GWAS 4: Bayesian inference

Matti Pirinen, University of Helsinki

23-Jan-2019

The slide set referred to in this document is “GWAS 4”.

The standard statistical inference typically used in GWAS is based on a fixed significance threshold and only determines whether P-value falls below the GWS threshold. Impicitly, however, GWAS make a more efficient use of the observed data than just labelling it significant or not significant as the exact P-values are reported. Intuitively, we feel that if two variants have P-values of 4e-8 and 2e-200, the latter is certain to be a true association, whereas the former might still go away when we collect more data. But we also know from last week that, at least based purely on its definition, P-value cannot be interpreted as probability of an association being real. Let’s look what information and assumptions we would need in addition to the P-value in order to talk about probability of a non-zero effect. A review on Bayesian methods in GWAS by Stephens & Balding.

From significance to the probability of association

In order to talk about probability of a hypothesis, we need to define the set of all possible hypotheses whose combined probability is 1, i.e., we work under the assumption that one of the hypotheses is true.

(For P-value calculation we only ever define the null hypothesis and therefore we are unable to talk about the probability of the null hypothesis since we haven’t defined what else is possible: Even if the data seem unlikely under the null hypothesis, if the data are even more unlikely under other possible hypotheses, then the null hypothesis might still be quite probable, and such considerations cannot be done from the P-value alone.)

In our case, we consider only two hypotheses: $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$. Next we need to quantify the **prior probabilities** of these hypotheses. These answer to the question: What would I consider as probability of each hypothesis before I have seen the data. The phrase “What would *I*” is there on purpose: prior probabilities are subjective. They are based on whatever knowledge *I* have available. Therefore different persons may have different prior probabilities for the same hypothesis and my prior can (and will!) change as I learn more about the question. For example, $P(H_1) = 10^{-5}$ could be a reasonable prior for a non-zero effect based on what I know about GWAS. (Last week we saw how a magnitude more stringent assumption $P(H_1) = 10^{-6}$ led us to the common GWS threshold.)

Then we observe the data $\mathcal{D}$ and our interest is in the probabilites of each hypothesis after we have seen the data. This is the core question of **Bayesian inference**: How does observing the data update our beliefs from our current state of knowledge, described by prior probabilities $P(H_i)$, into our **posterior probabilities** $P(H_i|\mathcal{D})$? In short: How do we learn from data? Not surprisingly, the answer is the Bayes rule.

4.1 Bayes rule

To write down the Bayes rule (also called Bayes theorem, Bayes formula), we just remember that we are considering joint distributions of two variables. Here the two variables are the hypothesis H_i and the observed data $\mathcal{D}$. We expand their joint probability $P(H_i, \mathcal{D})$ using conditional probability rule in both ways possible

$$P(H_i|\mathcal{D})P(\mathcal{D}) = P(H_i, \mathcal{D}) = P(\mathcal{D}|H_i)P(H_i),$$

from which we can solve the conditional probability that the hypothesis holds given that we have observed the data $\mathcal{D}$:

$$P(H_i|\mathcal{D}) = \frac{P(\mathcal{D}|H_i)P(H_i)}{P(\mathcal{D})}, \qquad \text{for } i = 0, 1.$$

This is the Bayes rule. The conditional probability on the left hand side is also called the **posterior probability** of the hypothesis given the data.

Since the term $P(\mathcal{D})$ (the marginal probability of the observed data) does not depend on hypothesis H_i, we can get rid of it by taking the ratio of the posteriors of the two hypotheses:

$$\frac{P(H_1|\mathcal{D})}{P(H_0|\mathcal{D})} = \frac{P(\mathcal{D}|H_1)P(H_1)}{P(\mathcal{D}|H_0)P(H_0)}.$$

Hence, in order to compute the posterior probability ratio for the hypotheses, we will still need the terms $P(\mathcal{D}|H_i)$ in addition to the prior probabilities. $P(\mathcal{D}|H_i)$ describes what kind of data sets we are likely to see under each hypothesis and with which probability. After these probability densities are specified, the inference is about letting the possible hypotheses compete both in how well they explain the observed data (terms $P(\mathcal{D}|H_1)$ and $P(\mathcal{D}|H_0)$) and in how probable they are *a priori* (prior probablity terms $P(H_1)$ and $P(H_0)$).

Example 4.1. The Bayesian inference shows that both the observed data AND the prior knowledge is crucial for a complete inference. Suppose, for example, that a sequencing effort of my genome returns data $\mathcal{D}$ that seems completely missing chromosome 6. We have two hypotheses: H_0: "There is a technical error", or H_1: "I don't carry any copies of chromosome 6 (in the cells involved)". The observed result could have resulted from either of these options and hence under both hypotheses $P(\mathcal{D}|H_i)$ is similarly very high. So both hypotheses are consistent with the observations and P-values computed under either of the hypothesis as the null hypothesis would not show inconsistencies between observed data and the hypothesis. However, the prior odds of $P(H_1)/P(H_0)$ is pretty small if we think that it is more likely that there is a tecnical error than that I would be missing chr 6 (and still be pretty healthy). Hence the posterior conclusion that combines the prior probabilities and the observed data is that it is more likely that we have an error somewhere in the process than that I don't carry chr 6, even though the observation alone couldn't tell the two hypotheses apart.

Example 4.2. Interpretation of a medical test result is the standard example of the use of Bayes rule. Let's apply it to a case where we try to determine whether an individual has a disease given his genotype.

Suppose that each copy of *HLA-DRB1*1501* allele on chromosome 6 increases the risk of *multiple sclerosis* by OR=3. Prevalence of MS-diseases is $K = 0.001$ and population frequency of DRB1*1501 is 0.20. What is probability of ever getting the disease for each genotype (i.e. 0,1 or 2 copies of DRB*1501)?

Answer. Denote by D the event of getting the disease and by X the genotype. Here D has the role of a hypothesis and X the role of observed data in the above formulation of Bayes rule. Bayes rule says that for each genotype $x \in \{0, 1, 2\}$:

$$P(D\,|\,X = x) = \frac{P(D)P(X = x\,|\,D)}{P(X = x)}.$$

We know that $P(D) = K = 0.001$ and we can assume that the control frequencies are approximately the population frequencies since the diseases has so low prevalence. Assuming HWE, the population frequencies are $P(X = 0) = 0.8^2 = 0.64$, $P(X = 1) = 2 \cdot 0.8 \cdot 0.2 = 0.32$ and $P(X = 2) = 0.2^2 = 0.04$. It follows that case frequencies are

```r
K = 0.001
or = 3
q = c(0.64, 0.32, 0.04) #controls are like population for a low prevalence disease
a = c(1, q[2]/q[1]*or, q[3]/q[1]*or^2) #see GWAS1 for how to get case freqs from controls and OR
f0 = 1/sum(a) #P(X=0 | D)
f = f0*a #case frequencies
rbind(cases=f,controls=q)
```

```
##                 [,1]      [,2]      [,3]
## cases    0.3265306 0.4897959 0.1836735
## controls 0.6400000 0.3200000 0.0400000
```

The risk to get the disease given the genotype is, according to Bayes rule

```
rbind(genotype=c(0,1,2),risk=K*f/q)
```

```
##                  [,1]        [,2]        [,3]
## genotype 0.0000000000 1.000000000 2.000000000
## risk     0.0005102041 0.001530612 0.004591837
```

So even though the disease status has a large effect on the genotype distributions, still even the high risk group has risk $< 0.5\%$ and the genotype doesn't give a practically useful predictive accuracy at the level of an individual from the population because the disease is so rare.

4.2 Probability model for observed GWAS data

To use Bayes rule in GWAS setting, we need to define probability density for observed data under both the null hypothesis and the alternative hypothesis.

In a linear regression GWAS model $y = \mu + x\beta + \varepsilon$, the observed data $\mathcal{D}_k = (\boldsymbol{y}, \boldsymbol{x}_k)$ consist of the phenotype vector $\boldsymbol{y}$ and the vector of genotypes $\boldsymbol{x}_k$ at the tested variant k. When we assume Gaussian errors (i.e. each ε_i having a Normal distribution with same variance σ^2) the probability density of data for a fixed value of effect size β and error variance σ^2 is

$$p\left(\mathcal{D}_k \,|\, \beta, \sigma^2\right) = \mathcal{N}(\boldsymbol{y};\, \boldsymbol{x}_k\beta,\, \sigma^2 I) \propto \exp\left(-(\boldsymbol{y} - \boldsymbol{x}_k\beta)^T(\boldsymbol{y} - \boldsymbol{x}_k\beta)/(2\sigma^2)\right).$$

(We have ignored μ here by assuming that y and x_k are mean-centered values; this just simplifies the notation but doesn't affect the results.)

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 a **prior distribution** for β, for example, by saying that under the alternative model $\beta \sim \mathcal{N}(b_1, \tau_1^2)$. This means that if the alternative model holds, then the true effect sizes are distributed around value b_1 with sd of τ_1. With this prior distribution, the probability of data under H_1 is given by weighting the above likelihood function by the prior probability of each possible value of β:

$$p(\mathcal{D}_k \,|\, H_1) = \int_\beta p\left(\mathcal{D}_k \,|\, \beta, \sigma^2\right) p(\beta \,|\, H_1) d\beta = \int_\beta \mathcal{N}\left(\boldsymbol{y};\, \boldsymbol{x}_k\beta,\, \sigma^2\right) \mathcal{N}\left(\beta;\, b_1,\, \tau_1^2\right) d\beta.$$

(In both models we typically fix σ^2 to its empirical maximum likelihood estimate as the competing regression models do not typically differ in their prior on 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

$$P(\mathcal{D}_k \,|\, H_1) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \text{SE}^2\right),$$

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

$$P(\mathcal{D}_k \,|\, H_0) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \text{SE}^2\right).$$

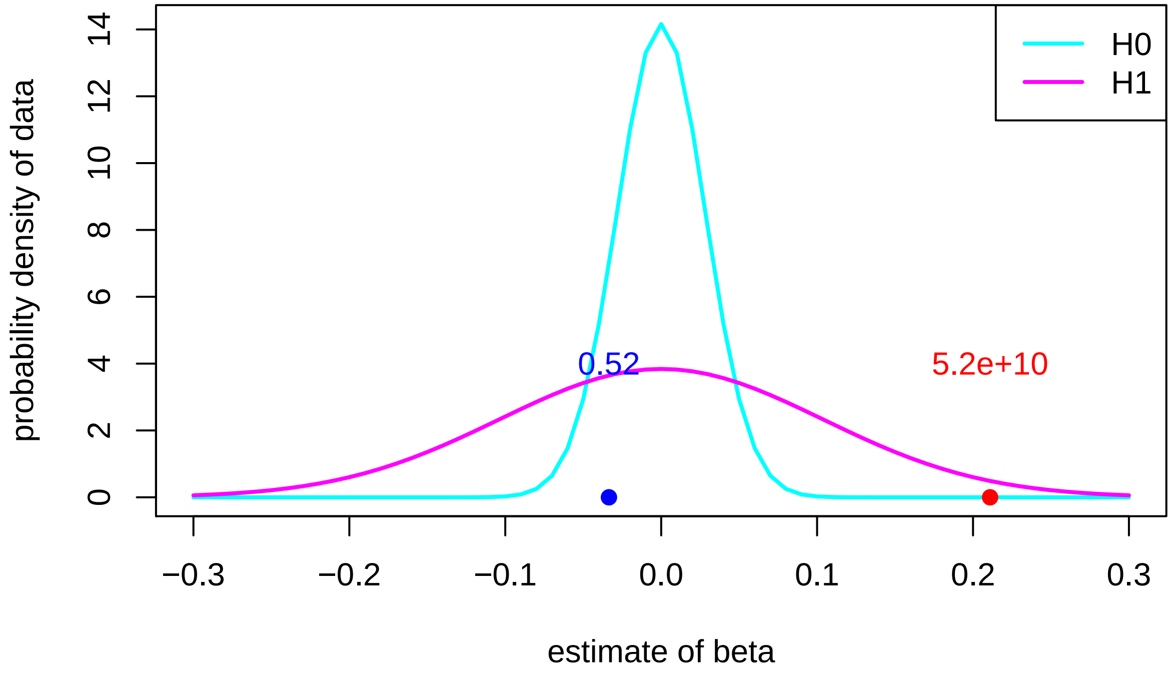
These results tell that we can quantify how well each model explains the data, by asking how well each model can explain the MLE $\widehat{\beta}$. Let's demonstrate this.

Example 4.3.

```
n = 3000 #sample size for SE calculation
f = 0.3 #MAF for SE calculation
sigma = 1 #error SD
se = sigma / sqrt(2*f*(1-f)*n) #SE for QT GWAS
tau = 0.1 #prior standard deviation for effect size beta under H1

#Let's draw probability densities of "data" under the two models, H0 and H1
#as a function of MLE estimate
x = seq(-0.3, 0.3, by = 0.01)
y1 = dnorm(x, 0, sqrt(tau^2 + se^2) )
y0 = dnorm(x, 0, se)
plot(x, y0, t = "l", col = "cyan", lwd = 2, xlab = "estimate of beta", ylab = "probability density of da
lines(x, y1, col = "magenta", lwd = 2)
legend("topright", c("H0","H1"), col=c("cyan","magenta"), lwd = 2)

#We make a shortcut and don't simulate data at all, but we simulate estimates
#Suppose we have two cases, first is null, second is alternative (true betas are 0 and 0.2)
b =c(0, 0.2)
b.est = rnorm(2, b, se) #these are simulated estimates: true means and Gaussian noise determined by SE
points(b.est, c(0,0), pch = 19, col = c("blue","red") )
#Next: log of Bayes factor of H1 vs H0, explained below
#       use log-scale to avoid inacuracies.
log.bf.10 = dnorm(b.est, 0, sqrt(tau^2 + se^2), log=T ) - dnorm(b.est, 0, se, log=T)
bf.10 = exp(log.bf.10) #then turn from log-scale to Bayes factor scale
text(b.est[1], 4, signif(bf.10[1], 2), col = "blue")
text(b.est[2], 4, signif(bf.10[2], 2), col = "red")
```



The distribution of H_0 desribes what kind of effect estimates we can get with this sample size and MAF when true effect is exactly 0. Any deviation from 0 is then by statistical sampling effect (as quantified by SE).

The distribution of H_1 describes what kind of effect estimates we expect when we have BOTH a true non-zero effect (whose expected range is described by standard deviation of τ_1) and we have ALSO a statistical sampling effect (as quantified by SE).

If observed estimate $\widehat{\beta}$ 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 $|\widehat{\beta}| \geq 0.05$. Values close to zero are relatively more likely under the null than under the alternative because the alternative model can also explain observations that are farther away from the zero and hence its probability mass is less concentrated around 0 than that of the null which can only explain well data sets with effect estimates near zero.

4.3 Bayes factor

Two points shown in the plot above are examples of possible estimates that could result either under H_0 (blue) or H_1 (red). The values are ratios of $P(\mathcal{D}|H_1)/P(\mathcal{D}|H_0)$ computed at these two points. When this ratio < 1 the null model H_0 explains the data better and when it is > 1 the opposite is true. This ratio is called **Bayes factor (BF)** and it is the factor that multiplies the prior odds to get to the posterior odds:

$$\underbrace{\frac{P(H_1|\mathcal{D})}{P(H_0|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{prior odds}}.$$

To interpret the Bayes factor, we can think that in order that the posterior probability of the alternative model would be at least 10 times higher than that of the null, the BF needs to be at least 10 times higher than the inverse of the prior odds. If prior odds are 1e-5, then a BF of 1e+6 would give posterior odds > 10.

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. Last week we saw that current genomewide-significant level can be thought to correspond to a prior probability of about $P(H_1) = 10^{-6}$. We know that this seems very stringent compared to true architecture behind complex traits and diseases, but it gives us a conservative reference point. With this prior probability, the posterior odds and posterior probabilities for the alternative model are:

```
post.odds = bf.10*1e-6/(1-1e-6) #P(H_1|D) / P(H_0|D)
post.prob = post.odds/(1+post.odds) #P(H_1|D)
paste(post.prob)
```

```
## [1] "5.22229033295724e-07" "0.999980639798069"
```

For illustration, let's check the P-values corresponding to these two data sets using a Wald statistic $\widehat{\beta}$/SE:

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

```
## [1] 2.342612e-01 6.999659e-14
```

So P-value of the first one is quite close to 0.2, whereas the second is way beyond 5e-8. And the Bayesian analysis said that the first one is almost certain to be null whereas the second one is almost certain to have a non-zero effect. Thus, there is no difference in the conclusion of the standard P-value based GWAS inference and the Bayesian inference, which is typically the case when there is enough data. Conceptually, however, there is a large difference between the P-value, which is probability of at least as extreme data under the null, and the posterior probability of the hypothesis itself.

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 remain useful simple summaries of data that can be computed easily. The important thing is to know what P-values are and what they are not, and that what kind of additional pieces of information would be needed in order to appropriately quantify the probabilities of the hypotheses.

4.4 Approximate Bayes factor in GWAS

The calculation of the Bayes factor above, that was based on the maximum likelihood estimate $\widehat{\beta}$ and its SE, was proposed by Jon Wakefield in 2009. The formula is

$$\text{ABF} \approx \frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)} \approx \frac{\mathcal{N}\left(\widehat{\beta};\,0,\,\tau_1^2+\text{SE}^2\right)}{\mathcal{N}\left(\widehat{\beta};\,0,\,\text{SE}^2\right)} = \frac{(2\pi)^{-0.5}(\tau_1^2+\text{SE}^2)^{-0.5}\exp\left(-\frac{1}{2}\frac{\widehat{\beta}^2}{\tau_1^2+\text{SE}^2}\right)}{(2\pi)^{-0.5}(\text{SE}^2)^{-0.5}\exp\left(-\frac{1}{2}\frac{\widehat{\beta}^2}{\text{SE}^2}\right)}$$

$$= \sqrt{\frac{\text{SE}^2}{\tau_1^2+\text{SE}^2}}\exp\left(\frac{1}{2}\frac{\widehat{\beta}^2}{\text{SE}^2}\frac{\tau_1^2}{\tau_1^2+\text{SE}^2}\right),$$

where the alternative model is specified by effect size prior $H_1 : \beta \sim \mathcal{N}(0, \tau_1^2)$. We have presented the ABF in the form where the alternative is in the numerator and null in the denominator. Hence large ABF means strong evidence in favor of the alternative model. (Wakefield's paper used the inverse of this quantity as Bayes factor, i.e., it computes Bayes factor comparing null to the alternative whereas we compare alternative to null.)

In R, this is easy to compute using `dnorm()` function, as we did above. It is always good to do the ratio of densities on log-scale to avoid possible numerical underflows/overflows. In `dnorm()` this happens by adding `log=TRUE` to the command. Then the ratio of densities becomes a difference between log-densities:

```
log.bf = dnorm(b.est, 0, sqrt(tau^2 + se^2), log=T ) - dnorm(b.est, 0, se, log=T)
bf = exp(log.bf) #turn from log-scale to Bayes factor scale
```

The same formula can be used when $\widehat{\beta}$ and its SE originate from a disease study analyzed by logistic regression. In that case the formula is an approximation based on the assumption that the logistic regression likelihood has a shape of a Gaussian density function. Therefore this approach is generally called **Approximate Bayes Factor (ABF)**.

ABF can be computed from the observed GWAS data $(\widehat{\beta}, \text{SE})$ once we have chosen the variance τ_1^2 of the effect size distribution under the alternative. How should we do that?

Example 4.4. Let's assume that the non-zero effects have a distribution $\mathcal{N}(0, \tau_1^2)$ and we want to determine τ_1 in such a way that with 95% probability the effect (of a SNP with MAF = 0.25) explains less than proportion p of the phenotypic variance $v = \text{Var}(y)$. We will first compute the effect size β_p that explains exactly phenotypic variance of $p \cdot v$, and then we will find the sd parameter τ_1 for which 95% of the probability mass of $\mathcal{N}(0, \tau_1^2)$ is within the region $(-\beta_p, \beta_p)$.

```
v = 1 #Set this to the phenotypic variance
p = 0.01 #effect explains less than 1% of the trait variance,
target.prob = 0.95 #with this probability
maf = 0.25
#2*maf*(1-maf)*b^2 = p*v --> b = +/- sqrt(p*v/(2*maf*(1-maf)))
b = sqrt(p*v/(2*maf*(1-maf)))
tau.seq = seq(0,1,0.001) #grid to evaluate tau
x = pnorm(b, 0, tau.seq, lower=F) #what is the upper tail prob. at b for each value of tau?
tau.1 = tau.seq[which.min( abs(x-(1-target.prob)/2))] #which is closest to target prob?
#Check that the probability mass in (-b,b) is indeed close to target
print(paste0("tau.1=",tau.1," has mass ",signif(1-2*pnorm(b,0,tau.1, lower=F),3),
             " in (-",signif(b,4),", ",signif(b,4),")."))
```

```
## [1] "tau.1=0.083 has mass 0.951 in (-0.1633, 0.1633)."
```

Example 4.5. For case-control GWAS, we want to find such τ_1 that with 95% probability a variant can increase risk by at most OR of 1.30. Now we get the critical point β_p directly as $\log(1.30)$ and then proceed as above.

```
or = 1.30 #effect is at most this large
target.prob = 0.95 # with this probability
```

```
b = log(or)
tau.seq = seq(0,1,0.001) #grid to evaluate tau
x = pnorm(b, 0, tau.seq, lower=F) #what is the upper tail prob. at b for each value of tau?
tau.1 = tau.seq[which.min( abs(x-(1-target.prob)/2))] #which is closest to target prob?
#Check that the probability mass in (-b,b) is indeed close to target
print(paste0("tau.1=",tau.1," has mass ",signif(1-2*pnorm(b,0,tau.1, lower=F),3),
             " in (-",signif(b,4),", ",signif(b,4),")."))
```

```
## [1] "tau.1=0.134 has mass 0.95 in (-0.2624, 0.2624)."
```

Note that the above choices of τ_1 do not model very large effect sizes. There are variants that explain, say, over 10% of the variation of a quatitative trait or that have OR of 3 for some disease. To properly model them in the Bayesian framework, one would need to use several prior distributions and average the results (Bayesian model averaging).