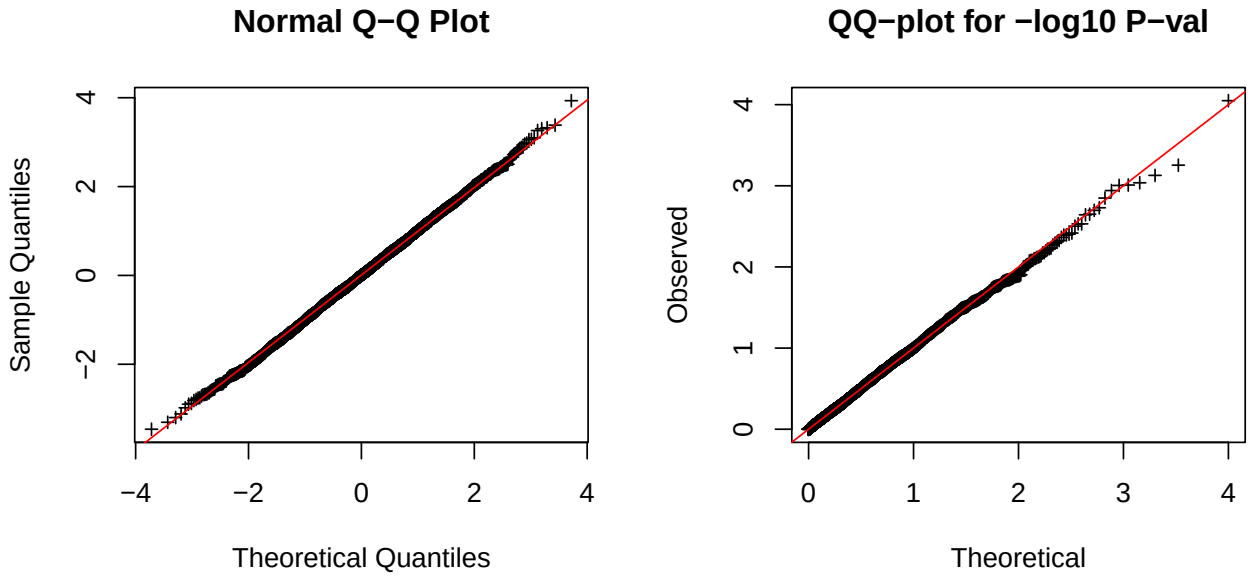
```r
qbinom(c(0.025,0.975), size = p, prob = 1/50)
```

```
## [1]  81 120
```

Thus, the variation in the histogram is consistent with the null distribution.)

Let's also compare the distributions via QQ-plots. First t-statistics against the normal distribution and then P-values against the uniform distribution. To see particularly well the smallest P-values, which are often the most interesting, we will show P-values on -log10 scale.

```r
par(mfrow=c(1,2)) #Let's make qqplots for t-stats and for P-values
qqnorm(lm.res[3,], cex = 0.7, pch = 3); qqline(lm.res[3,], col = "red")
#((1:p)-0.5) / p gives us
#p equally spaced values in (0,1) to represent quantiles of Uniform(0,1).
qqplot(-log10( ((1:p)-0.5) / p), -log10(pval), xlab = "Theoretical",
       ylab = "Observed", main = "QQ-plot for -log10 P-val", cex = 0.7, pch = 3)
abline(0, 1, col = "red")
```



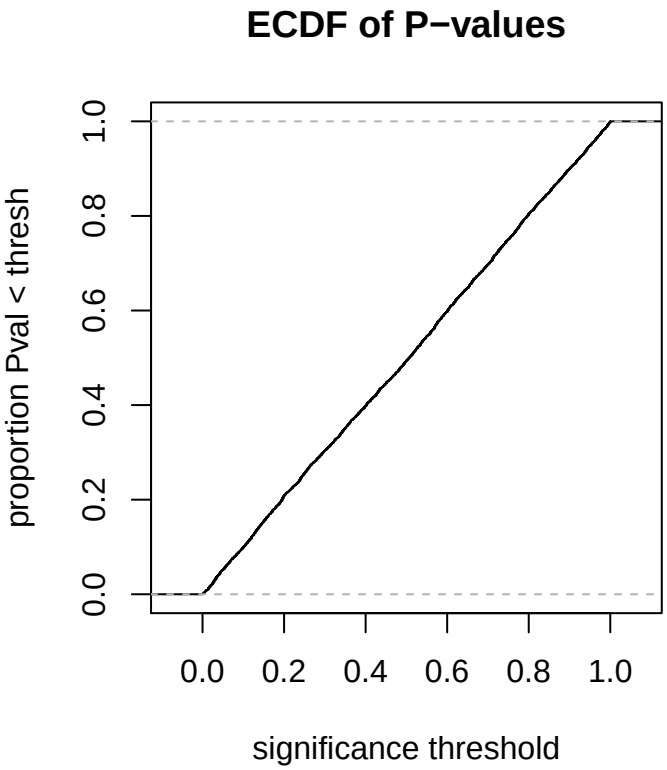
What are QQ-plots? In QQ-plot quantiles of two distributions are plotted against each other. If the distributions are similar, then the differences between adjacent quantiles are proportional between the distributions and QQ-plot forms a line. In the above plots, the red line can be used to assess visually whether the points seem to be close to it, in which case the two distributions are similar. We conclude that here t-statistics follows well the standard Normal distribution and P-values follow the Uniform(0,1).

Why are P-values distributed as Uniform(0,1) under the null?

$$\Pr(P \leq q_0 \,|\, \text{NULL}) = \Pr(\text{test stat falls within the most extreme region of prob. mass } q_0 \,|\, \text{NULL}) = q_0,$$

thus the cumulative density function (cdf) of P-value is $F(x) = x, 0 \leq x \leq 1$, which equals the cdf of Uniform(0,1). Let's use empirical cdf of the P-values to demonstrate this.

```r
par(pty = "s")
plot(ecdf(pval), xlab = "significance threshold", ylab="proportion Pval < thresh",
     main = "ECDF of P-values")
```



We have named the x-axis as "significance threshold" following the traditional framework where certain cutoff for P-values is used to label P-values as "significant" or "non-significant". From y-axis we can read which proportion of P-values of these random predictors reach each possible significance threshold in (0,1). This confirms empirically that, for any given threshold α, the proportion of P-values from the null that are $\leq \alpha$ is expected to be α.

For example, if we would use a standard significance threshold $\alpha = 0.05$ to determine "statistical significance" of each predictor, we would here label

```
sum( pval < 0.05 )
```

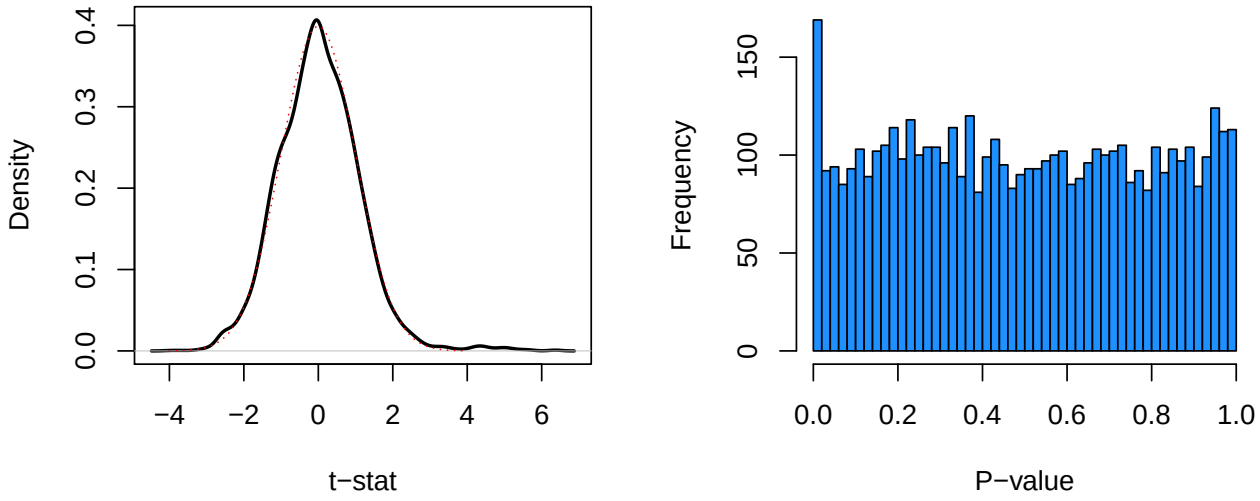
```
## [1] 259
```

$\approx 250 = 0.05 \cdot 5000$ predictors as "significant" even though they were all just random noise. If we had had done the test for $p = 10000$ predictors, we would expect 500 of them to have reached $P < 0.05$ even when none of them truly had non-zero effects, and so on. This increasing flood of false positives with an increasing number p of tests is a **multiple testing problem** arising in the standard hypothesis testing framework, when the significance level α is kept fixed while p grows.

Let's then add some ($m = 50$) predictors that have non-zero effects on the outcome y. Now our data will have m predictors with non-zero effects and $p - m$ predictors with zero effects.

```
set.seed(6102017)
n = 1000 #individuals
p = 5000 #variables measured on each individual
m = 50 #number of predictors that have an effect: they are x_1,...,x_m.
b = 0.5 #effect size of predictors that have an effect
X = matrix(rnorm(n*p), n, p) # random predictors
y = X[,1:m] %*% rep(b,m) + rnorm(n) #outcome variable that is associated with x_1,...,x_m

#by mean-centering y and each x, we can ignore intercept terms (since they are 0)
X = as.matrix(scale(X, scale = F)) #mean-centers columns of X
```

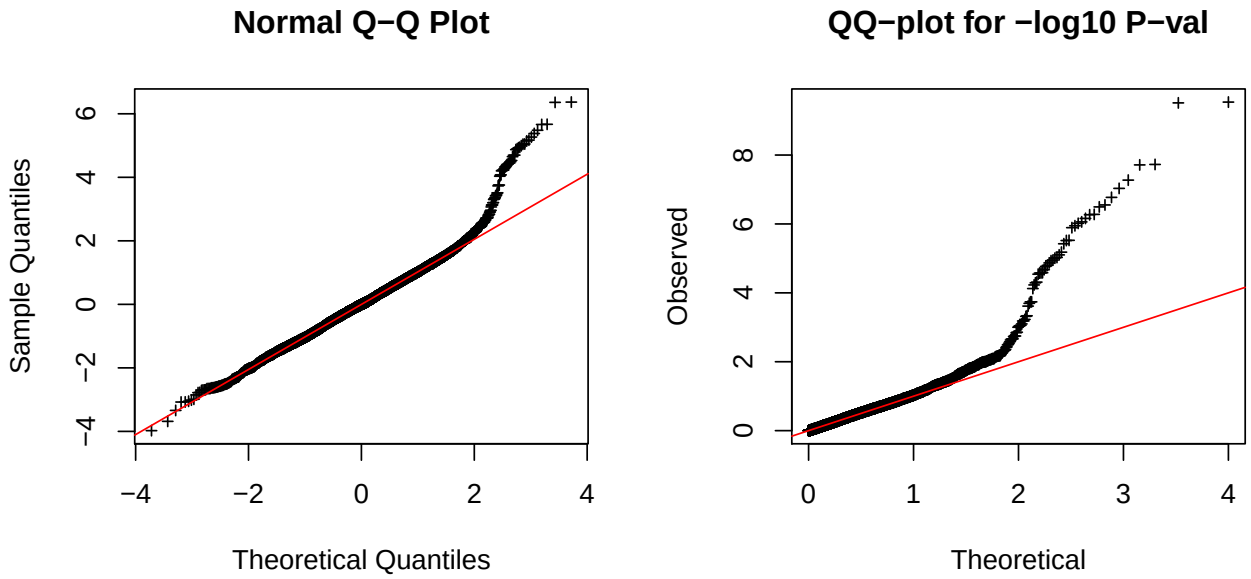
```r
y = as.vector(scale(y, scale = F))

#apply lm to each column of X separately and without intercept
lm.res = apply(X, 2 , function(x) summary(lm(y ~ -1 + x))$coeff[1,])
#has 4 rows: beta, SE, t-stat and pval
pval = lm.res[4,]

par(mfrow = c(1,2))
plot(density(lm.res[3,]), sub = "", xlab = "t-stat", main = "", lwd = 2) #under null is t with n-2 df
curve(dnorm(x), -4, 4, col = "red", lty = 3, add = T) #normal distribution in red
hist(pval, breaks = 50, xlab = "P-value", main = "", col = "dodgerblue")
```



The left plot shows that the density function of the t-statistic has longer tail to the right than the Normal distribution. This is because the non-zero predictors were chosen to have positive effects in these data. Since the proportion of the non-zero predictors is small (1%), the longer right tail is not that clearly observable in the density function. The histogram, instead, shows clearly that the P-value distribution differs from the null assumption of uniform distribution by an enrichment of the smallest P-values.

Let's try QQ-plots.

```r
par(mfrow = c(1,2)) #Let's make qqplots for t-stats and for P-values
qqnorm(lm.res[3,], cex = 0.6, pch = 3)
qqline(lm.res[3,], col = "red")
#Now we use ppoints() to give the p quantiles from the uniform:
qqplot(-log10(ppoints(p)), -log10(pval), xlab = "Theoretical",
       ylab = "Observed", main = "QQ-plot for -log10 P-val", cex = 0.6, pch = 3)
abline(0, 1, col = "red")
```



With QQ-plots, we see a clear deviation from the null distribution at the right tail of the test statistic (left panel) as well as in the smallest P-values (right panel). The right-side QQ-plot shows that, under the null, we would expect that the smallest P-value would be around 10^{-4}, whereas we observe at least ~30 P-values $< 10^{-4}$ with the smallest P-values around 10^{-10}.

The next question is how can we make a statistically sound inference about which ones of the tested predictors are non-zero effects?

Multiple testing framework

Let's introduce some notation using traditional terminology from hypothesis testing. Let H_j = "predictor j is null", be the null hypothesis for predictor x_j, $j = 1, \ldots, p$. We "reject H_j" if there is statistical evidence that x_j may not be null, otherwise we "accept H_j". Another names for "rejecting null" are labeling predictor "statistically significant" or "significant", or calling it a "discovery". Typically we mean by these that the predictor is interesting enough to deserve further examination/replication attempts etc. It typically does NOT mean that we are highly certain about the non-null status. We test p predictors and assume that p_0 of them are truly null (but of course we can't know either p_0, nor which ones of the predictors are truly null). Let's use the following notation:

Test result / True state of the world	null	not null	Total
positive (significant, discovery)	FD	TD	D
negative (not significant)	TN	FN	$p - D$
Total	p_0	$p - p_0$	p

- p is the total number of hypotheses tested.
- p_0 is the number of true null hypotheses, an unknown parameter.
- $p - p_0$ is the number of true not null ("alternative") hypotheses.
- FD is the number of false discoveries or false positives. (Type I errors.)
- TD is the number of true discoveries or true positives.
- FN is the number of false negatives or false non-discoveries. (Type II errors.)
- TN is the number of true negatives or true non-discoveries.
- $D = FD + TD$ is the number of discoveries, i.e., rejected null hypotheses.

Of these we only observe p and D and will make statistical inference on the rest.

P-values and family-wise error rate (FWER)

The simplest inference procedure is to fix a statistical significance threshold α and call each predictor **significant** (at significance level α) if its P-value $\leq \alpha$. We know that under the null the distribution of P-values is uniform, i.e., $\Pr(P \leq \alpha \mid \text{NULL}) = \alpha$ for $\alpha \in [0, 1]$. Thus α is also the **Type I error rate**, or **false positive rate**, the rate at which null predictors are labelled as significant. This can be written as $\mathrm{E}(FD/p_0) = \alpha$. The traditional significance level testing just controls the false positive rate.

For example, as we saw above, the common significance threshold $\alpha = 0.05$ means that 1/20 null predictors will reach $P \leq \alpha$ and therefore we expect that $\alpha \cdot p_0$ predictors reach significance level α, when we test p_0 independent null predictors. If, in a high-dimensional problem, we test $p_0 \approx 10^6$ predictors, we expect 50,000 significant results at significance level $\alpha = 0.05$ already even if there are no true positives to be found at all. Since increasing number of false discoveries is a problem, methods that control a much more stringent **family-wise error rate (FWER)** are often used in the multiple testing setting.

FWER is the probability of making at least one false discovery across all the tests carried out in a multiple testing setting:

$$\text{FWER} = \Pr(FD \geq 1) = \mathrm{E}(FD \geq 1).$$

Of these $\Pr(FD \geq 1)$ seems a more natural definition; the latter expectation formulation means the expected value of the indicator function of the event $\{FD \geq 1\}$, and it is added here for comparison with other methods that are often defined through expectations.

Example 1.3. Suppose that 10 independent groups do a clinical trial on the same drug on the same disease and one of the groups observes an effect at significance level 0.05 and publishes the result (while other groups don't observe the effect). What is the FWER of such a procedure under the null hypothesis? That is, what is the probability for the observation "at least one P-value ≤ 0.05 out of 10 independent P-values" under the null hypothesis that there is no real effect in any study?

$$\Pr(\text{at least one } P \leq 0.05 \,|\, \text{NULL}) = 1 - \Pr(\text{all } P > 0.05 \,|\, \text{NULL}) = 1 - (1 - 0.05)^{10} \approx 0.401.$$

So when we do 10 independent tests, each at significance level 0.05, the overall FWER of this set of tests is 0.401, that is, 8-fold compared to 0.05. This example shows why it is problematic that only "significant" results tend to get published: A proper assessment of the drug should be done based on all available 10 studies, NOT only on the one that happened to give "significant" result, since that study is likely to be biased towards larger effect size given that other studies didn't report "significant" results.

Bonferroni correction

The simplest way to control FWER at level α is to apply significance threshold $\alpha_B = \alpha/p$ for each test, i.e., report as significant the predictors whose P-value $\leq \alpha_B$. This is called the Bonferroni correction for multiple testing (after Italian mathematician Carlo Bonferroni). Proof that it does the job is

$$\text{FWER} = \Pr\left(\bigcup_{j=1}^{p_0} \{P_j \leq \alpha_B\} \,\middle|\, \text{NULL} \right) \leq \sum_{j=1}^{p_0} \Pr\left(P_j \leq \alpha_B \,|\, \text{NULL}\right) = p_0 \cdot \alpha_B = p_0 \frac{\alpha}{p} \leq p \frac{\alpha}{p} = \alpha.$$

This procedure does not assume anything about the dependency between separate tests or the proportion of truly null hypotheses. Its advantages are thus complete generality and very simple form that is easy to apply in practice.

Exercise. In genome-wide association studies (GWAS) a significance threshold 5×10^{-8} has become commonly used. If you think it as a result of Bonferroni correction to achieve FWER of 0.05, which assumption has been made about the number of null tests done in a GWAS?

Exercise. Assume that you test the association of 10 clinical variables (such as cholesterol levels or blood pressure) for association with a disease outcome Y, and the smallest P-value you get is 0.008 for variable X. What would you report as statistical evidence from this experiment?

Bonferroni correction controls FWER, but it is very stringent and hence has low statistical power to detect true effects. This has motivated a lot of research on how to improve power. As an example of such work, let's consider **Holm method**. It has only a small improvement on power over Bonferroni correction, but it serves as our introduction to step-wise testing procedures.

Holm method

- Order the P-values from the lowest to the highest: $P_{(1)} \leq \ldots \leq P_{(p)}$, and let the corresponding hypotheses be $H_{(1)}, \ldots, H_{(p)}$.
- For a given significance level α, let j be the smallest index such that $P_{(j)} > \frac{\alpha}{p+1-j}$.
- Reject the null hypotheses $H_{(1)}, \ldots, H_{(j-1)}$ and do not reject $H_{(j)}, \ldots, H_{(p)}$.
- If $j = 1$ then do not reject any of the null hypotheses and if no such j exist then reject all of the null hypotheses.

Proof that Holm method controls FWER. Let I_0 be the set of p_0 indexes of the true null hypotheses. Let k be the index of the first true null hypothesis among the order sequence of P-values, i.e., $H_{(1)}, \ldots, H_{(k-1)}$ are false but $H_{(k)}$ is true. We want to show that probability that $H_{(k)}$ is rejected is $\leq \alpha$. Since there are $p_0 - 1$ true nulls in the ordered sequence of hypothesis after $H_{(k)}$, it follows that

$$k + p_0 - 1 \leq p \implies \frac{1}{p+1-k} \leq \frac{1}{p_0} \implies \frac{\alpha}{p+1-k} \leq \frac{\alpha}{p_0}.$$

$$\Pr\left(P_{(k)} \leq \frac{\alpha}{p+1-k}\right) \leq \Pr\left(P_{(k)} \leq \frac{\alpha}{p_0}\right) = \Pr\left(\bigcup_{i \in I_0}\left(P_i \leq \frac{\alpha}{p_0}\right)\right) \leq \sum_{i \in I_0} \Pr\left(P_i \leq \frac{\alpha}{p_0}\right) = p_0 \frac{\alpha}{p_0} = \alpha.$$

Example 1.4. Suppose we have 5 P-values $\{0.4, 0.001, 0.8, 0.011, 0.12\}$. Which hypotheses would be rejected at FWER of 0.05 using Bonferroni method or using Holm method?

```
fwer = 0.05
p.ex = 5 #use ".ex" to not mix up with p=5000 existing variables that we will reuse later
pval.ex = c(0.4, 0.001, 0.8, 0.011, 0.12)
#Bonferroni rejects:
(pval.ex <= fwer/p.ex)
```

```
## [1] FALSE  TRUE FALSE FALSE FALSE
```

```
#For Holm we first sort P values in ascending order
sorted.pval = sort(pval.ex)
#we compute individual rejection threshold for EACH hypothesis in ascending order
alpha.holm = fwer/( p.ex + 1 - (1:p.ex) )
rbind(sorted.pval, alpha.holm)
```

```
##              [,1]   [,2]       [,3]  [,4] [,5]
## sorted.pval 0.001 0.0110 0.12000000 0.400 0.80
## alpha.holm  0.010 0.0125 0.01666667 0.025 0.05
```

```
#Let's find min index where P-value > holm threshold. We reject smaller indexes.
i = min( which(sorted.pval > alpha.holm) )
paste("reject:", paste(sorted.pval[1:(i-1)], collapse=" ") )
```

```
## [1] "reject: 0.001 0.011"
```

Here Bonferroni rejected only 0.001 while Holm rejected 0.001 and 0.011.

Holm method is (slightly) more powerful than Bonferroni, because the P-value threshold for rejecting the null is higher except for the hypothesis having the smallest P-value, in which case the threshold is the same in both methods (α/p).

Exercise. Assume that both Bonferroni and Holm methods have rejected the hypotheses corresponding to the k smallest P-values. What is the ratio of P-value thresholds that is needed to reject the next hypothesis with these methods?

How do we do Bonferroni and Holm corrections in R? Let's use our existing data where we had $p = 5000$ predictors and $m = 50$ of them were true effects and $p_0 = p - m = 4950$ were null and P-values are stored in `pval`.

```
p.thresh = 0.5 #this is very liberal significance level for raw P-values, but not after FWER adjustment
sum( pval < p.thresh )
```

```
## [1] 2545
```

```
sum( p.adjust(pval, method = "holm") < p.thresh )
```

```
## [1] 37
```

```
sum( p.adjust(pval, method = "bonferroni") < p.thresh )
```

```
## [1] 37
```

```
#Let's see how many true and false positives we have
signif.tests = (pval < p.thresh)
S = sum(signif.tests[1:m]) #True positives
V = sum(signif.tests[(m+1):p]) #False positives
print(paste("Raw P-values: TP =",S,"FP =",V))
```

```
## [1] "Raw P-values: TP = 50 FP = 2495"
```

```
signif.tests = (p.adjust(pval, method="holm") < p.thresh)
S = sum(signif.tests[1:m]) #True positives
V = sum(signif.tests[(m+1):p]) #False positives
print(paste("Holm: TP =",S,"FP =",V))
```

```
## [1] "Holm: TP = 35 FP = 2"
```

Bonferroni and Holm methods gave the same inference here. By controlling FWER at 0.5 we got 35 of true positives and only 2 false positives. By looking at P-values directly, all 50 true positives and over 2495 false positives(!) reached the level 0.5.

Note that `p.adjust` literally adjusts the P-values, i.e., it multiplies the P-value by the correction factor that for Bonferroni method is p and for Holm method is $p + 1 - k$ for the kth smallest P-value. In addition, for Holm method, it makes sure that the adjusted P-values have the same order as the unadjusted ones, by taking maximum over all adjusted P-values on the left hand side of each adjusted P-value when the adjusted P-values are in ascending order.

Criticism towards FWER control

By controlling FWER we can clearly keep the number of false positives low in the total experiment. However, the price for requiring so stringent statistical evidence is that we may also lose a lot of true positives. Eventually, the balance between avoiding false positives and catching true positives needs to be set by lossess associated with each of these types of errors. Often FWER control becomes problematic when there are many discoveries to be made and not a large penalty for making a few false positives as well. FWER methods simply do not have power to make those discoveries because they are so much afraid of making false positives. Therefore, less stringent multiple testing correction methods have been developed to control the *false discovery rates* (FDRs), which will be our next topic.

Another, and maybe even larger conceptual problem with FWER is to justify why the goal is to control *simultaneously* particularly this one set of possibly *independent* hypotheses in a frequentist manner. The world is full of all kinds of hypotheses, why do we want to control for some of them jointly, but not all of

them? How can we choose which to consider jointly? We will come back to this conceptual problem later and seek answers from Bayesian inference.