Multiple testing*

- Conducting hypothesis test. First, we define the null and alternative hypotheses. Next we construct a test statistic that summarizes the strength of evidence against the null hypothesis. We then compute a p-value that quantifies the probability of having obtained a comparable or more extreme value of the test statistic under the null hypothesis. Finally, based on the p-value, we decide to reject the null. When conducting multiple testing, we should be careful in order to avoid to reject too many null.
- we usually use normal distribution, t-distribution, χ^2 distribution, F distribution.

Decision \ Truth	H_0	H_a	
Reject H_0	Type I ✓	✓ caused ↑ power.	R
Do not reject H_0	✓ caused ↓	Type II ✓	m − R
	m_0	$m - m_0$	m

- Type I : incorrectly rejecting H_0.
- The power of the hypothesis test is defined as the probability of not making a Type II error given that H_a holds : i.e the probability of correctly rejecting H_0

- when we perform many test, we are bound to make a large number of type I errors.

Family-wise error Rate

- FWER is the probability of making at least one type I error. $\Rightarrow FWER = P_r(V \geq 1)$. A strategy of rejecting any null hypothesis for which the p-value is below α, leads to $FWER(\alpha) = 1 - P_r(V = 0)$
  $= 1 - P_r(\text{do not falsely reject } H_0)$
  $= 1 - P_r(\bigcap_{j=1}^{m} \text{falsely reject } H_j)$
  $= 1 - \prod_{j=1}^{m} (1-\alpha) = 1 - (1-\alpha)^m$
- for instance $\alpha = 0,05 \Rightarrow FWER = 1 - (1-0,05)^{100} = 0,994$. we are $m = 100$ eventually guaranteed to make at least one type I error.

- Bonferroni method and Holm's step down procedure, are general purpose approaches for controlling the FWER that can be applied whenever m p-values have been computed regardless of the form of the null hypotheses the choice of test statistics, or the (in)dependence of the p-value.

Bonferroni Method

$FWER = P_r(\text{falsely reject at least one null hypothesis}) = P_r(\bigcup_{j=1}^{m} A_j) \leq \sum_{j=1}^{m} P_r(A_j) \Rightarrow FWER(\alpha/m) \leq m \times \frac{\alpha}{m} = \alpha$

this correction can be quite conservative, in the sense that the true FWER is often quite a bit lower that the nominal (or target) FWER. this results from the inequality. By construct a less conservative procedure might allow us to control the FWER while rejecting more null hypotheses, and therefore making fewer type II errors.

Holm's step-down procedure

This method is less conservative than Bonferroni in the sense it will reject more null hypotheses, typically resulting in fewer type II errors and hence greater power. The threshold used to reject each null hypothesis $P(L)$ depend on the values of all m p-values.

Holm will always reject more tests than Bonferroni

1. Specify α, the level at which to control the FWER
2. Compute p-values, $P_1, \dots P_m$ for the m null hypotheses $H_0, \dots H_m$
3. Order the m p-values so that $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$
4. Define $L = \min\left\{ j : P_{(j)} > \frac{\alpha}{m+1-j} \right\}$
5. Reject all null hypothesis H_j for which $P_j < P_{(L)}$.

- In certain special cases, more powerful procedures for multiple testing correction may be available, in order to control the FWER while achieving higher power (i.e committing fewer type II errors) than would be possible using Holm or Bonferroni → Tukey's method
  → Scheffé's method
- In general there is a tradeoff between the FWER threshold we choose and the power. In practice, when m is large, we may be willing to tolerate a few false positive, in the interest of making more discoveries. i.e more rejections of the null hypotheses, this is the motivation behind the false discovery rate.

False discovery rate

- When m is large, trying to prevent any false positive (as in FWER control) is simply too stringent. Instead we might try to make sure that the ratio of false positive (V) to total positive ($V+S = R$) is sufficiently low, so that most of the rejected null hypotheses are not false positives. The ratio V/R is the false discovery proportion (FDP). $\Rightarrow$ $FDR = E(FDP) = E(V/R)$ (the choice of FDR threshold is context dependent).
- The Benjamini-Hochberg procedure gives us a very easy way to determine, given a set of m p-values, which null hypotheses to reject in order to control the FDR at any pre-specified level q ($FDR \leq q$)
  1. Specify q, the level at which to control the FDR
  2. Compute p-values, $p_1, \dots, p_m$ for the m null hypotheses $H_{01}, \dots H_{0m}$
  3. Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
  4. Define $L = \max\{j : p_{(j)} < qj/m\}$
  5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$.

Re-sampling

- If our null hypothesis H_0 or test statistic T is somewhat unusual, then it may be the case that no theoretical null distribution exists, then we may be wary of relying upon it, perhaps because some assumption that is required for it to hold is violated (i.e sample size too small).
- Using re-sampling approach gave us substantially different results from using the theoretical p-value. There are two setting in which a re-sampling approach is useful:
  - Perhaps no theoretical null distribution is available. This may be the case if you are testing an unusual null hypothesis H_0, or using an unusual test statistic T.
  - Perhaps a theoretical null distribution is available but the assumptions required for its validity do not hold ~~for evidence~~
- In general, if you can come up with a way to resample or permute the observations in order to generate data that follows the null distribution, then you can compute p-values or estimate the FDR using methods. In many real-world settings this provides a powerful tool for hypothesis testing when no out-of-the-box hypothesis test are available, or when the key assumption underlying those out of the box tests are violated.