# DS 102 Discussion 2

Monday, February 3, 2020

1. **Distribution of $p$-values.** Understanding $p$-values and their distribution can help develop intuition about how to control FDR. We want to test two hypotheses. Under $H_0$ the data is generated by the null distribution $\mathcal{N}(0,1)$, and under the alternative $H_1$, the data is generated by a distribution $\mathcal{N}(\mu,1)$ where $\mu > 0$.

   Suppose you have a test $T(X)$, where $X$ is your data. Recall that the p-value $P$ of the test is:

   $$P(X) = \mathbb{P}(T > t | T(X) = t)$$

   where $\mathbb{P}$ denotes the probability with respect to data drawn from the null distribution. This is the probability that, under the null distribution, you see a result at least as extreme as from your data. Remember that a $p$-value is a random variable, since it is a function of your data.

   (a) Let $F(t)$ be the cumulative density function (CDF) of the test statistic $T(X)$ under the null distribution. Recall that $F$ is a monotonically increasing function (if that's not clear, draw a picture to see why). Suppose that $F$ is invertible, meaning that there exists a function $F^{-1}$ such that:

   $$F^{-1}(F(t)) = t.$$

   What is the CDF of the $p$-value? What can we conclude about $p$-values under the null?

   **Solution:** Let us first write the $p$-value as a function of the CDF of $T$.

   $$p = \mathbb{P}(t > T(X)) = 1 - F(T(X))$$

   Given this, the CDF of $p$ is given by:

   $$\begin{aligned}
   \mathbb{P}(p \leq a) &= \mathbb{P}(1 - F(T(X)) \leq a) \\
   &= \mathbb{P}(F(T(X)) \geq 1 - a) \\
   &= \mathbb{P}(T(X) \geq F^{-1}(1-a)) \\
   &= 1 - \mathbb{P}(T(X) \leq F^{-1}(1-a)) \\
   &= 1 - F(F^{-1}(1-a))) \\
   &= a
   \end{aligned}$$

   Since this is the CDF of a uniform distribution over $[0,1]$, we conclude that $p$-values are uniformly distributed under the null.

   (b) What part of the proof above fails if data do not come from the null distribution?

**Solution:** We need to account for the two different distributions. In particular, the second to last line does not hold, since

$$F_1(F_0^{-1}(p)) \neq p$$

(c) Suppose you have two independent $p$-values $p_1$ and $p_2$. If on both hypotheses you choose a $0 < \alpha < 1$ and use the naive decision rule:

$$\delta(p;\alpha) = \begin{cases} \text{reject null} & p \leq \alpha \\ \text{accept null} & p > \alpha \end{cases}$$

what is the probability of making at least one false discovery? This probability is also known as the family-wise error rate (FWER).

**Solution:** A false discovery occurs if you reject the null hypothesis when the data comes from the null distribution. That is, $p_i < \alpha$ when the null is true. From previous parts of this discussion, we know that under the null distribution, $p$-values are uniformly distributed. Therefore we have:

$$\mathbb{P}(\text{false discovery}) = \mathbb{P}(p \leq \alpha) = \alpha$$

If we have two p-values, the probability that we make a false discovery is:

$$\begin{aligned} \mathbb{P}(\text{at least one false discovery}) &= 1 - \mathbb{P}(\text{no false discovery}) \\ &= 1 - (1-\alpha)^2 \\ &= \alpha(2-\alpha) \end{aligned}$$

(d) Does this decision rule keep the probability of a false discovery below $\alpha$?

**Solution:** No, since $\alpha(2-\alpha) \geq \alpha$ for $0 < \alpha < 1$, we have that:

$$\mathbb{P}(\text{at least one false discovery}) > \alpha.$$

The naive decision rule doesn't even work with two $p$-values!

2. **False discoveries in genome-wide association studies (GWAS).**

In "classical" genetics, researchers identify one candidate mutation they believe is associated with a trait (*e.g.* a disease), then conduct an experiment to test that hypothesis. In contrast, the recent explosion of genomic data has led to the rise of genome-wide association studies (GWAS), where researchers search for associations between a trait and

mutations across the entire genome. For example, in one of the first GWAS (Klein et al. 2005), 103,611 mutations across the genome were tested for association with macular degeneration. Methods for controlling FWER or false-discovery rate (FDR) are critical for ensuring the findings of GWAS are meaningful and actionable: today, the mutation identified in that first GWAS is showing promising results as a therapeutic target.

(a) The Bonferroni correction, which uses the decision rule

$$\delta\left(p; \frac{\alpha}{n}\right)$$

controls the FWER described in the previous problem. Suppose you have $n$ independent $p$-values: $p_1, \ldots, p_n$. Show that the Bonferroni correction controls the probability of at least one false discovery. *Hint: Let $E_i$ be the event that $p_i < \frac{\alpha}{n}$.*

**Solution:** Using the union bound,

$$\begin{aligned}\mathbb{P}(\text{at least one false discovery}) &= \mathbb{P}(\cup_{i=1}^{n} E_i) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(E_i) \\ &\leq n\frac{\alpha}{n} = \alpha\end{aligned}$$

(b) Klein and his colleagues tested $103,611$ hypotheses in their GWAS. Using the Bonferroni correction, what threshold should be used for the decision rule such that the FWER is less than 0.05?

**Solution:**

$$\frac{\alpha}{n} = \frac{0.05}{103611} = 4.83 \times 10^{-7}.$$

Fig. 1 shows Figure 1(a) from Klein et al. (2005), which plots the negative log of $p$-values of all tested mutations and the Bonferroni-corrected threshold of $4.8 \times 10^{-7}$. Two mutations passed the threshold, and were considered candidates for therapeutic targets in subsequent studies.

# References

R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, J. Hoh. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720): 385-389.

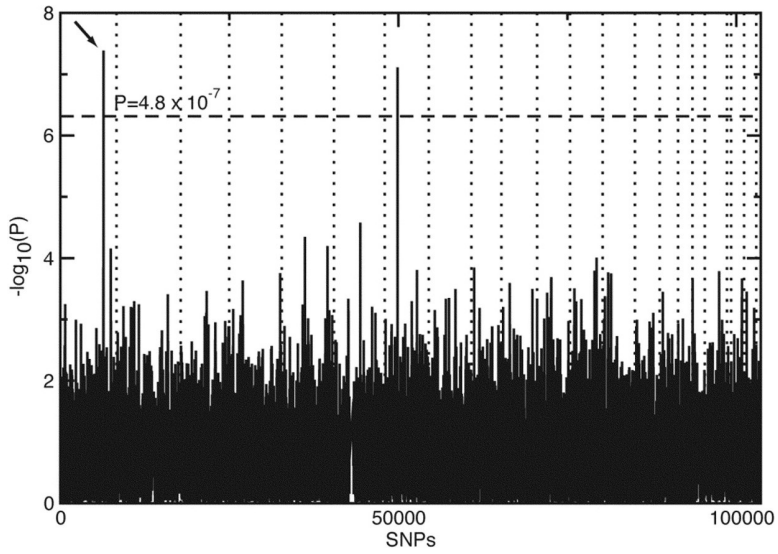


Figure 1: Figure 1(a) from Klein et al. (2005). From their caption: *$P$ values of genome-wide association scan for genes that affect risk of developing macular degeneration. ... The dotted horizontal line shows the cutoff for $P = 0.05$ after Bonferroni correction. The vertical dotted lines show chromosomal boundaries. The arrow indicates the peak for SNP rs380390, the most significant association, which was studied further.*