# Multiple Testing

## 1 Executive Summary

Suppose we have a large number of features and we test each one to see if it is significant. Even if we set a test threshold so that each individual test has a small false-positive probability, there may still be a large probability that one or more of the tests produces a false-positive. To avoid this, the test threshold needs to be increased to compensate for the multiple testing problem.

## 2 Gaussian Model

Suppose we have $m$ random variables $x_1, x_2, \ldots, x_m$ and for each we consider the following hypothesis test:

$$\begin{aligned} H_0 &: x_i \sim \mathcal{N}(0,1) \\ H_1 &: x_i \sim \mathcal{N}(\mu,1)\ ,\ \mu > 0 \end{aligned}$$

$\mathcal{N}(0,1)$ is called the null distribution (irrelevant feature) and $\mathcal{N}(\mu,1)$ is the alternative distribution (for relevant features). The log LRT for each feature has the form $\mu x - \mu^2/2 \underset{H_0}{\overset{H_1}{\gtrless}} \tau$. We can absorb $\mu^2/2$ into the threshold and since (the unknown) $\mu > 0$ we can divide both sides by $\mu$ and arrive at the test

$$x_i \underset{H_0}{\overset{H_1}{\gtrless}} t\ .$$

The *family-wise error rate* FWER is the probability of one or more false-positives occurring in the $m$ tests. The event that there are one or more false-positives is given by

$$\max_i x_i \geq t\ .$$

If we define $C(t,m) = \sum_{i=1}^m \mathbb{1}(x_i \geq t)$, then this event can also be written as

$$C(t,m) \geq 1\ .$$

We would like to control $C(t,m)$ by choosing the threshold $t$ to be large enough. To do this, we will use the following upper and lower bounds on the tail of the Gaussian density.

$$\frac{1}{\sqrt{2\pi t^2}}(1-t^{-2})e^{-t^2/2} \ \leq\ \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \ \leq\ \frac{1}{2}e^{-t^2/2}\ .$$

Now suppose all the $x_i \sim \mathcal{N}(0,1)$ and we want to have few if any false-positives. Consider the expected value of $C(t,m)$.

$$\begin{aligned} \mathbb{E}[C(t,m)] &= \sum_{i=1}^m \mathbb{E}[\mathbb{1}(x_i \geq t)] \ =\ m\mathbb{P}(x_1 \geq t)\ ,\ \text{since the } x_i \text{ are iid} \\ &\leq \frac{m}{2}e^{-t^2/2} \ =\ \frac{1}{2}e^{-(t^2/2 - \log m)}\ ,\ \text{using the upper bound above.} \end{aligned}$$

Thus, taking $t \geq \sqrt{2\log m}$ guarantees that $\lim_{m\to\infty} \mathbb{E}[C(t,m)] \leq 1/2$. This suggests that our threshold should grow like $t = \sqrt{2\log m}$. Also notice that the FWER is simply $\mathbb{P}(C(t,m) \geq 1) \leq \mathbb{E}[C(t,m)]$, by Markov's inequality. Thus, $t = \sqrt{2\log m}$ indeed controls the FWER.

The problem with this threshold is that it grows with $m$, meaning that we may detect fewer true positives as $m$ grows. We might hope that our bounding isn't so sharp and we can use a lower threhold. To see that this is not the case, consider

$$\begin{aligned}
\mathbb{E}[C(t,m)] &= \sum_{i=1}^{m} \mathbb{E}[\mathbb{1}(x_i \geq t)] = m\mathbb{P}(x_1 \geq t) \text{ , since the } x_i \text{ are iid} \\
&\geq m\frac{1}{\sqrt{2\pi t^2}}(1-t^{-2})e^{-t^2/2} \text{ , using the lower bound above.}
\end{aligned}$$

Now take $t = \sqrt{\beta \log m}$ for some $0 < \beta < 2$. Plugging this in give us the bound

$$\mathbb{E}[C(t,m)] \geq m^{1-\beta/2}\frac{1}{\sqrt{2\pi\beta\log m}}\left(1 - \frac{1}{\beta\log m}\right) .$$

Since $\beta < 2$, this diverges as $m \to \infty$. Thus $t = \sqrt{2\log m}$ is necessary and sufficient to (asymptotically) control the FWER.