Multiple tests, Bonferroni correction, FDR

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Multiple tests

If we do multiple tests of the same hypothesis, the chance of finding one or more of the tests to be positive increases:
Under the null hypothesis, with n tests, where each has a probability α of a false positive (a Type I error):

- The probability of a false positive in one test is α

Multiple tests

If we do multiple tests of the same hypothesis, the chance of finding one or more of the tests to be positive increases:
Under the null hypothesis, with n tests, where each has a probability α of a false positive (a Type I error):

- The probability of a false positive in one test is α
- The probability of all n of them failing to show a false positive is (if they are independent tests) $(1-\alpha)^n$

Multiple tests

If we do multiple tests of the same hypothesis, the chance of finding one or more of the tests to be positive increases:
Under the null hypothesis, with n tests, where each has a probability α of a false positive (a Type I error):

- The probability of a false positive in one test is α
- The probability of all n of them failing to show a false positive is (if they are independent tests) $(1-\alpha)^n$
- The probability that at least one of them is false positive (so that we reject the null hypothesis) is then, by subtraction

$$\text{Prob (At least one is (falsely) positive)} = 1-(1-\alpha)^n$$

Multiple tests

If we do multiple tests of the same hypothesis, the chance of finding one or more of the tests to be positive increases:
Under the null hypothesis, with n tests, where each has a probability α of a false positive (a Type I error):

- The probability of a false positive in one test is α
- The probability of all n of them failing to show a false positive is (if they are independent tests) $(1-\alpha)^n$
- The probability that at least one of them is false positive (so that we reject the null hypothesis) is then, by subtraction

$$\text{Prob (At least one is (falsely) positive)} = 1-(1-\alpha)^n$$

Probability of a false positive with multiple tests

So the probability of a false positive can get fairly high:

Number of tests	Prob(false positive)
1	0.05
2	0.0975
3	0.142625
4	0.1854938
5	0.2262
10	0.40126
15	0.5367
20	0.6415
50	0.9231
100	0.9941

Independent tests and the Bonferroni correction

To set α so that the probability of rejecting the null hypothesis when there are n independent tests, just take the formula $P = 1 - (1-\alpha)^n$ and solve for α in terms of P, where usually $P = 0.05$. The result is of course $\alpha = 1 - (1-P)^{1/n}$.

We reject the null hypothesis if any of the tests reaches the tail probability α (i.e. if the most significant test reaches it).

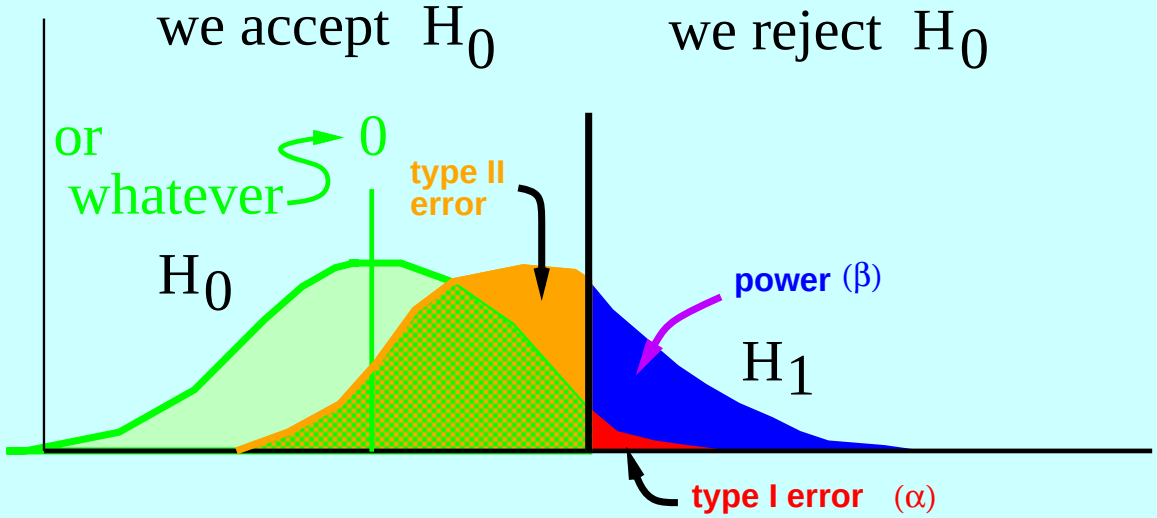
A very good (and even a little conservative) approximation (due to Carlo Emilio Bonferroni and George Boole) is simply P/n:

Number of tests	α if tests independent	Bonferroni $(0.05/n)$
1	0.05	0.05
2	0.02532	0.025
3	0.01695	0.01666
4	0.01274	0.0125
5	0.01021	0.01
10	0.00512	0.005
20	0.00256	0.0025
100	0.000513	0.0005

Interestingly, even if the tests aren't entirely independent, Boole's Inequality used in the Bonferroni correction makes it conservative.

Type I and Type II errors and all that

If we know the alternative hypothesis and how the data are distributed on it, you can calculate the probability of concluding that H_0 is true or false, and the probability that H_1 is true of false:



The probability of (correctly) rejecting the null hypothesis is the *power*, called β (blue area in figure).

A type I error is false rejecting the null hypothesis (red in the figure). A Type II error is false rejecting the alternative hypothesis (orange).

Sensitivity, specificity, FPR, TPR

	Truth is:	
We conclude:	H_1	H_0
H_1	TP	FP
H_0	FN	TN

- False Positive Rate: the fraction of cases which appear to have H_1 true among those which actually have H_0 true ($=?$)

Sensitivity, specificity, FPR, TPR

	Truth is:	
We conclude:	H_1	H_0
H_1	TP	FP
H_0	FN	TN

- False Positive Rate: the fraction of cases which appear to have H_1 true among those which actually have H_0 true ($=?$)
- $FPR = FP/(FP+TN)$ $\quad (= 1 - \text{specificity} = \alpha)$

Sensitivity, specificity, FPR, TPR

	Truth is:	
We conclude:	H_1	H_0
H_1	TP	FP
H_0	FN	TN

- False Positive Rate: the fraction of cases which appear to have H_1 true among those which actually have H_0 true ($=?$)
- $FPR = FP/(FP+TN)$ $\quad (= 1 - \text{specificity} = \alpha)$
- True Positive Rate: the fraction of apparent positives among true positives ($=?$)

Sensitivity, specificity, FPR, TPR

	Truth is:	
We conclude:	H_1	H_0
H_1	TP	FP
H_0	FN	TN

- False Positive Rate: the fraction of cases which appear to have H_1 true among those which actually have H_0 true (=?)
- $FPR = FP/(FP+TN)$ $(= 1 - \text{specificity} = \alpha)$
- True Positive Rate: the fraction of apparent positives among true positives (=?)
- $TPR = TP/(TP+FN)$ $(= \text{sensitivity} = \text{power} = \beta)$

The False Discovery Rate

$$FDR = FP/(FP+TP)$$

Contrary to what you will hear, **it is for a different case**:

- Bonferroni: Testing the same hypothesis many times, or many different ways. (Some tests may be much less powerful than others). If even one test really shows that the null hypothesis is wrong, then it is dead.
- FDR: We are testing many hypotheses (in a similar way). We want to know which ones are likely to be true.

The idea of the considering the FDR is to see what the fraction of false positives might be among among all those that appear to be true.

Holm's method

Another way of choosing promising hypotheses is to accept the most significant test, then the next most significant, and so on until one fails.

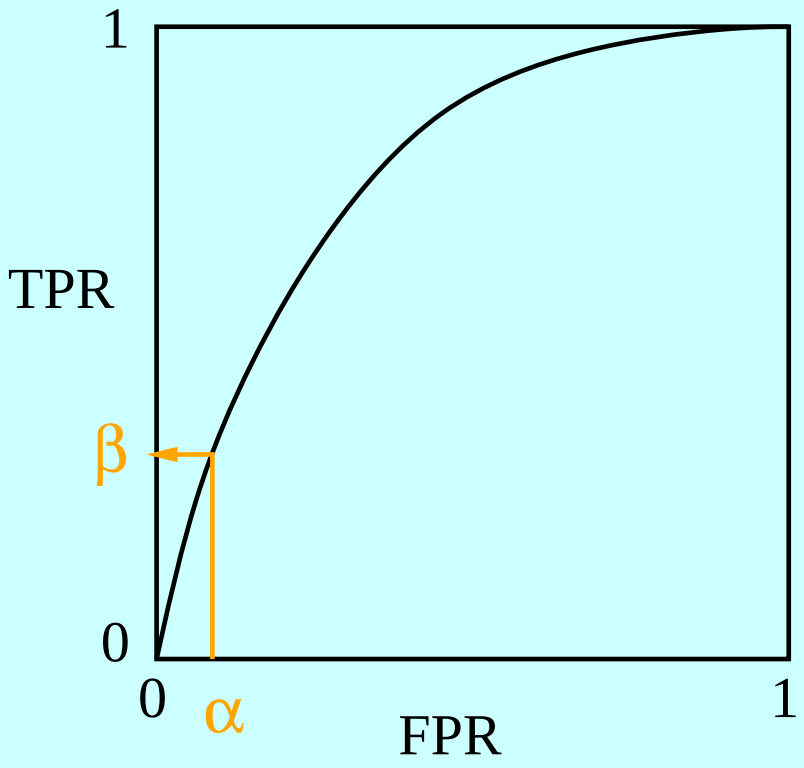
Holm showed that the proper way to do this, to have a probability α of accepting any test when the null hypothesis is actually true is, say if there were 120 tests of different hypotheses:

- Test the one with the smallest P value against $\alpha/120$
- Test the one with the next smallest P value against $\alpha/119$
- Test the one with the next smallest P value against $\alpha/118$
- ... and so on, until one is not significant.

This is an alternative to the FDR machinery.

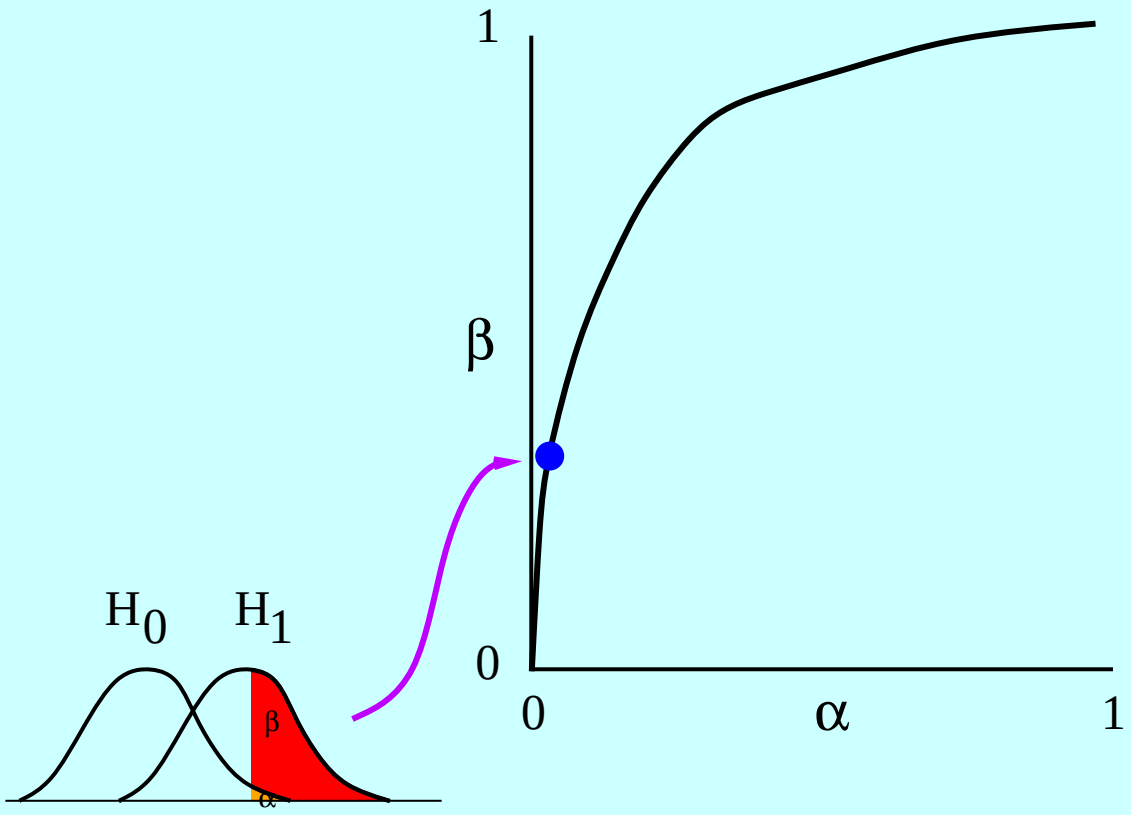
The ROC Curve

In the (mostly unrealistic) cases where we know the distributions of data under the null hypothesis and the alternative hypothesis, we can plot the TPR as a function of the FPR, for different P values we might use. For historical reasons (radar in World War II) this is called the Receiver Operating Characteristic curve:



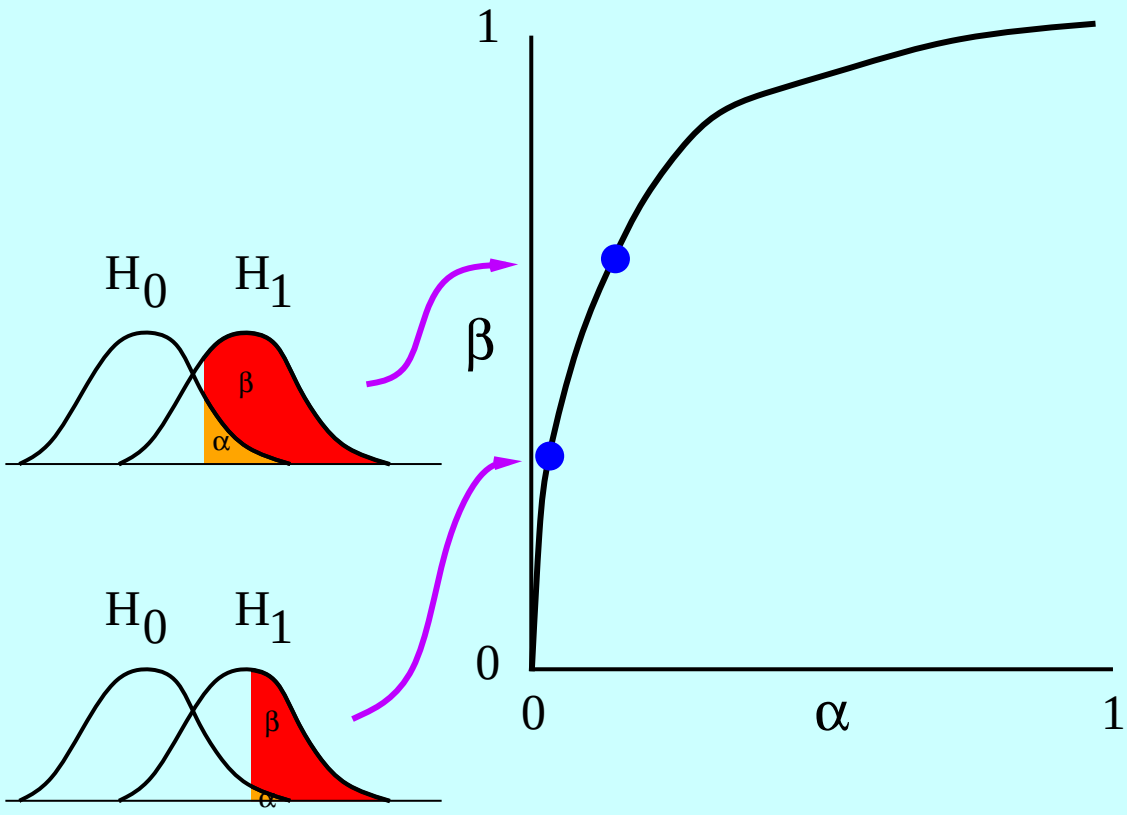
The ROC Curve and the type I and type II errors

Here is a diagram showing plotting the fraction of True Positives against the power (which we could do if we knew everything):



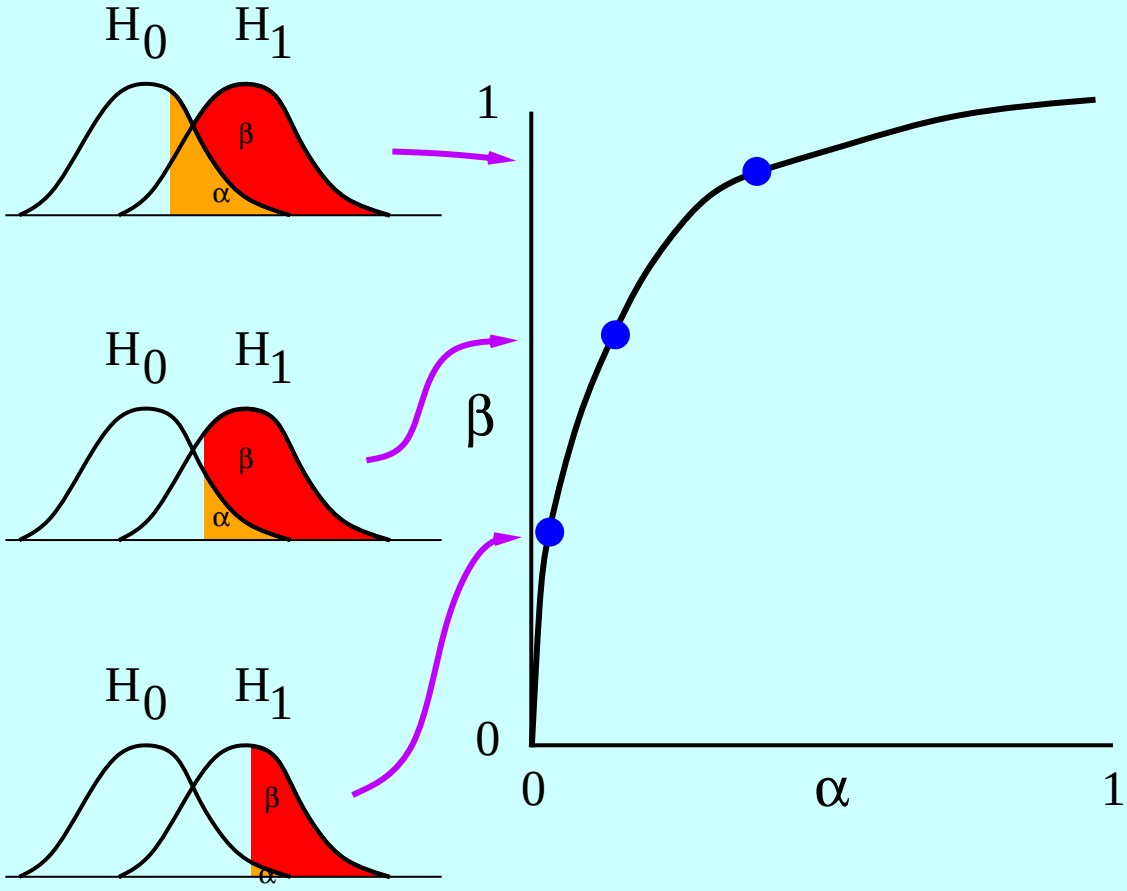
The ROC Curve and the type I and type II errors

Here is a diagram showing plotting the fraction of True Positives against the power (which we could do if we knew everything):



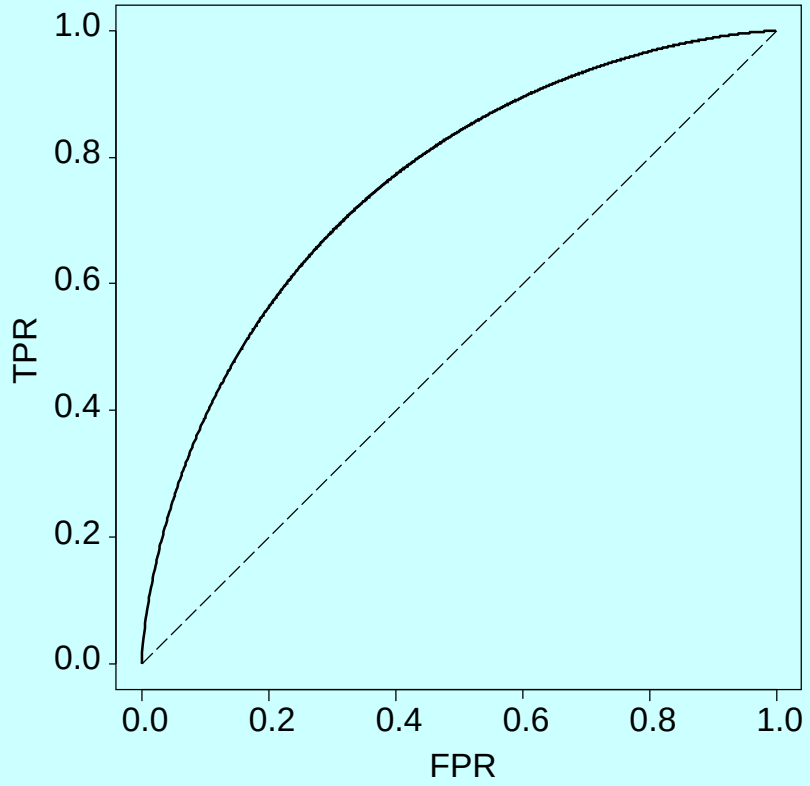
The ROC Curve and the type I and type II errors

Here is a diagram showing plotting the fraction of True Positives against the power (which we could do if we knew everything):



The ROC Curve – an example

If the null is Normal(0,1) and the alternative is Normal(1,1), and the data is one observation, here is the ROC:



The distribution of P is flat under the null hypothesis

- The probability that P just reaches the 0.05 level is (duh) 0.05.

The distribution of P is flat under the null hypothesis

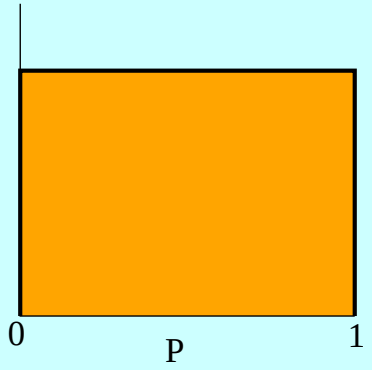
- The probability that P just reaches the 0.05 level is (duh) 0.05.
- The probability that P just reaches the 0.2 level is 20%.

The distribution of P is flat under the null hypothesis

- The probability that P just reaches the 0.05 level is (duh) 0.05.
- The probability that P just reaches the 0.2 level is 20%.
- And similarly for any other value.

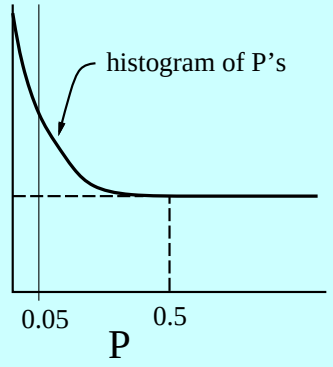
The distribution of P is flat under the null hypothesis

- The probability that P just reaches the 0.05 level is (duh) 0.05.
- The probability that P just reaches the 0.2 level is 20%.
- And similarly for any other value.
- So the distribution of the P if the null hypothesis is true is a flat (rectangular) distribution between 0 and 1:



Storey's way of calculating the FDR

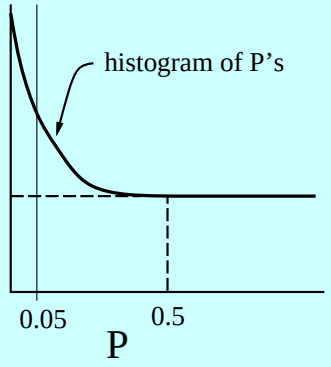
(A simplified version). When there are many tests (say 1000) we plot the P values for all tests.



- If all are false, they should be uniformly distributed.

Storey's way of calculating the FDR

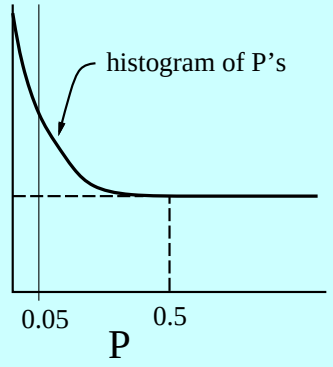
(A simplified version). When there are many tests (say 1000) we plot the P values for all tests.



- If all are false, they should be uniformly distributed.
- If some are from the alternative hypothesis, those should be concentrated near 0.

Storey's way of calculating the FDR

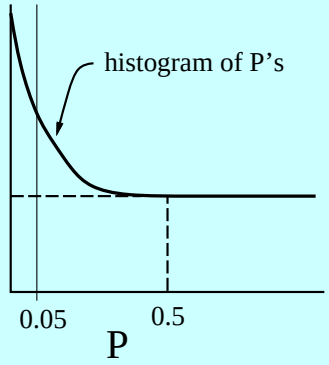
(A simplified version). When there are many tests (say 1000) we plot the P values for all tests.



- If all are false, they should be uniformly distributed.
- If some are from the alternative hypothesis, those should be concentrated near 0.
- Estimate the number that are from the null hypothesis by doubling the number of those that are above P = 0.5. The remaining number are probably from the alternative hypothesis.

Storey's way of calculating the FDR

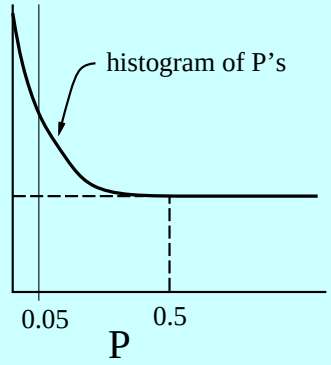
(A simplified version). When there are many tests (say 1000) we plot the P values for all tests.



- If all are false, they should be uniformly distributed.
- If some are from the alternative hypothesis, those should be concentrated near 0.
- Estimate the number that are from the null hypothesis by doubling the number of those that are above P = 0.5. The remaining number are probably from the alternative hypothesis.
- From this you can estimate, for any cutoff value (α) what fraction of the positives will be false. (You can't tell which ones they are).

Storey's way of calculating the FDR

(A simplified version). When there are many tests (say 1000) we plot the P values for all tests.



- If all are false, they should be uniformly distributed.
- If some are from the alternative hypothesis, those should be concentrated near 0.
- Estimate the number that are from the null hypothesis by doubling the number of those that are above P = 0.5. The remaining number are probably from the alternative hypothesis.
- From this you can estimate, for any cutoff value (α) what fraction of the positives will be false. (You can't tell which ones they are).
- For example, if out of T tests A are above 0.5, another A are estimated to be uniformly distributed below 0.5. Below 0.05 will be A/10. So the FDR is A/10 divided by the number you see below 0.05, which we could call B.

If you’re a Bayesian, however ...

... you don’t need any of the above, just use your priors and the posteriors.