HDS 4. Bayesian inference

Matti Pirinen, University of Helsinki

9.11.2021

Statistical threshold measures P-value and Q-value are defined through **frequentist** properties of certain procedures. Namely,

- P-value of 0.05 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 0.05.
- Q-value of 0.05 means that if we label this particular variable with Q-value of 0.05 as significant (i.e. as a discovery), then the expected proportion of false discoveries among all the variables with Q-values $\leq$ 0.05 is 0.05.

Two questions that come up:

1. These definitions seem quite clumsy. In the end, we just want to know what is the evidence that the null hypothesis holds for this particular variable j. For example, we want to know a probability $\Pr(H_j \,|\, \text{Data})$. Neither P-value, nor Q-value answers that key question. Obviously, this question requires more information than is used for computation of P-value or Q-value, and we will look soon what exactly is needed.
2. What if I want to do inference on variable j and do not think that another variable k, that I just happened to have observed at the same experiment, should affect my inference on H_j in any way? Both P-value-based FWER control as well as FDR control rely on the concept of a set of hypotheses for which some error rates can be controlled after considering them as a single entity.

Example 4.1. Suppose that I plan to download data from the UK Biobank on 500,000 individuals and their 10M genetic variants to test the association with diabetes status that is also available in the biobank. However, due to technical issues, I am able to download only one small genotype file while the others are unavailable and hence I have data only for 1000 genetic variants instead of 10M. I do the regression for each of 1000 variants available and for one of them, variant v, I observe a P-value of 10^{-5}. Should I call v significant?

Discussion. Let's use FWER control at 5% level as then we would seem to have a strong control over how likely we are to make any false discoveries. If I had observed all 10^7 variants and had done that many tests, then FWER would say that significance threshold should be $0.05/10^7 = 5 \times 10^{-9}$. However, if I compute the threshold for the observed 1000 variables, it is 10,000 times higher, $0.05/1000 = 5 \times 10^{-5}$. It seems that whether I label this variant v as a discovery depends on whether I was successful in downloading the other files. This seems unsatisfactory, when my interest is simply to determine how much evidence do I have that this particular variant v is null. It does not seem conceptually sound that for the same observed data for variable v, my inference on whether I think it is interesting depends on how many other things have been tested "in the same experiment". Deciding significance threshold based only on the number of tests done is rarely conceptually satisfactory from scientific point of view. Similarly, Q-value is not a complete quantity to measure the status of any *particular one* variable, because Q-value of a particular variable may change when more/less/other variables are included in the same study. (This being said, Q-value uses the data on the other variables to learn about common properties of all variables, e.g., in estimating π_0, and hence

Q-value is, in many settings, an improvement over using the data on the other variables only through their total count, as is done in Bonferroni correction.)

A fix to this conceptual problem can be formulated by Bayesian statistics and we will use Bayesian terminology below. Same ideas go by name "**local false discovery rates**" in literature to emphasize that the focus is now put on the particular test and not anymore on the whole tail of "at least as extreme data sets", as it was with P-values, FWERs and FDRs. Later we will have a look at how local false discovery rate has been implemented in `qvalue` package.

Notation Throughout these notes we use $\Pr(\cdot)$ as a function name that can mean either

- probability of an event, e.g., $\Pr(H_0)$ is probability of the null hypothesis.
- probability density of a continuous variable, e.g., $\Pr(X = x \,|\, H_0) = \Pr(x \,|\, H_0)$ is the value of the probability density function of random variable X under hypothesis H_0, evaluated at point x.

Probability and Bayes rule We denote probability of event A as $\Pr(A)$ and conditional probability of event A given that event B has occurred by $\Pr(A \,|\, B)$ which can be represented in terms of ratio of probabilities of two events as $\Pr(A \cap B)/\Pr(B)$. We also denote more simply $\Pr(A, B) = \Pr(A \cap B)$.

Bayes rule can be derived by writing the joint probability $\Pr(A, B)$ using expansions through conditional probabilities in both ways possible:

$$\Pr(B)\Pr(A|B) = \Pr(A, B) = \Pr(A)\Pr(B|A) \implies \Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}.$$

Bayes rule tells how observing event B updates probability of A by a multiplicative factor $\Pr(B|A)/\Pr(B)$, hence its central role when quantifying how we learn from observations. Marginal probability of $\Pr(A)$ is the probability of event A when we have not observed any information about event B (or any other known events). In Bayes rule, where event B also appears, the interpetation of $\Pr(A)$ is the **prior** probability of event A, i.e., probability of A *before* we have learned about B. (And similar interpretation applies to $\Pr(B)$ by changing the roles of A and B.) $\Pr(A|B)$ is called the **posterior** probability of A, i.e., probability of A after we have observed event B. Bayes rule tells exactly how the observations made determine the transition from a prior probability to a posterior probability.

Example 4.2. Suppose we have a medical test for a disease D which has sensitivity of $\alpha = 0.99$ (that is, gives a positive result in 99% of true cases) and specificity of $\beta = 0.99$ (that is, gives a false positive result in 1% of healthy individuals). What is the probability that individual who in a population screening tests positive (+) truly has the disease, when the prevalence K of the disease is (a) 10% or (b) 0.1%?

Answer 4.2.

- $\Pr(+ \,|\, D) = \alpha$ (sensitivity is α)
- $\Pr(+ \,|\, \text{no}D) = 1 - \beta$ (specificity is β)
- $\Pr(D) = K$ (prevalence is K, this is a prior probability of having D *before* accessing the test result)
- $\Pr(+) = \Pr(+ \,|\, D)\Pr(D) + \Pr(+ \,|\, \text{no}D)\Pr(\text{no}D) = \alpha K + (1 - \beta)(1 - K)$

Bayes rule says that the posterior probability of D (called **positive predictive value** (PPV) in epidemiology) is

$$\Pr(D \,|\, +) = \Pr(D)\frac{\Pr(+ \,|\, D)}{\Pr(+)} = \frac{\alpha p}{\alpha p + (1 - \beta)(1 - p)}.$$

When a test is 99% sensitive and 99% specific, PPV is, depending on whether the prevalence is 10% or 0.1%,

```
a = 0.99
b = 0.99
K = c(0.1, 0.001)
cbind(preval = K, PPV = a*K/(a*K+(1-b)*(1-K)))
```

```
##      preval        PPV
## [1,]  0.100 0.91666667
## [2,]  0.001 0.09016393
```

How come that only 9% of positives have the disease (in the low prevalence setting) even though the test captures well (99%) both the true positives and the true negatives? Let's consider 10,000 individuals. Out of them, only 10 have D. Assume all those 10 test positive due to high sensitivity of the test. The remaining 9990 do not have D. But out of them 1%, i.e., ~100 still test positive. Thus, out of all 110 positive tests, only 10 (~9%) were true positives and 100 (~91%) were false positives.

Using terminology from the previous lectures, we may formulate this result as saying that the FDR of this screening procedure is about 91% when prevalence is 0.1% and about 8.4% when prevalence is 10%. In particular, the practical usefulness of the test strongly depends on the population screened, and Bayes formula is the way to determine how.

Example 4.3. Let's apply Bayesian inference to the parameter θ that represents the probability of a coin landing heads up in a coin toss. We start by defining our prior probability distribution on θ. Suppose that Uniform(0,1) (which is also the Beta(1,1) distribution) is a description of our prior beliefs.

Then we start tossing the coin and report value $H(i)$ that is the number of heads in first i tosses. Our sampling model for $H(i)$ is

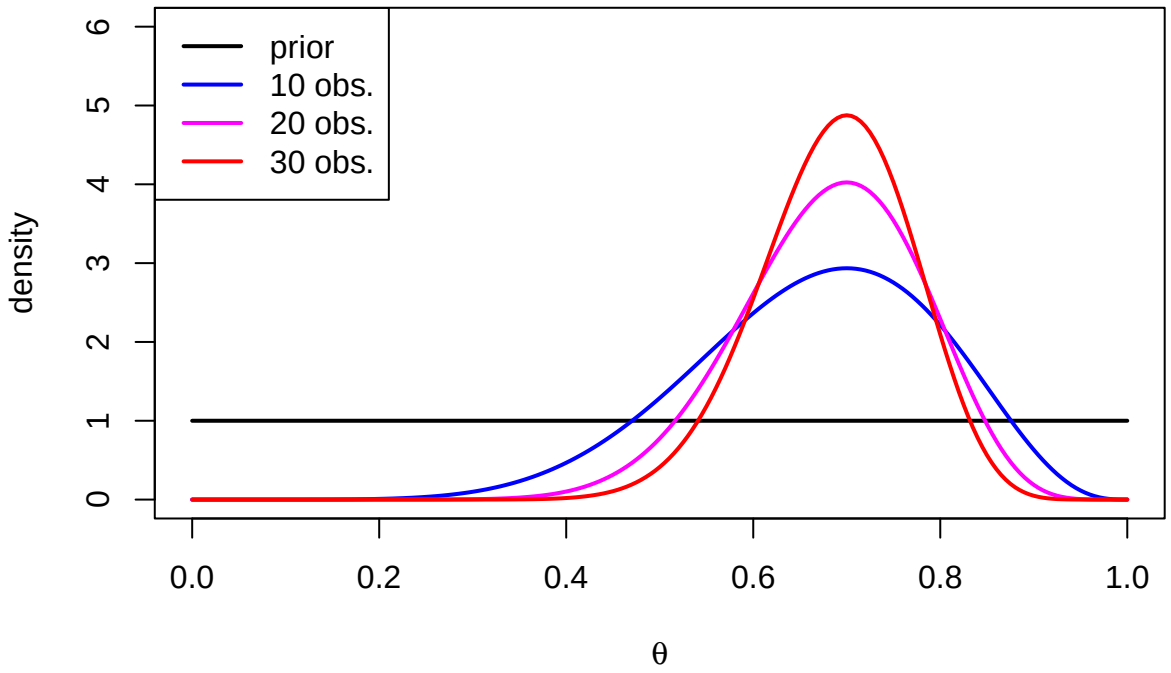
$$H(i) \,|\, \theta \sim \text{Binomial}(i, \theta).$$

It follows (from a course on Bayesian inference) that the posterior distribution for θ is

$$\theta \,|\, H(i) \sim \text{Beta}(H(i) + 1, i - H(i) + 1).$$

Let's draw the prior and the posterior distributions via their density functions after the following series of observations: $H(10) = 7$, $H(20) = 14$ and $H(30) = 21$.

```
n = c(0, 10, 20, 30)
H = c(0, 7, 14, 21)
cols = c("black", "blue", "magenta", "red")
x = seq(0, 1, 0.005)
plot(NULL, xlim = c(0,1), ylim = c(0,6), xlab = expression(theta), ylab = "density")
for(ii in 1:length(n)){
  y = dbeta(x, H[ii]+1, n[ii]-H[ii]+1)
  lines(x, y, col = cols[ii], lwd = 2)
}
legend("topleft", leg = c("prior","10 obs.","20 obs.","30 obs."), col = cols, lwd = 2)
```



We see that the point estimate from data $H(i)/i = 0.7$ at all three observations and that the posterior distribution narrows down around that value as the sample size increases. There is little numerical difference to pure maximum likelihood inference about θ when there are tens of tosses but conceptually there is a big difference when we consider, e.g., Bayesian posterior probability intervals (credible intervals) vs. traditional confidence intervals, as studied in courses on Bayesian inference.

Significance threshold and probability of the null hypothesis Let's next see how a standard inference procedure based on a fixed significance level 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 the event of a true non-null effect as $T = \{H_j \text{ does not hold}\}$ and its complement of a null effect: $N = \{H_j \text{ holds}\}$. Naturally, $\Pr(T) = 1 - \Pr(N)$ and we are interested in $\Pr(T \mid S)$. Bayes rule gives

$$\Pr(T \mid S) = \frac{\Pr(T)\Pr(S \mid T)}{\Pr(S)} \text{ and}$$

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

By dividing the first equation by the second we have

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

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

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

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. 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 possible small prior odds of a true effect in high-dimensional problems. Note however, that power will also drop when we decrease the significance threshold so we cannot ignore it forever.

Example 4.4. Suppose that we are looking for genetic mutations that at least double the risk of diabetes compared to normal variant at that position of genome. We have a large sample size so that we are very well powered to find such variants. We think that there are not many such mutations around, maybe only 10 or so in the genome that has 10^7 variants. Thus we say that our prior probability that any one variant increases risk is $P(T) = 10/10^7 = 10^{-6}$. What should our significance threshold be if we want to be over 95% certain that a significant finding is truly a real effect?

```
p.T = 1e-6
prior.odds = p.T / (1 - p.T)
pwr = 1.0 #assume full power
post.odds = 0.95 / (1 - 0.95)
alpha = prior.odds*pwr/post.odds
paste(signif(alpha,3))
```

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

Above we used the size of the genome to derive a prior probability of a true effect. Even though this may coincide with the number of tests carried out in the actual analysis, a big conceptual difference is that the derivation of α above is independent of the actual number of tests carried out. This derivation makes clear that the requirement of a small significance threshold, that we often encounter in high-dimensional problems, is not primarily because of the number of tests carried out, but because of a small prior probability that any one of our measured variables is a true effect. Importantly, this derivation removes the problem in the earlier Example 4.1: the significance threshold required should not change with the number of tests done, but should be determined by the prior-odds and power of the study. In particular, the threshold does not change depending on whether I analyse 1000 or 10^7 variants "in the same experiment". (Note, however, that if there is prior knowledge that the 1000 variants are more likely to be non-zero because of their properties by which they have been chosen among all 10^7 variants, then I can loosen the significance threshold for them, but that is because the prior-odds are now different, not because the number of tests is different.)

Questions.

1. In which cases we could use different prior odds for different variables?
2. What would smaller/higher prior odds mean in terms of significance level required for a fixed posterior odds given that everything else remains constant?
3. What is the smallest possible value for "power" in formula above, and what does Bayes formula tell about learning from data if we have a setting with that minimum possible power?

From significance to the observed data The methods that we have considered so far only determine whether tests are "significant" or not. We could and should also make more efficient use of the observed data than just labelling things significant or not significant. Or would you think that variables with P-values 0.04 and 10^{-10} should have the same posterior probability of being non-null?

Let's now apply Bayes rule to the null hypothesis testing problem in a way where we condition on the full observed data $\mathcal{D}$ and not just on whether P-value is below some threshold as we did above. Let's mark the null hypothesis by H_0 and the alternative hypothesis by H_1.

$$\Pr(H_0 \,|\, \mathcal{D}) = \frac{\Pr(H_0)\Pr(\mathcal{D} \,|\, H_0)}{\Pr(\mathcal{D})} = \frac{\Pr(H_0)\Pr(\mathcal{D} \,|\, H_0)}{\Pr(\mathcal{D} \,|\, H_0)\Pr(H_0) + \Pr(\mathcal{D} \,|\, H_1)\Pr(H_1)}$$

We need to specify probabilistic models for the observed data under both hypothesis. After these models are specified, the infrence is about letting the possible models compete both in how well they explain the observed data (terms $\Pr(\mathcal{D} \mid H_1)$ and $\Pr(\mathcal{D} \mid H_0)$) and in how probable they are *a priori* (terms $\Pr(H_1)$ and $\Pr(H_0)$).

Example 4.5. The Bayesian inference shows that both the observed data AND the prior knowledge is crucial for complete inference. Suppose, for example, that one morning you wake up and it is completely dark outside. This darkness could results either from H_1: "Sun has disappeared" or from H_0: "You have woken up > 1 hour earlier than usually", and under both models $\Pr(\mathcal{D} \mid H_i)$ is similarly very high. So both models are consistent with the observations and P-values computed under either of the models as null hypothesis would not show inconsistencies between the observed data and either of the models. However, the prior odds of $\Pr(H_0)/\Pr(H_1)$ is extremely large and hence your posterior conclusion would be that you are extremely more likely to have woken up early than woken up to the world without the sun.

In a simple linear regression model the observed data contain the outcome vector $\boldsymbol{y}$ and the tested variable $\boldsymbol{x}_j$, $\mathcal{D}_j = (\boldsymbol{y}, \boldsymbol{x}_j)$, and when we assume Gaussian errors, the model for a fixed value of regression coefficient β and error variance σ^2 is

$$\Pr\left(\mathcal{D} \mid \beta, \sigma^2\right) = \mathcal{N}(\boldsymbol{y} - \boldsymbol{x}_j\beta;\, 0,\, \sigma^2 I) \propto \exp\left(-(\boldsymbol{y} - \boldsymbol{x}_j\beta)^T(\boldsymbol{y} - \boldsymbol{x}_j\beta)/(2\sigma^2)\right)$$

Under the null model, we set $\beta = 0$ and in the alternative model, we can set β to some other value b_1. If we do not want to specify our model of true effects by a single value b_1, we can use the Bayesian prior distribution for β, for example, by saying that $\beta \sim \mathcal{N}(b_1, \tau_1^2)$. With this prior, the probability density of data under H_1 is given by weighting the above likelihood by prior probability density of each possible value of β:

$$\Pr(\mathcal{D} \mid H_1) = \int_\beta \Pr\left(\mathcal{D} \mid \beta, \sigma^2\right) \Pr(\beta \mid H_1) d\beta = \int_\beta \mathcal{N}\left(\boldsymbol{y} - \boldsymbol{x}_j\beta;\, 0,\, \sigma^2\right) \mathcal{N}\left(\beta;\, b_1,\, \tau_1^2\right) d\beta.$$

(In both models we typically fix σ^2 to its empirical maximum likelihood estimate as the competing regression models do not typically differ in σ^2, and hence we are less interested in it than in β.)

If we assume that in the Gaussian prior of β the mean parameter $b_1 = 0$, then the integral can be done analytically to give

$$\Pr(\mathcal{D} \mid H_1) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \mathrm{SE}^2\right),$$

where c is a constant and $\widehat{\beta}$ is the MLE of β and SE the corresponding standard error. Note that by replacing τ_1 with 0, we have

$$\Pr(\mathcal{D} \mid H_0) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \mathrm{SE}^2\right).$$

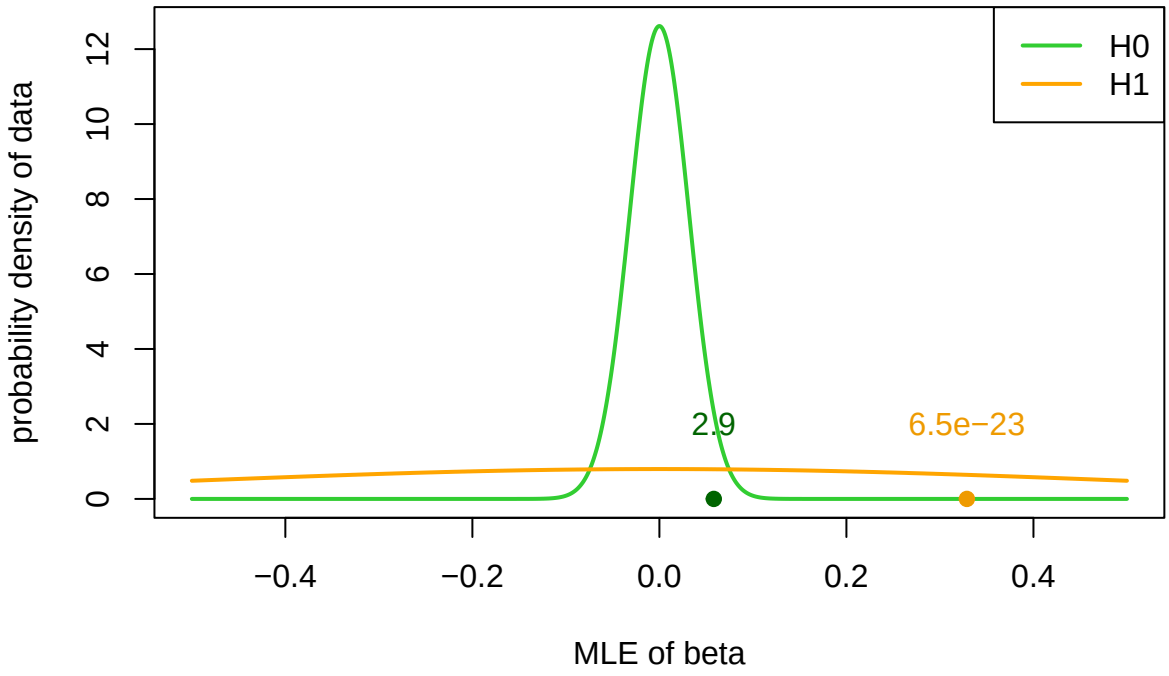
These results tell that we can quantify how well each model explains the data, by asking how well each model can explain the MLE $\widehat{\beta}$. Let's demonstrate this by (1) plotting probability densities of data under both models H_0 (green) and H_1 (orange) and by simulating one data set from each model and by showing where the parameter estimate from each data set fall.

```
set.seed(16102017)
n = 1000
#sample size for SE calculation
sigma = var.x = 1
se = sigma/sqrt(n*var.x) #see HDS 0 notes for SE in linear model
tau = 0.5 #prior standard deviation for H1

#Let's draw probability densities of "data" under the two models, H0 and H1
#as a function of MLE estimate x
x = seq(-0.5, 0.5, by = 0.001)
y1 = dnorm(x, 0, sqrt(tau^2 + se^2) )
y0 = dnorm(x, 0, se)
```

```r
plot(x, y0, t = "l", col = "limegreen", lwd = 2,
     xlab = "MLE of beta", ylab = "probability density of data")
lines(x, y1, col = "orange", lwd = 2)
legend("topright", legend = c("H0","H1"), col = c("limegreen","orange"), lwd = 2)

#We make a shortcut and don't simulate data at all, but we simulate MLEs
#Suppose we have two cases, first is null, second is alternative (beta=0.3)
b =c(0, 0.3)
b.mle = rnorm(2, b, se) #these are simulated MLE estimates
points(b.mle, c(0,0), pch = 19, col = c("darkgreen","orange2"))
bf.01 = dnorm(b.mle, 0, se) / dnorm(b.mle, 0, sqrt(tau^2+se^2)) #Bayes factor between H0 and H1
text(b.mle[1], 2, signif(bf.01[1],2), col = "darkgreen")
text(b.mle[2], 2, signif(bf.01[2],2), col = "orange2")
```



If $\widehat{\beta}$ is close to 0, then H_0 explains the data better than H_1, whereas the opposite is true when $\widehat{\beta}$ is farther away from 0. With these parameters H_1 starts to dominate about when $|\widehat{\beta}| \geq 0.1$.

Two points shown in plot are examples of possible maximum-likelihood estimates that could result either under H_0 (green) or H_1 (orange). The values shown are ratios of $\Pr(\mathcal{D} \mid H_0)/\Pr(\mathcal{D} \mid H_1)$ computed at these two points. When this ratio > 1, the null model H_0 explains the data better than the alternative H_1 and when it is < 1 the opposite is true. This ratio is called **Bayes factor (BF)** and it is a multiplicative factor that multiplies the prior odds to result in the posterior odds:

$$\frac{\Pr(H_0 \mid \mathcal{D})}{\Pr(H_1 \mid \mathcal{D})} = \frac{\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D} \mid H_1)} \frac{\Pr(H_0)}{\Pr(H_1)}$$

We are almost there having calculated a proper probability for the null hypothesis. We still need to agree on the prior probability of the null model. For example, by continuing with our genomics application, we had $\Pr(H_0) = 0.999999$ because we expected that there was only a very small probability (about 10^{-6}) that there was a true effect for any one variant. So posterior odds and posterior probabilities for the null model are:

```
post.odds = bf.01*(1 - 1e-6) / 1e-6
post.prob = post.odds / (1 + post.odds)
cbind(data = c("green","orange"), post.H0 = post.prob, post.odds = post.odds)
```

```
##      data     post.H0                post.odds
## [1,] "green"  "0.999999659543705"    "2937233.5737576"
## [2,] "orange" "6.49242701456545e-17" "6.49242701456545e-17"
```

For illustration, let's check the P-values corresponding to these two data sets:

```
pchisq( (b.mle / se)^2, df = 1, lower = F)
```

```
## [1] 6.583224e-02 2.512191e-25
```

So P-value of the first one is quite close to 0.05, but still the probabilistic analysis says that the observed data have even made the null model more likely than it was prior to seeing the data. This is an example how P-value does not compare probability of data under the null model to any other model, it simply measures how likely the data are to raise under the null model. In principle, data can be quite unlikely under model H_0, but if it is even more unlikely under model H_1, then we do not yet have evidence to prefer the alternative model over the null model.

Note that there were several assumptions made in the Bayesian anaysis about the effect sizes under H_1 and also on the prior probabilities of the models, and the posterior probabilities will change when these assumptions are changed. Therefore, P-values and Q-values remain useful simple summaries of data that can be computed easily and with little additional assumptions. The important thing is to know what P-values and Q-values are and what they are not (they are not probabilities of null hypothesis!), and that what kinds of additional pieces of information would be needed in order to do more complete probabilistic inference.

Questions.

1. What is the main conceptual benefit of having posterior probability of null hypothesis compared to having a P-value under the null hypothesis?
2. What are the main practical complications of computing posterior probabilities compared to computing P-values?

The role of marginal likelihood in Bayesian inference (Adapted from Rasmussen & Williams.)

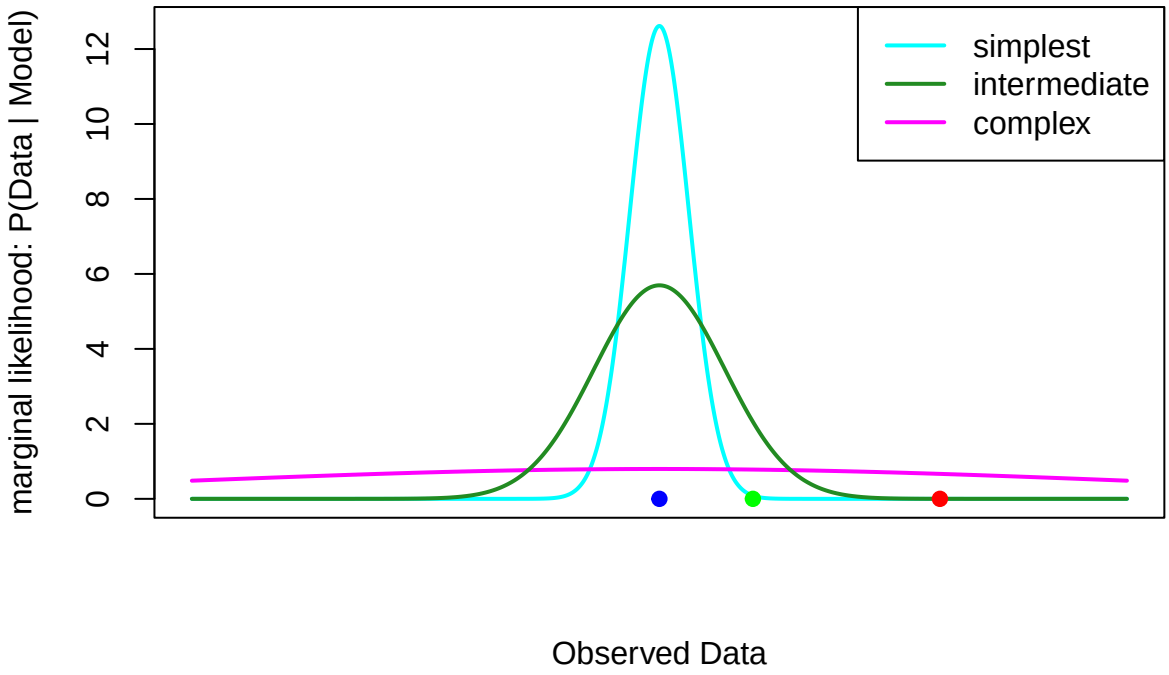
In Bayesian analysis, the probability density of the data given the model, $\Pr(\mathcal{D} \,|\, \mathcal{M})$, describes how well a model (or "hypothesis") describes the data. This term is called **marginal likelihood**. It is called marginal because the parameters of the model that appear in the likelihood function do not appear in the marginal likelihood. Typically, marginal likelihood is computed as an integral over the parameter space of the product of the likelihood function $\Pr(\mathcal{D} \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of parameters of the model, and the prior distribution of the parameters $\Pr(\boldsymbol{\theta} \,|\, \mathcal{M})$ specified by the model $\mathcal{M}$.

$$\Pr(\mathcal{D} \,|\, \mathcal{M}) = \int \Pr(\mathcal{D} \,|\, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \,|\, \mathcal{M}) d\boldsymbol{\theta} :$$

For example, the null model may say that the slope parameter is exactly 0 whereas the alternative model may say that the prior on slope is $\mathcal{N}(0, 1)$, and in both cases the likelihood function could be the likelihood from the linear regression model with Gaussian errors.

Above we saw marginal likelihood values for two models H_0 and H_1 as a function of observed data (that were summarized by the estimate $\widehat{\beta}$ and its SE). In Figure below we see an illustration of the behaviour of the

marginal likelihood for three different models with differing levels of complexity. The horizontal axis is an idealized representation of all possible outcome vectors y that we could observe, and the vertical axis plots the marginal likelihood. A simple model can only account for a limited range of possible sets of outcome values, but since the marginal likelihood is a probability distribution over y it must normalize to unity, and therefore the data sets which the model does account for have a large value of the marginal likelihood. Conversely for a complex model: it is capable of accounting for a wider range of data sets, and consequently the marginal likelihood doesn't attain equally large values for any given data set that the simple model explains well. For example, the simple model could be a linear model, and the complex model a large neural network. The figure illustrates why the marginal likelihood doesn't simply favor the models that are most complex in cases where the simple model already can explain the data.



It is a property of the marginal likelihood that it automatically incorporates a trade-off between model fit and model complexity. That is, it has an inherent guard against overfitting. This is the reason why the marginal likelihood is valuable in solving the model selection problem.

Note the difference between maximum lilkeilhood, where parameter values are estimated by optimizing the likelihood, and marginal likelihood, where the parameter values are integrated out from the model to reveal the descriptive power of the model in explaining the observed data set. Pure maximum likelihood method always tends to favor more complex models because more complex models can be made fit better to any observed data. However, more complex models also spread their explanatory power over much larger sets of possible data sets than simpler models, and therefore their marginal likelihood for a certain data set can be much lower than that of a simpler model.

In above Figure, which model would be the best explanation for each of the three data sets (blue, green and red points)?

Example 4.6. Let's consider two series of coin tosses of length n, where the number of observed heads are h_1 and h_2. We want to infer whether the two series are conducted with a similar coin. Let's apply Bayesian inference through marginal likelihoods and let's demonstrate the diffrence to maximized likelihood.

Denote by θ_1 and θ_2 the proportion of heads for coins 1 and 2. The likelihood function for the whole experiment is

$$\Pr(h_1, h_2 \,|\, \theta_1, \theta_2) = \prod_{i=1}^{2} \binom{n}{h_i} \theta_i^{h_i} (1-\theta_i)^{n-h_i}$$

Let's define two models. Model H_0 states that there is only one parameter $\theta_1 = \theta_2$ and in the Bayesian version that parameter has Uniform(0,1) prior distribution. Model H_1 says that parameters θ_1 and θ_2 are not necessarily the same and they are given Uniform(0,1) prior each independently. We can compute marginal likelihood for both models using Beta function (technically this is an integration problem and the Beta function is defined as a shorthand to exactly these types of integrals):

$$\Pr(h_1, h_2 \,|\, H_0) = \int_0^1 \binom{n}{h_1}\binom{n}{h_2}\theta^{h_1+h_2}(1-\theta)^{2n-h_1-h_2}d\theta = \binom{n}{h_1}\binom{n}{h_2}B(h_1+h_2+1, 2n-h_1-h_2+1)$$

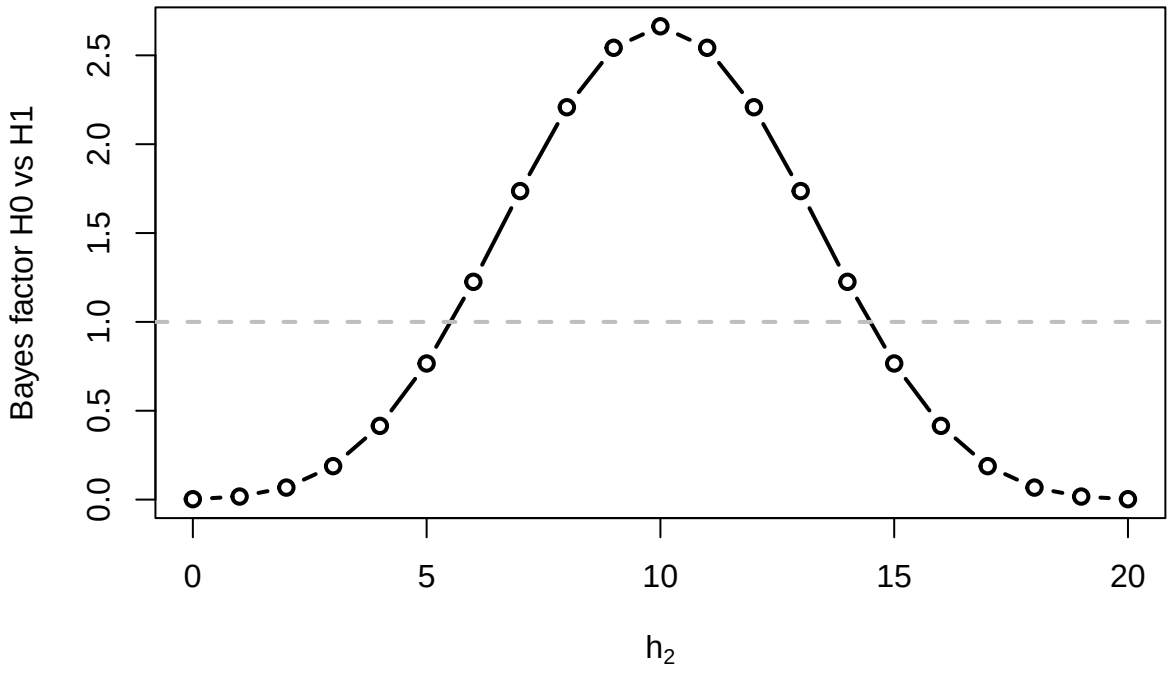
$$\Pr(h_1, h_2 \,|\, H_1) = \int_0^1 \prod_{i=1}^{2}\binom{n}{h_i}\theta_i^{h_i}(1-\theta_i)^{n-h_i}d\theta_i = \prod_{i=1}^{2}\binom{n}{h_i}B(h_i+1, n-h_i+1)$$

Thus the Bayes factor of model 0 against model 1 is

$$\mathrm{BF}_{01} = \frac{\Pr(h_1, h_2 \,|\, H_0)}{\Pr(h_1, h_2 \,|\, H_1)} = \frac{B(h_1+h_2+1, 2n-h_1-h_2+1)}{B(h_1+1, n-h_1+1)B(h_2+1, n-h_2+1)}$$

Let's draw picture of BF when $n = 20$ and $h_1 = 10$ as $h_2 = 0, \ldots, 20$.

```
n = 20
h1 = 10
h2 = 0:n
bf = beta(h1 + h2 + 1, 2*n - h1 - h2 + 1) / beta(h1 + 1, n - h1 + 1) / beta(h2 + 1, n - h2 + 1)
plot(h2, bf, t = "b", lwd = 2, xlab = expression(h[2]), ylab ="Bayes factor H0 vs H1")
abline(h = 1, lty = 2, col = "gray", lwd=2)
```



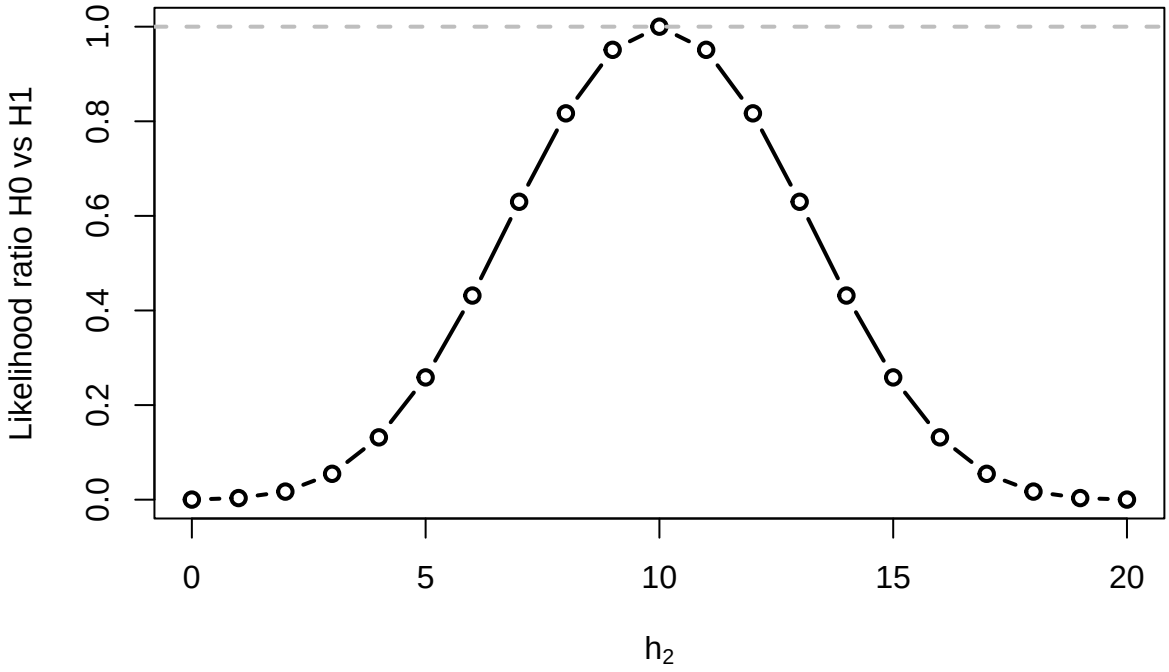
We see that the marginal likelihood favors the null model when h_2 is in range [6,14] that includes the observed $h_1 = 10$. There, according to marginal likelihood, the simpler model with a single parameter is better in explaining the observed data than the more complex model with two parameters. When h_2 is farther away from h_1, then the marginal likelihood says that we should rather use the two parameter model to explain the data.

Why is this not yet a complete Bayesian analysis? We haven't considered the prior probabilities of H_0 and H_1. If we assume them equal, then the Bayes factor is also the posterior odds. But if we have strong prior

beliefs that the coins are / aren't similar, then the inference will be affected by the prior odds $\Pr(H_0)/\Pr(H_1)$ as well.

What about maximized likelihood of the two models? MLE is $\widehat{\theta}_1 = \widehat{\theta}_2 = \frac{h_1+h_2}{2n}$ for H_0 and $(\widehat{\theta}_1 = \frac{h_1}{n}, \widehat{\theta}_2 = \frac{h_2}{n})$ for H_1. Let's plot the ratio of maximized likelihoods under the same assumptions as we did above for the ratio of marginal likelihoods.

```
lratio = ((h1 + h2)/(2*n))^(h1 + h2) * (1 - (h1 + h2)/(2*n))^(2*n - h1 - h2) /
((h1/n)^h1 * (1 - h1/n)^(n - h1)) / ((h2/n)^h2 * (1 - h2/n)^(n - h2))
plot(h2, lratio, t = "b", lwd = 2, xlab = expression(h[2]), ylab ="Likelihood ratio H0 vs H1")
abline(h = 1, lty = 2, col = "gray", lwd = 2)
```



We see that the maximized likelihood of the more complex model H_1 is always larger than the maximized likelihood of the simpler model H_0 (except in one point where both are equal). Thus the ratio of maximized likelihoods does not have similar automatic adjustment for model complexity as the ratio of marginal likelihoods. In high-dimensional models, the automatic adjustment for model complexity is a valuable property of Bayesian inference as it helps to avoid overfitting.

Local False Discovery Rate in `qvalue` package Let's see how these Bayesian ideas have been implemented in the `lfdr` method of `qvalue` package. The idea is to compute, for each tested hypothesis H_j, $\text{lfdr}_j = \Pr(H_j \text{ holds} \,|\, \text{all P-values})$, i.e., the probability of the null hypothesis given the distribution of all P-values. Note that this is different from Q_j that estimates FDR among all tests with smaller or equal Q-values, and does not talk specifically about the probability of hypothesis j. lfdr_j can be seen as a posterior probability of the null hypothesis H_j given the distribution of P-values.

The method assumes that the null P-values follow a Uniform(0,1) distribution and estimates the proportion of true null hypotheses $\widehat{\pi}_0$ as we studied earlier in lecture notes HDS3. Then the method forms an estimate of the marginal density of the observed P-values, $\widehat{f}(\cdot)$. Because $f(P) = \pi_0 \cdot 1 + (1-\pi_0)g(P)$ where g is the density of non-null P-values, it follows that

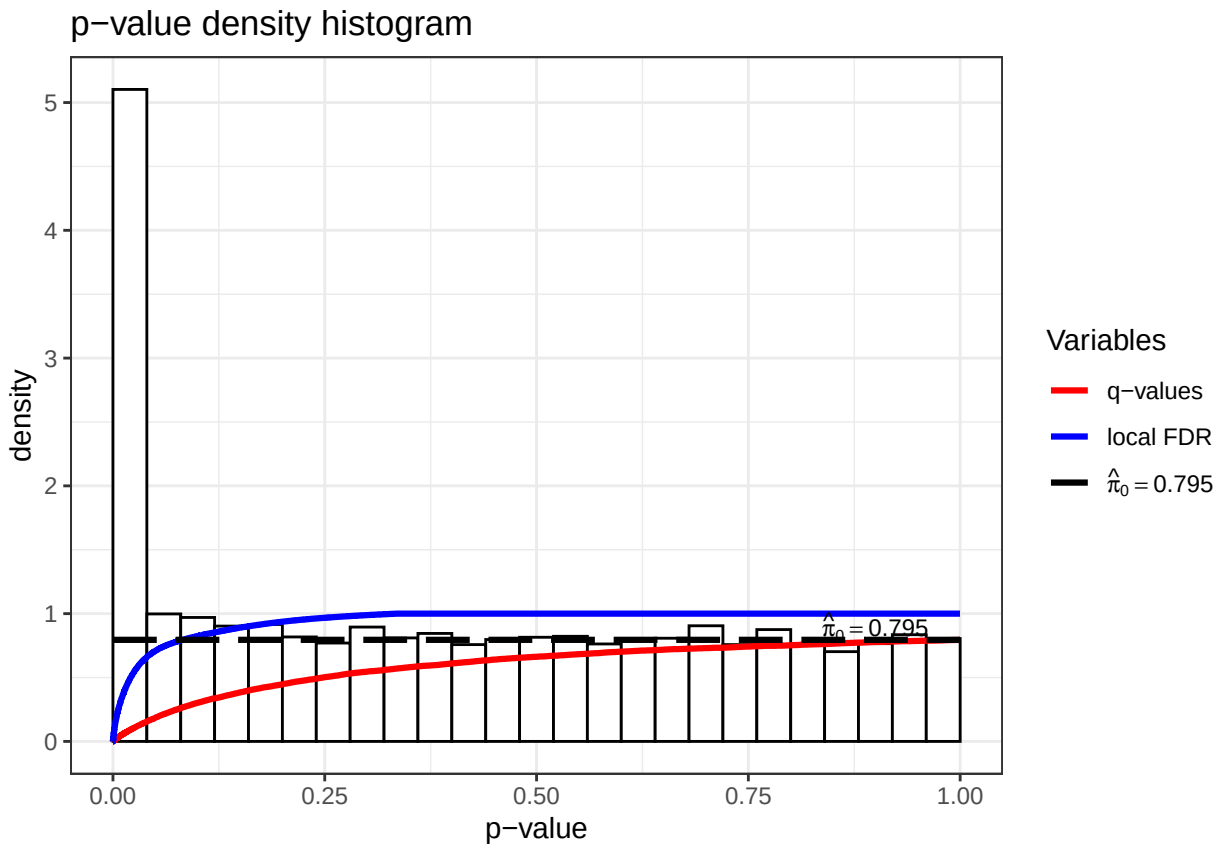
$$\text{lfdr}_j = \Pr(H_j \,|\, P_j, \widehat{f}, \widehat{\pi}_0) = \frac{\Pr(H_j \,|\, \widehat{f}, \widehat{\pi}_0) \cdot \Pr(P_j \,|\, H_j, \widehat{f}, \widehat{\pi}_0)}{\Pr(P_j \,|\, \widehat{f}, \widehat{\pi}_0)} = \frac{\widehat{\pi}_0 \cdot 1}{\widehat{f}(P_j)} = \frac{\widehat{\pi}_0}{\widehat{f}(P_j)}.$$

Since this posterior probability is based on empirical estimates of $\widehat{f}$ and $\widehat{\pi}_0$, it is called an **empirical Bayes** method.

Intuitively, lfdr_j can be interpreted as comparing the number of null P-values near the observed P-value P_j (that is, $p\,\widehat{\pi}_0\,dP$) to the number of all P-values near P_j, (that is, $p\,\widehat{f}(P_j)\,dP$), where dP is a small interval around P-value P_j. The ratio of these two numbers estimates a probability that a P-value near P_j is null.

Example 4.7. Let's repeat the P-value distribution from lecture 3 where $m = 2000$ of P-values came from Beta(0.1,4.9) and $p_0 = 8000$ from the null distribution. Let's also compute lfdr for every P-value and show the histogram of P-values and lfdr_j values as implemented with `hist()` applied to output from `qvalue()`.

```
set.seed(566)
p = 10000
m = 2000
beta.1 = 0.1 # weight for unit interval's end point 1
beta.0 = 4.9 # weight for unit interval's end point 0
null.pval = runif(p - m, 0, 1)
alt.pval = rbeta(m, beta.1, beta.0)#rbeta(m, beta.1, beta.0) #non-null = alternative distribution
pval = c(alt.pval, null.pval) #all P-values together
library(qvalue)
hist(qvalue(pval))
```



At every observed P-value p_j, lfdr_j is the ratio of value $\widehat{\pi}_0$ to the density from the histogram and varies from near 0 to near 1 from the smallest P-values to the largest P-values.

Example 4.8. Suppose that we are studying relationship between genetic variants and heart disease. We already know from previous studies 10 variants that have strong effects on the disease and therefore they also have very low P-values (let's say 10^{-10}) in our new data. We have tested 100 additional variants to

determine whether they are also associated with the disease. Suppose that all new variants actually are null. Let's see how Q-values and lfdr values differ in their robustness to inclusion of the previously known variants among the tested variants.

```
set.seed(11)
p0 = 100
pval0 = sort(runif(p0)) # null P-vals from smallest to largest
pval.all = c(pval0, rep(1e-10, 10)) #add 10 known variants at the end of vector
pval.all[1:5]
```

```
## [1] 0.0005183129 0.0137805955 0.0140479084 0.0143792452 0.0162145779
```

```
Qval0 = qvalue(pval0)
Qval.all = qvalue(pval.all)
rbind(Qval0$qvalues[1:5], Qval.all$qvalues[1:5]) # Q-values
```

```
##             [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] 0.015487016 0.09689723 0.09689723 0.09689723 0.09689723
## [2,] 0.001407911 0.03068907 0.03068907 0.03068907 0.03229908
```

```
rbind(Qval0$lfdr[1:5], Qval.all$lfdr[1:5]) # lfdr values
```

```
##            [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 0.05114749 0.1812101 0.1822572 0.1835420 0.1903847
## [2,] 0.05122124 0.1799140 0.1809315 0.1821767 0.1887580
```

We see that Q-values of the smallest (null) P-values change quite a lot dependening on whether the 10 previously known strong effects were included in the analysis. lfdr values change much less. This is reflecting the fact that Q-value talks about ALL P-values smaller than the focal P-value, whereas lfdr talks about the status of the focal P-value.

There is nothing wrong with Q-values here. They work as expected here, i.e., they give smaller Q-values for the null variables with the smallest P-values after there are 10 strong true effects included because inclusion of these 10 true effects will allow us to also do more false discoveries for any given FDR level α_F. The reason why lfdr values are much less affected by the inclusion of additional variables is that lfdr directly measures the probability that each particular variable is null, and does not say anything about status of those "more extreme" observations than this particular one that we are considering.

Question.

1. Explain what does the values P-value = 0.03, Q-value=0.16 and lfdr=0.47 each tell about the hypothesis shown at the 5th column above? (The 5th column above corresponds to the 15th smallest P-value after we have added 10 P-values with value 1e-10 at the end of the P-value vector.)

Null hypothesis testing vs. effect sizes

So far we have been equating statistical inference to testing the null hypothesis. This has been because in high-dimensional problems our first interest is often to identify the important variables. However, in most applications, eventually, we should be primarily interested in the value of the effect size, rather than the probability whether the effect is zero, let alone on its P-value. Furthermore, since almost all effects are non-zero when we look carefully enough, it follows that if we simply increase our sample size, we will see a statistically significant difference virtually in every possible comparison we can think of at least in fields such

as social sciences, humanities etc. that study very complex phenomena that are affected by almost an infinite number of factors. In those cases, statistically highly significant result may not be at all significant in real life. For example, suppose we can show that group A statistically significantly (P-value < 0.05) earn more than group B when both groups have same education. If this result has been achieved from a large sample of individuals (say millions), and the difference in earnings is very small, say 1%, then this result is unlikely to be surprising: groups A and B are different in some identifiable way, (otherwise you could not tell who belong to A and who to B), and therefore we also expect *at least some small* differences between them in also in other things we can measure. The real question is how large the difference is. If the difference between the groups turns out to be large enough, say 10%, then we may want to seek for a further explanation for it.

Null hypothesis testing can be more informative in natural sciences such as physics, chemistry or genetics where clear and plausible null hypotheses can be formulated and then tested (e.g. mass-energy equivalence in particle physics). They are also useful in ranking predictors to produce sparse models that, as we will see soon, are key methods behind successful high dimensional statistical models.