GWAS 2: P-values in GWAS

Matti Pirinen, University of Helsinki

Updated: March 6, 2025

The slide set referred to in this document is "GWAS 2".

A GWAS conducted by the additive regression model returns three quantities for each variant:

- $\widehat{\beta}$, the effect size estimate for allele 1 ("*effect allele*");
- SE of $\widehat{\beta}$, describing uncertainty in $\widehat{\beta}$;
- P-value.

We saw earlier how to interpret β as a difference between the means of the adjacent genotype groups, and how SE describes the uncertainty of the estimate $\widehat{\beta}$ (slide 2). These two parameters tie the data to the actual phenotypic change in a concrete way. For example, from these quantities we could infer that a variant increases LDL levels by 0.5 mmol/L (95%CI 0.42..0.58) or that another variant increases odds of MS-disease by 20% (95%CI 16%...24%). While these are the most concrete and detailed information about the effect of a variant available in GWAS output, still, typically, the first statistic we look at is the P-value.

The P-value is widely used as a central summary statistic in GWAS because the null hypothesis of effect size being exactly zero is thought to be a realistic hypothesis for a large majority of all variants in the genome, and a crude assessment of zero vs. non-zero effect size in the GWAS context can be done based on the P-value. Additionally, a genotype-phenotype association can be interesting already because it points to biology behind the phenotype, no matter how large the actual effect size of the variant is. Indeed, some parts of the human genome, that contribute to the risk of certain disease, may tolerate only such genetic variation that has a very small effect on the disease risk, since natural selection may efficiently remove any risk variants with larger effects from the population. Still, by discovering such parts of the genome, one may be able to create therapeutic interventions that affect the same biological disease mechanisms with much stronger (beneficial) effects than what exist in the genomes of a natural population. Therefore, the magnitude of the effect size has a smaller role in statistical inference in the GWAS context than in many other contexts, such as, e.g., in social sciences, where it is unrealistic to expect that any effect size is exactly zero. In those other research fields, the focus should be first and foremost on the size of the effect whereas there the P-value is a less relevant quantity.

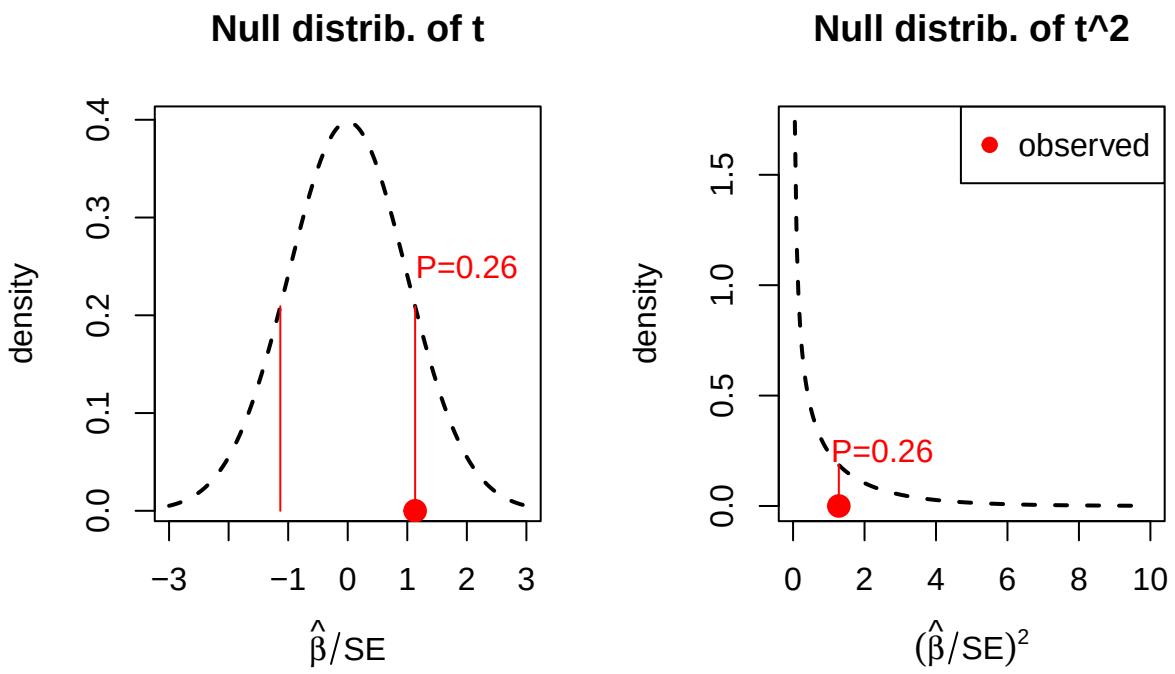
2.1 What is the P-value? (slides 5-8 & 20-23)

The purpose of using the P-value is to evaluate whether the observed data seem inconsistent with the null hypothesis. Typically, the null hypothesis states that the variant is not important, or technically, that its effect size is 0. We have one null hypothesis per each variant. The P-value is a probability of getting something "at least as extreme" as what has been observed, if the null hypothesis was true. Therefore, a small P-value is taken as evidence that the null hypothesis may not be true. The logic goes that if the P-value is very small, then it would be very unlikely to observe "those kinds of data sets" under the null hypothesis – and therefore either the null hypothesis is not true or we have encountered an unlikely event. The P-value is thus a simple numerical summary of the consistency between the null hypothesis and the

observed data. But note that the P-value **is not** a probability that the null hypothesis is true: The P-value is a probability of certain types of data sets given a hypothesis, and the P-value **is not** a probability of the hypothesis given data (slides 10-11).

Let's do one linear regression and put its P-value in its place in the null distribution of t-statistic. The goal is to study whether the effect β of the additive GWAS model $y = \mu + x\beta + \varepsilon$ is zero. The null hypothesis is $H_0 : \beta = 0$. Here, the P-value tells that if the true slope $\beta = 0$, what is the probability that we would observe a data set from which the computed slope is at least as large (in absolute value) as the observed estimate $\widehat{\beta}$. Most often, we don't look at the null distribution of $\widehat{\beta}$, which depends on the sample size and variances of x and ε, but instead we look at the null distribution of the t-statistics $t = \widehat{\beta}/\text{SE}$ which is $t(n-2)$, i.e., the t-distribution with $n-2$ degrees of freedom, where n is the sample size. When $n > 50$, we can approximate $t(n-2)$ with the standard normal distribution $\mathcal{N}(0,1)$, and this approximation is typically done in the GWAS setting, where the t-statistic is often called z-score that refers to a variable that has a standard normal distribution. The other way to express the same normal approximation is to say that the square of the t-statistic follows a chi-square distribution with 1 degree of freedom: $t^2 \sim \chi^2_1$. By squaring t, this approach ignores the sign of the estimate, which is not a problem in the GWAS setting, since we do not, by default, have any prior knowledge on the direction of the effect. Let's draw pictures about P-values using both t and t^2 statistics.

```
n = 100
f = 0.3 #MAF
x = rbinom(n, 2, f) #example genotypes for n individuals
y = rnorm(n) #outcome that is independent of x
lm.fit = lm( y ~ x )
par( mfrow = c(1,2) ) #draw 2 panels on the grid with 1 row and 2 cols
#1st on t-statistic's scale
x.grid = seq(-3, 3, 0.05) #we need this 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 distrib. of t") #null distr. of t-stat.
t.stat = summary(lm.fit)$coeff[2,3] #t-statistic: Estimate/SE
points(t.stat, 0, pch = 19, cex = 1.5, col = "red")
segments(t.stat*c(1,-1), c(0,0), t.stat*c(1,-1), rep( dt( t.stat, df = n - 2), 2 ), col = "red")
text(2, 0.25, paste0("P=",signif(summary(lm.fit)$coeff[2,4],3)), col = "red")

#2nd on t^2 statitstic's scale
x.grid = seq(0, 10, 0.05)
plot(x.grid, dchisq( x.grid, df = 1 ), lty = 2, lwd = 2, t = "l",
     xlab = expression(( hat(beta)/SE)^2 ), ylab = "density",
     main = "Null distrib. of t^2") #null distribution of t^2-stat.
t2.stat = summary(lm.fit)$coeff[2,3]^2 #t^2-statistic: (Estimate/SE)^2
points(t2.stat, 0, pch = 19, cex = 1.5, col = "red")
segments(t2.stat, 0, t2.stat, dchisq(t2.stat, df = 1), col = "red")
text(2.5, 0.25, paste0("P=", signif(summary(lm.fit)$coeff[2,4],3)), col = "red")
legend("topright", pch = 19, col = "red", leg = "observed" )
```



The P-value is the probability mass outside the red segments, i.e., on the left panel it is the sum of the two tail probabilities and on the right panel it is the right tail probability. It tells how probable, under the null, it is to get at least as extreme (from 0) observation as we have got here. Note that the chi-square distribution leads to a simpler set-up since we need to consider only the right-hand tail of the distribution to compute the P-value. For the normal distribution we need to account also for the other tail that has the different sign from the observed value of the statistic (here the left tail).

Let's make sure that we understand where the P-value came from and let's compute it manually from both the standard normal distribution and from the chi-square distribution using the cumulative density functions.

```
z = summary(lm.fit)$coeff[2,3] #t-statistic also called z-score under Normal approximation
pnorm(-abs(z), 0, 1, lower = T) + pnorm(abs(z), 0, 1, lower = F) #P-value from N(0,1): left + right tai
```

```
## [1] 0.2577372
```

```
pchisq(z^2, df = 1, lower = F) #P-value from chi-square is the upper tail
```

```
## [1] 0.2577372
```

In this example, the P-value was 0.26. To interpret this value through a frequentist inference framework, we imagine an experiment where we were repeatedly sampling new data at this SNP without a real genotype-phenotype association, i.e. the true effect size is exactly zero. We would observe that in about 26% of those data sets the effect size estimate would be at least as large in absolute value as what we have observed here. We consider that these data do not give us any clear indication of a true genotype-phenotype association, because effect size estimates at least this large in absolute value are present already in about 1 in 4 of the null data sets. Hence, these data seem completely plausible to have originated from the null.

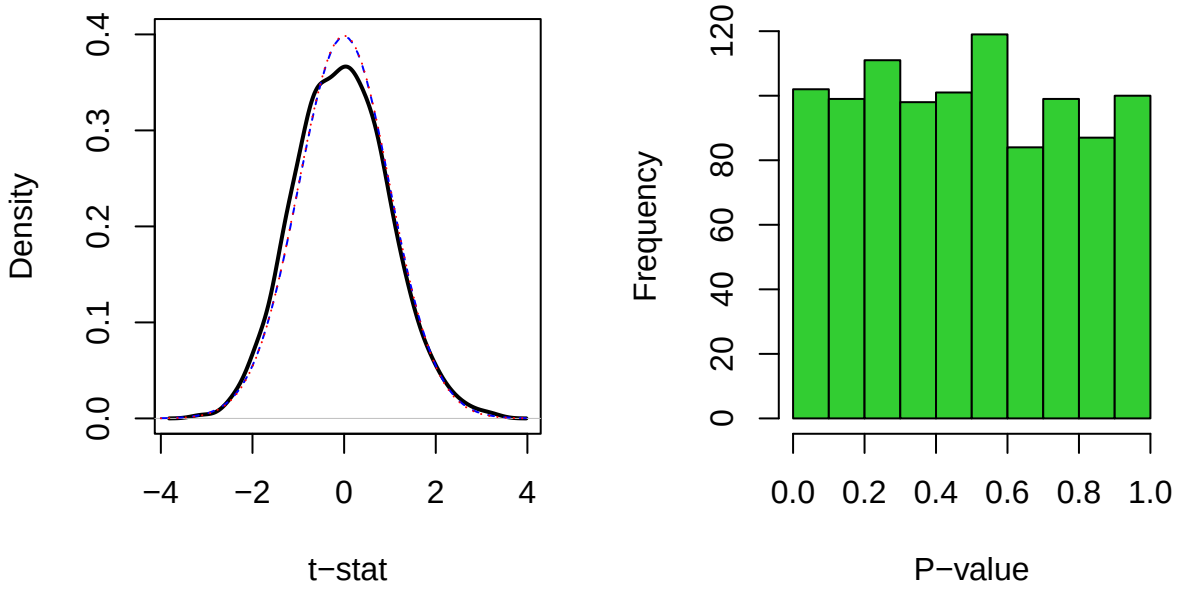
IMPORTANT: Make sure you understand how the above interpretation of the P-value is different from a **wrong interpretation** that based on the P-value only we could say that the probability that this SNP is null is 26% (slides 10-11).

If P=0.26 is not yet a convincing association, then how small a P-value should be to make us reasonably convinced that there is an interesting genotype-phenotype association?

2.2 Distribution of P-values

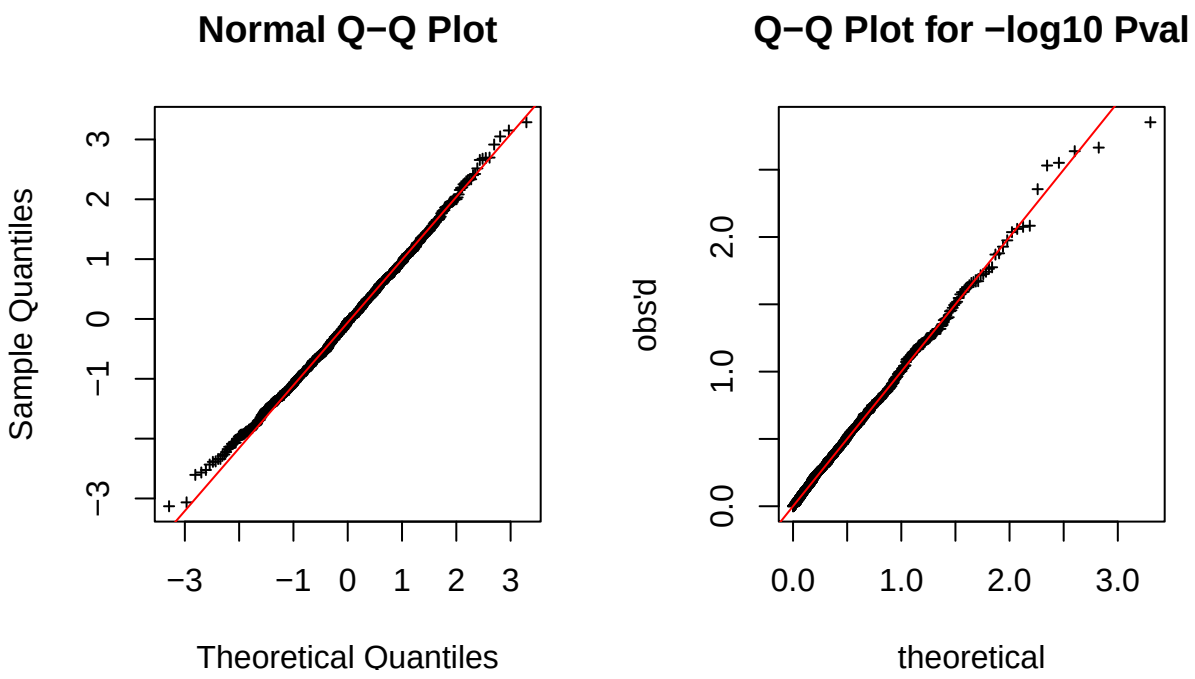
Let's make data on 1000 null variants, whose true effect size is zero, each measured on 100 individuals and look at their P-value distributions.

```
set.seed(39)
n = 100 #individuals
p = 1000 #variants measured on each individual
f = 0.4 #MAF is assumed constant across variants; doesn't actually matter here
X = matrix(rbinom(n*p, 2, f), nrow = n, ncol = p) #just random genotypes
y = rnorm(n) #phenotype that is not associated with any of the variants

#apply lm to each column of X separately and collect results for genotype (row 2 of coeff)
lm.res = apply(X, 2 , function(x) summary(lm(y ~ x))$coeff[2,])
#result has 4 rows: beta, SE, t-stat and pval
pval = lm.res[4,] #pick pvalues

par(mfrow = c(1,2))
plot(density(lm.res[3,]), ylim = c(0,0.4), sub = "",
     xlab = "t-stat", main = "", lwd = 2) #should be t with n-2 df
x.seq = seq(-4, 4, 0.1) #x-coordinates for plotting
lines(x.seq, dt(x.seq, df = n - 2), col = "blue", lty = 2) #t distribution in blue
lines(x.seq, dnorm(x.seq), col = "red", lty = 3) #normal distribution in red
hist(pval, breaks = 10, xlab = "P-value", main = "", col = "limegreen") #should be uniformly distribute
```




```
par(mfrow = c(1,2)) #Let's make qqplots for t-stats and for P-values
qqnorm(lm.res[3,], cex = 0.5, pch = 3) #t with ~100 df ~ normal, hence qqnorm()
qqline(lm.res[3,], col = "red")

# For P-values, we want to compare to the Uniform(0,1) distribution:
# We use ppoints(p) to get
# p equally spaced values in (0,1) to represent quantiles of Uniform(0,1).
# we take -log10 transformation to see the small P-values particularly well
qqplot(-log10(ppoints(p)), -log10(pval), xlab = "theoretical",
```

```
       ylab = "obs'd", main = "Q-Q Plot for -log10 Pval", cex = 0.5, pch = 3)
abline(0, 1, col = "red")
```



What are QQ-plots?

Why are P-values distributed as Uniform(0,1) under the null? The cumulative distribution function (cdf) of P-values under the null is

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

which is also the cdf of the Uniform(0,1) distribution.

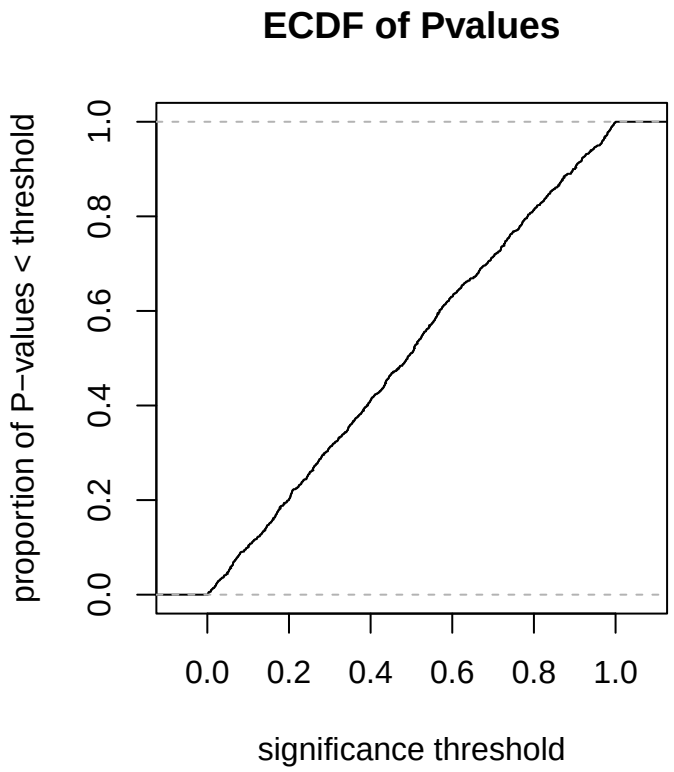
Conclusion: We have just seen that under the null hypothesis the t-statistic of effect size estimate is essentially standard normal (shown by both the density function and QQ-plot) and that the P-values under the null hypothesis are uniformly distributed between 0 and 1.

Significance threshold P-value quantifies incompatibility between observed data and the null hypothesis. Throughout applied statistics, it is also common to use the P-value to label some of the observations as showing "**statistically significant**" deviation from the null. This happens by fixing a reference threshold for the P-value **before data are observed**, and then calling the observations that reach the reference threshold as "significant" or "discoveries". The idea is to highlight those observations that seem interesting given their possible inconsistency with the null hypothesis. But the idea is NOT that some fixed "significance" threshold should be used to declare truth (slide 12). Additionally, the exact P-value is always a more valuable piece of information than the binary classification into a "significant" or "non-significant" P-value.

For such significance testing, the P-value threshold of 0.05 has become the most common reference threshold (also called the significance threshold or significance level). By definition, we expect that 1 in 20 of the null data sets will result in a P-value lower than 0.05. Let's see in practice which proportion of the P-values of the random genotypes at 1000 SNPs we generated earlier reach each significance threshold. We plot an empirical cumulative distribution function (ecdf) of the P-values.

```
par(pty = "s")
plot(ecdf(pval), xlab="significance threshold",
```

```
     ylab="proportion of P-values < threshold",
     main="ECDF of Pvalues")
```




```
#For example, check how many are < 0.05?
sum( pval < 0.05 )
```

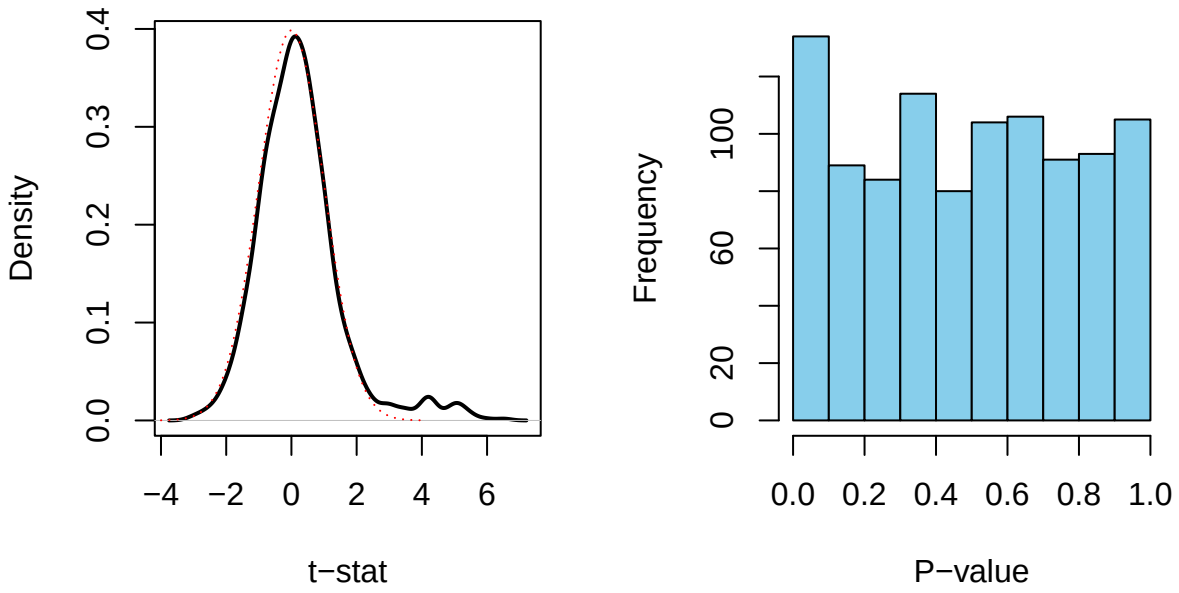
```
## [1] 47
```

This ecdf looks like the cdf of the Uniform(0,1), as expected, and if we used a standard significance threshold $\alpha = 0.05$ to label SNPs as "statistically significant", we would label about 50 variants out of all 1000 as significant even though these "significant" associations all originated from random data generation under the null. If we had analyzed $p = 1,000,000$ variants, as is typical in a GWAS, we would had expected 50,000 of them to reach $P < 0.05$ even in the case that none of them truly had a non-zero effect. This flood of false positives is the **multiple testing problem** arising from the standard significant testing framework. This is a considerable problem, particularly in settings like GWAS, where the number of true positives is small compared to the number of all tests, because, in that case, almost all "significant" discoveries will be false discoveries.

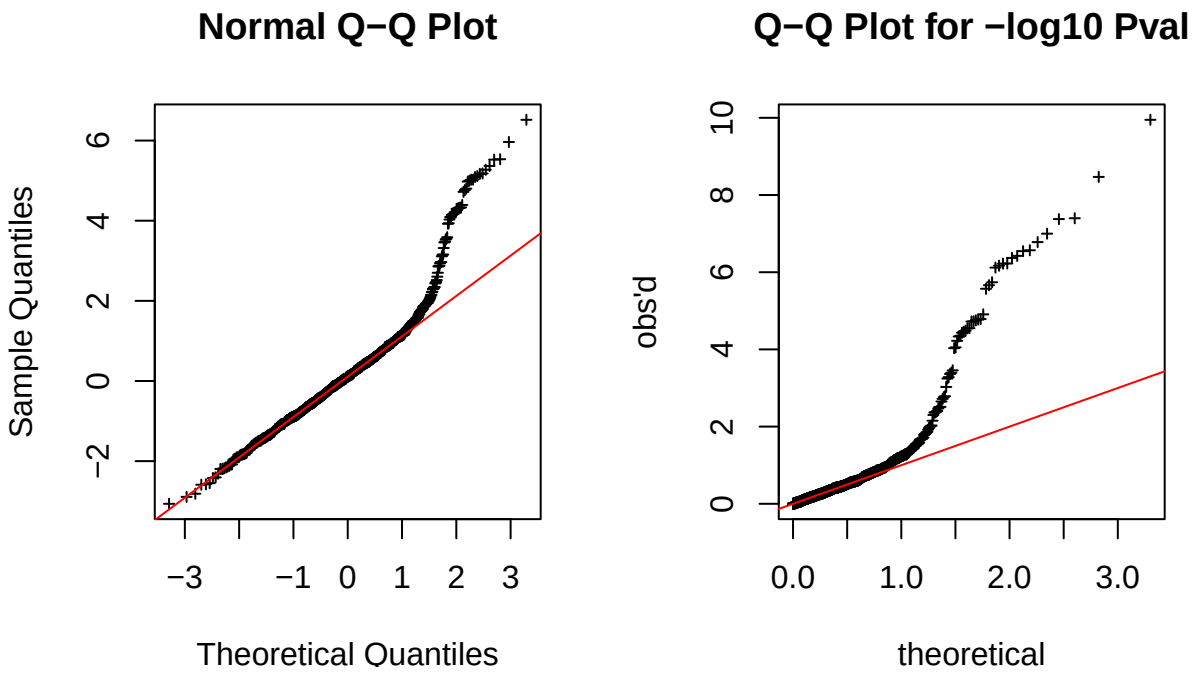
As a side note, conceptually, "multiple testing" is a slightly misleading term in the GWAS context, since there the problem of a large proportion of false discoveries for liberal thresholds like 0.05 would not occur primarily because of "multiple tests" but rather because a liberal threshold would not account properly for the small prior probability of any one variant to be truly associated with the phenotype. We'll talk more about this topic later.

After we have seen how the P-values of null variants behave, let's make things more interesting by adding $m = 50$ variants, that truly have an effect on the phenotype, among a set of $p = 1000$ variants that we generate.

```r
set.seed(49)
n = 1000 # individuals
p = 1000 # genotypes measured on each individual
m = 50  # number of variants that have an effect: they will be x_1,...,x_m.
f = 0.4 # MAF
b = 0.5 # effect size of variants that have an effect
X = matrix(rbinom(n*p, 2, f), nrow = n, ncol = p) # just random genotypes at SNPs
y = X[,1:m] %*% rep(b, m) + rnorm(n) # phenotype that is associated with x_1,...,x_m

#apply lm to each column of X separately
lm.res = apply(X, 2 , function(x) summary(lm(y ~ x))$coeff[2,])
#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
lines(seq(-4, 4, 0.1), dnorm(seq(-4, 4, 0.1)), col = "red", lty = 3) #normal distribution in red
hist(pval, breaks = 10, xlab = "P-value", main = "", col="skyblue") #under null is uniformly distribute
```




```r
par(mfrow = c(1,2)) #Let's make qqplots for t-stats and for P-values
qqnorm(lm.res[3,], cex = 0.5, pch = 3)
qqline(lm.res[3,], col = "red")
qqplot(-log10(ppoints(p)), -log10(pval), xlab = "theoretical", ylab = "obs'd",
       main = "Q-Q Plot for -log10 Pval", cex = 0.5, pch = 3)
abline(0, 1, col = "red")
```



We see a strong deviation from the null distribution, as expected, since there are now $m = 50$ non-zero effects among $p = 1000$ SNPs. This is what we would like to see in our GWAS as well: Some clear non-zero effects while most SNPs seem to follow the null (slide 15). Ideally, we would like to have an inference procedure that could tell which are the true non-zero effects and which are the null variants. We will next look at how a multiple testing framework conceptualizes this goal.

Note also the usefulness of a QQ-plot as a visual comparison of two empirical distributions, which comes from the fact that each observation can be seen on its own. By comparing density plots or cumulative distribution functions between two distributions, the differences caused by a set of points that are far in the tail are more difficult to see than in the QQ-plot, especially when the number of exceptional points is small compared to the number of all points.

For GWAS Manhattan plots see slides 16-17.

2.3 Multiple testing framework

Let's introduce some notation using traditional terminology from hypothesis testing. Let $H_j = \{\text{variant } j \text{ is null}\}$, be the null hypothesis for variant x_j, $j = 1, \ldots, p$. We "reject H_j" if there is enough statistical evidence that x_j is not null, otherwise we "do not reject H_j". Other terminology for "rejecting null" is to label the variant "statistically significant" or "significant" or to call it a "discovery". Typically, we mean by these phrases that the variant is interesting enough to deserve further examination and replication attempts etc. We test p variants and assume that p_0 of them are truly null but of course we can't know the value of p_0. Let's use the following symbols:

Test result	Truth = null	Truth = not null	Total
significant	FD	TD	D
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 error).
- TD is the number of true discoveries, true positives.
- FN is the number of false negatives or false non-discoveries. (Type II error)
- TN is the number of true negatives or true non-discoveries.
- $D = \text{FD} + \text{TD}$ is the number of discoveries / rejected null hypotheses / significant variants.

Of these quantities, we only observe p and D.

P-values and family-wise error rate As discussed before, the simplest inference procedure is to fix a statistical significance threshold α before data are observed, and call each variant significant if its P-value turns out to be $\leq \alpha$. We just saw that under the null the distribution of P-values is uniform, i.e., $\Pr(P \leq \alpha \,|\, \text{Null}) = \alpha$ for all $\alpha \in [0, 1]$. Thus α is also the **Type I error rate**, the rate at which the null variants are (incorrectly) labelled as significant.

As we saw earlier, the common significance threshold of $\alpha = 0.05$ means that about 1 in 20 of all null variants will have $P \leq \alpha$ and therefore we expect that about αp_0 null variants will reach the significance level α, when we test p_0 independent null variants. If, in a GWAS, we test $p_0 \approx 10^6$ null variants, we expect about 50,000 significant results at the significance level $\alpha = 0.05$ already when there are no true positives at all among the variants. Since such an increasing number of false discoveries is a problem, inference approaches that control a much more stringent **family-wise error rate (FWER)** are used.

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

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

If we use an inference procedure that keeps FWER very small, then we can be confident that there cannot be many false discoveries made in the study.

Example 2.1. (See also slide 13.) Suppose that 5 groups test the same missense variant (a mutation that changes the amino acid sequence of a protein) for association with MS-disease in five independent case-control data sets. Assume that one of the groups reports a non-zero effect at significance threshold 0.05 and publishes the result while the other 4 groups don't observe a non-zero effect at this threshold. What is the FWER of such a procedure under the null hypothesis that there is no true association in any data set? In other words, what is the probability for the observation "at least one P-value ≤ 0.05 out of 5 independent P-values", under the null hypothesis that there is no real effect in any study?

$$\Pr(\text{at least one } P \leq 0.05 \,|\, \text{NULL}) = 1 - \Pr(\text{all } P > 0.05 \,|\, \text{NULL}) = 1 - (1 - 0.05)^5 = 0.226.$$

Thus, in more often than in 1 out of 5 null variants at least one of the 5 studies reports a P-value ≤ 0.05. The correct Type I error rate in this multiple testing setting is 0.226, not 0.05 that any one study group might had thought!

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

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

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. On the other hand, it is often a very stringent method and may miss many of the truly non-zero associations.

Genome-wide significance (GWS). In GWAS literature, a significance threshold 5×10^{-8} is commonly used as a reference value for a convincing association. It is often thought to result from the Bonferroni correction applied to achieve a FWER of 0.05 for a GWAS when the GWAS tests about one million *independent* variants. The term *genome-wide significant association* is used for variants that reach this threshold. An interpretation is that, on average, 1 in 20 GWAS that uses the significance threshold 5×10^{-8} will report at least one false discovery whereas the remaining 19 out of 20 do not report any false discoveries.

In practice, 5e-8 has proven out to work well in the sense that there are few false positives that reach this significance level in GWAS. Since the concept of GWS is typically introduced as the Bonferroni corrected significance threshold for one million tests, some studies take an approach to apply their own Bonferroni corrected threshold adjusted only by the number of SNPs they happen to test in their own data, which can be much less than one million. This does not seem a scientifically sound way to think about the GWAS analysis as the next example shows.

Example 2.2. Suppose that I plan to test 10^6 genetic variants for association with heart disease. The machine in the lab breaks up and produces data only for one chromosome and I only have 10^2 variants to analyze. I do the statistics for each of 100 variants available and for one of them, variant v, I observe a P-value of 10^{-4}. Should I call v significant?

Discussion. Let's use FWER control because then we would seem to have control over how likely we are to make any false discoveries. If I had observed all 10^6 variants and had done that many tests, then FWER would say that significance threshold should be $0.05/10^6 = 5 \times 10^{-8}$, our typical GWS threshold. However, if I compute the Bonferroni corrected threshold for the observed 100 variables, it is 10,000 times higher, $0.05/100 = 5 \times 10^{-4}$. It seems that whether I label this variant as a discovery depends on whether my lab machine happened to work for other chromosomes, which seems unsatisfactory, if my goal is simply to determine whether I have enough evidence that this particular variant v is interesting. It does not seem conceptually sound that for the same observed data for variant v, my inference on whether I think it is interesting depends on how many other things have been tested "in the same experiment". Deciding a significance threshold based only on the number of tests done "simultaneously" is rarely conceptually satisfactory. It may work well in practice, but the real reason for that is likely elsewhere than in the correction for the number of simultaneous tests. We will come to that real reason later.

Which quantity would we ideally want to have (instead of the P-value)? The P-value is defined through **frequentist** properties of data generating procedure of the null hypothesis. Namely, observing a P-value of 0.023 means that under the null hypothesis the probability of getting at least as extreme data set, in terms of a particular test statistic, as the one we have observed, is 0.023.

This definition seems quite clumsy. In the end, we would just want to know what is the evidence that the null hypothesis holds for this particular variant j after we have observed GWAS data, i.e., we would want to know a probability $P(H_j \,|\, \text{Data})$. P-value does not answer this question and any conceptually sound answer would require more information than is used for computation of P-value.

Significance threshold and probability of the null hypothesis Let's approach the probability of the null hypothesis by revealing how a standard inference procedure based on a fixed significance threshold α relates to the probability of null hypothesis. This inference procedure is simply to reject null hypothesis H_j and call variable j significant if the corresponding P-value is $\leq \alpha$. Let's take as our observed data simply the event of a significant P-value: $S = \{P_j \leq \alpha\}$. Let's also define event for a true effect as $T = \{H_j \text{ does not hold}\}$ and its complement of a null effect: $N = \{H_j \text{ holds}\}$. Naturally $P(T) = 1 - P(N)$ and we are interested in $P(T \,|\, S)$, i.e., probability that there is a true effect given that we observe a significant association. Bayes rule gives

$$P(T \,|\, S) = \frac{P(T)P(S \,|\, T)}{P(S)} \text{ and}$$

$$P(N \,|\, S) = \frac{P(N)P(S \,|\, N)}{P(S)}.$$

By dividing the first equation by the second we have

$$\frac{P(T \mid S)}{P(N \mid S)} = \frac{P(T)P(S \mid T)}{P(N)P(S \mid N)}.$$

This says that the odds of there being a true effect, after we have observed a significant P-value, are the prior odds of a true effect ($P(T)/P(N)$) times the ratio of probabilities of getting a significant result under the alternative model vs. the null model. By definition, $P(S \mid N) = \alpha$, i.e., under the null we get significant results with probability α. The term $P(S \mid T)$ is called **statistical power** of the study to observe a true effect. Thus,

$$\frac{P(T \mid S)}{P(N \mid S)} = \text{prior-odds} \times \frac{\text{power}}{\text{significance threshold}}.$$

If we assume that we have a well-powered study to detect effects we are interested in, say power is above 80%, we can replace power by ≈ 1 and ignore it for now but we will come back to power later on the course. We see that whether a significant result is more likely to be a true positive than a false positive depends on the ratio of prior-odds of true effect and significance threshold. If we want our inference procedure to produce significant results only for almost certain cases of true positives, we need to choose our significance threshold small enough that it can overcome a possibly small prior odds of a true effect in GWAS. Note, however, that power will also drop when we decrease the significance threshold so we cannot ignore it forever.

Example 2.3. Suppose that we are looking for common genetic variants that increase the odds of heart disease by at least 20% compared to the other allele at that position of genome. We think that there are not many such variants around, maybe only 10 or so among the 10^7 common variants. Thus we say that our prior probability that any one variant is this kind of a risk variant is $P(T) = 10/10^7 = 10^{-6}$. What should be our significance threshold if we want to be over 95% certain that an (average) significant finding is truly a real effect? (Here "average" because this consideration does not consider the exact properties of the variant but only average properties of those variants that reach significance threshold; later on the course we will dive into more precise inference for the particular individual variant that has been observed.)

We have that

$$\text{significance threshold } \alpha = \text{power} \times \frac{\text{prior-odds}}{\text{posterior odds}} \leq \frac{\text{prior-odds}}{\text{posterior odds}}.$$

We don't know power as it depends on the significance threshold α, but we can bound power from above by 1, and hence get an upper bound for α.

```
p.T = 1e-6
prior.odds = p.T/(1-p.T)
pwr = 1 #upper bound for power --> upper bound for alpha
post.odds = 0.95/(1-0.95)
alpha = prior.odds*pwr/post.odds
paste(signif(alpha,3))
```

```
## [1] "5.26e-08"
```

This is the standard genome-wide significance threshold arrived from a different angle compared to the direct Bonferroni correction. On the other hand, there is an obvious connection between the two approaches, since above we used the number of variants in the genome to derive the prior probability of a true effect and this may coincide with the number of tests carried out in a GWAS, which then drives the Bonferroni approach. However, a large conceptual difference is that the derivation of α above is independent of the actual number of tests carried out by any one researcher/GWAS. This derivation makes it clear that the requirement for a small significance threshold, that we encounter in GWAS, is not primarily because of the number of tests carried out in any one GWAS, but because of a small prior probability that any one of our measured variant has a non-zero effect. Importantly, this derivation removes the problem in the earlier Example: the significance threshold required should not change with the number of tests done, but should

be determined by the properties of prior-odds and power of the study. In particular, the threshold does not change depending whether I analyse 100 or 10^6 variants "in the same experiment". Note, however, that if there is prior knowledge that the 100 variants are more likely to be non-zero because of their functional annotations or other properties by which they have been pre-screened from among all 10^6 variants, then I could loosen the significance threshold for them, but that is because prior-odds is now different, not because the number of tests were different.

What do we do with the GWS variants? Even if the GWS threshold of 5e-8 has proven out to be so stringent that the number of false positives is very small, the GWS variants are just starting points for further work:

- **Replication** should always be done in an independent cohort, and preferrably in several different populations, to see that the association is not due to some technical or methodological bias in one study, and to assess whether the association exists also in other populations. (slide 18)
- **Biological effect** behind a new GWAS locus may be challenging to figure out because there are often up to hundreds of correlated GWS variants in the GWAS region (slide 19), of which only a few may truly have a causal effect, and we still know only a little about the functional consequences of the DNA variation. This area of reserach is in rapid expansion thanks to the new technologies to produce data on functional genomics, which provides us with exciting times for the interpretation of GWAS results. Examples of detective work on some GWAS loci: *C4* and schizophrenia by Sekar et al. 2016 and *RNF186* and ulcerative colitis by Rivas et al. 2016.

Statement of statistical significance and P-values (slide 14) We have seen how P-values are used in the GWAS context and how we account for multiple tests in defining the significance threshold. In recent years, with a flood of new data, the issues of multiple testing and resulting false positive findings have become a problem throughout many fields of science. This has created a "replicability crisis", where far too many published results do not seem to be replicable, and partly this is due to a misinterpretation of P-values, in particular, in the amount of evidence that they provide against the null hypothesis. In 2016, The American Statistical Association published a statement about P-values which is a good read for everyone interpreting statistical analyses.

Extra: How to compute a P-value

The test statistics from the same statistical model can be derived in many ways. For models with a likelihood function, such as regression models in GWAS, there are three main approaches to derive the test statistic. In the order of increasing computational demands, they are

1. **Score test.** Based on the distribution of the score statistic (that is, the derivative of the log-likelihood function with respect to the parameters to be tested, evaluated at the null value of the parameters, typically 0). This is often simple to compute because there is no need to optimize the parameters under the alternative model as we only evaluate the score statistic under the null. A downside is that this is very inaccurate when the likelihood is not close to the shape of the Normal density. A Normal approximation is not good when we are testing low-frequency variants (MAF $< 1\%$), or we have an unbalanced case-control data (proportion of cases $< 20\%$) or the sample size is small. In those cases, the other two tests must be used, and unfortunately this fact seems not to be very well known/appreciated across the GWAS community.
2. **Wald's test.** This is the default method to compute P-values in the output of R's `summary()` function. It is based on the t-statistic, i.e., the ratio $\widehat{\beta}/\text{SE}$, that under the null has an approximate standard normal distribution. Wald's test requires finding the maximum likelihood estimate $\widehat{\beta}$ and its standard error, which in turn requires computing the second derivative of the log-likelihood at the maximum.

3. **Likelihood ratio test.** This is typically the most accurate way to compute a P-value out of the three methods mentioned here. It is based on the likelihood ratio test statistic (LRT) $2L(\widehat{\beta}; \text{DATA}) - 2L(\beta_0; \text{DATA})$ where β_0 contains the fixed values of β under the null and L is the log-likelihood function. LRT follows approximately χ^2_k under the null, where k is the dimension of the tested parameter vector.