Hypothesis Testing

Yun (Renee) Zhang, PhD

Email: zhangy@jcvi.org

Monthly Biostatistics Discussion

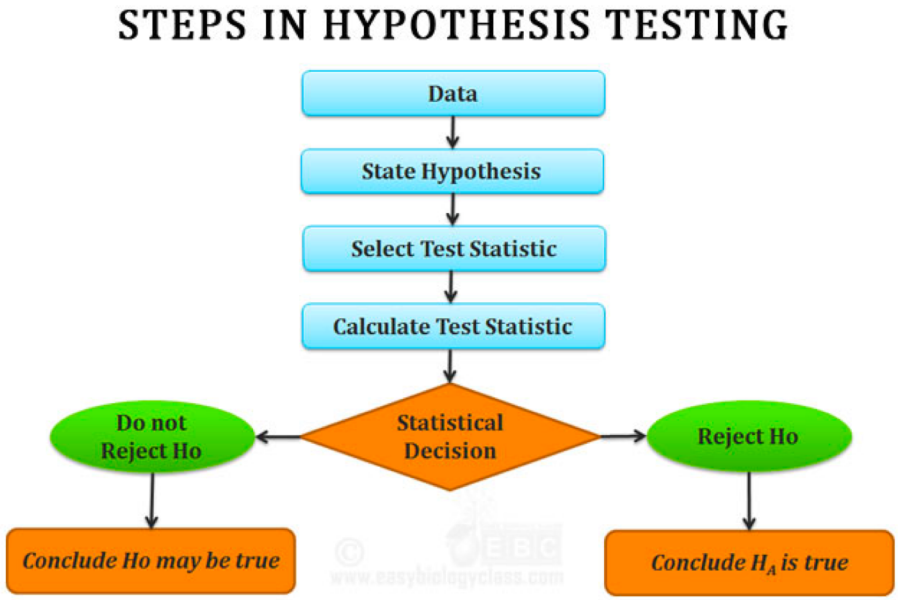
12-19-2018

Outline

1. Hypothesis Testing
2. Power and Sample Size
3. Multiple Hypothesis Testing Correction

Hypothesis Testing

Elements in Hypothesis Testing



- Null hypothesis (H_0) (e.g. H_0: $\mu = 0$)
- Alternative hypothesis (H_1) (e.g. H_1: $\mu \neq 0$)
- Test statistic (e.g. t-statistic $t = \frac{\bar{x}}{s/\sqrt{n}}$, calculated from data)
- Significance level (α) (usually $\alpha = 0.05$, pre-determined)
- Decision rule

Decision Rules

- Decision rule based on critical value / rejection region:

  Reject H_o if test statistic falls in rejection region

- Decision rule based on p-value:

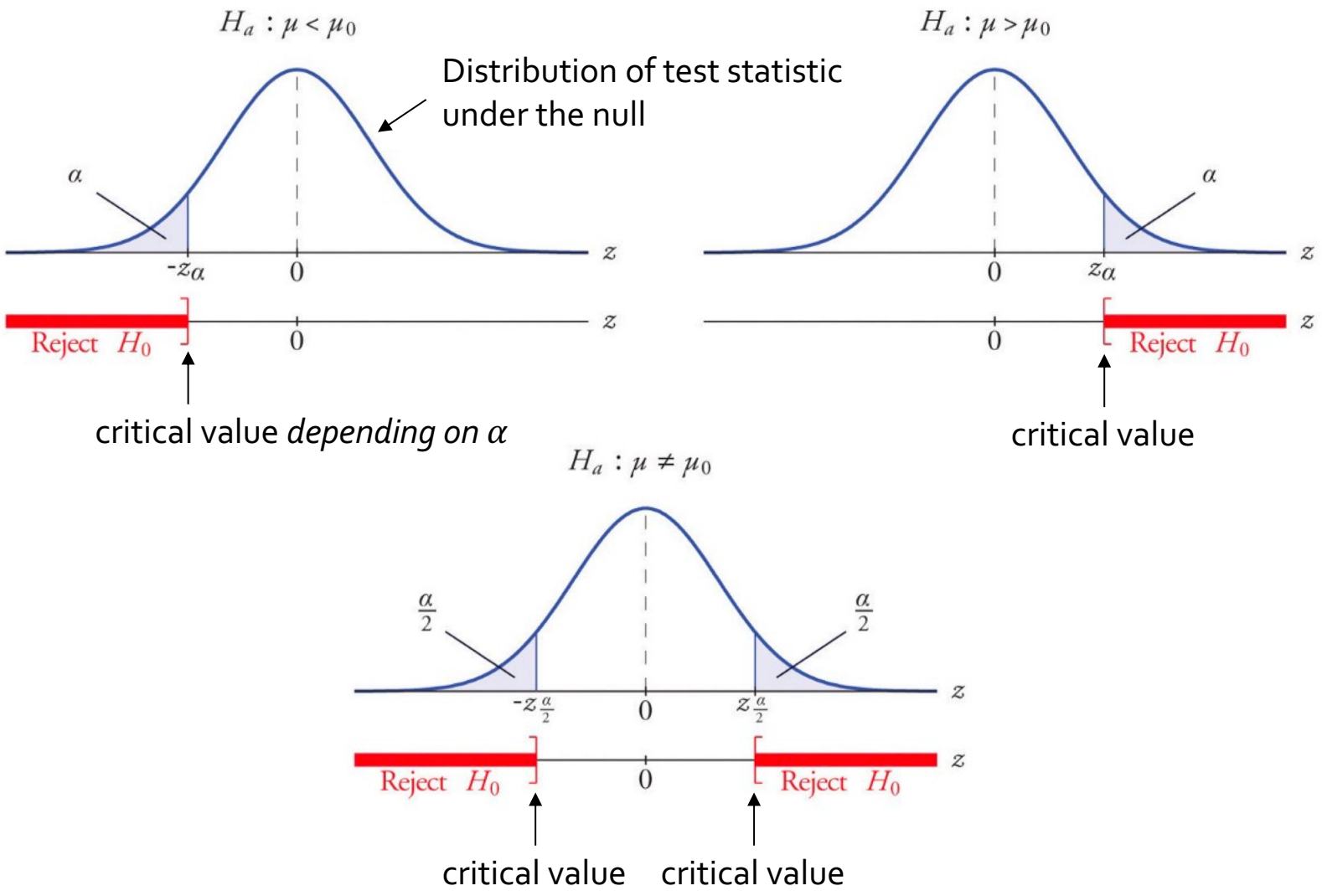
  Reject H_o if p-value $< \alpha$

- Decision rule based on confidence interval:

  Reject H_o if $(1-\alpha)$-confidence interval does not contain the value in H_o

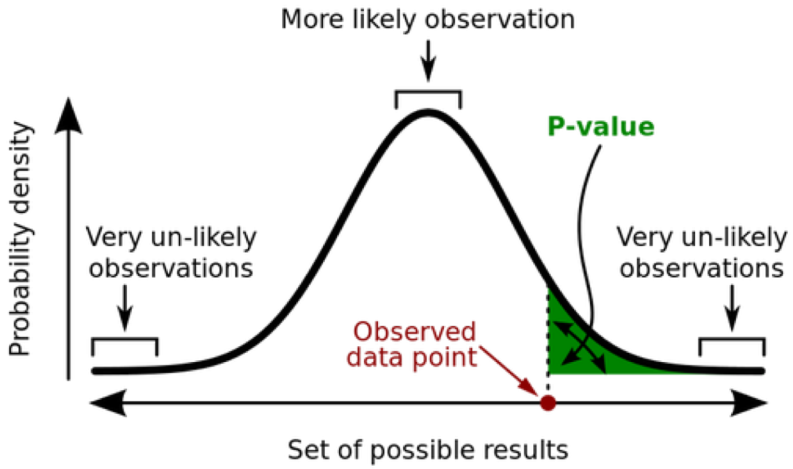
- All decision rules are equivalent

Critical Value and Rejection Region



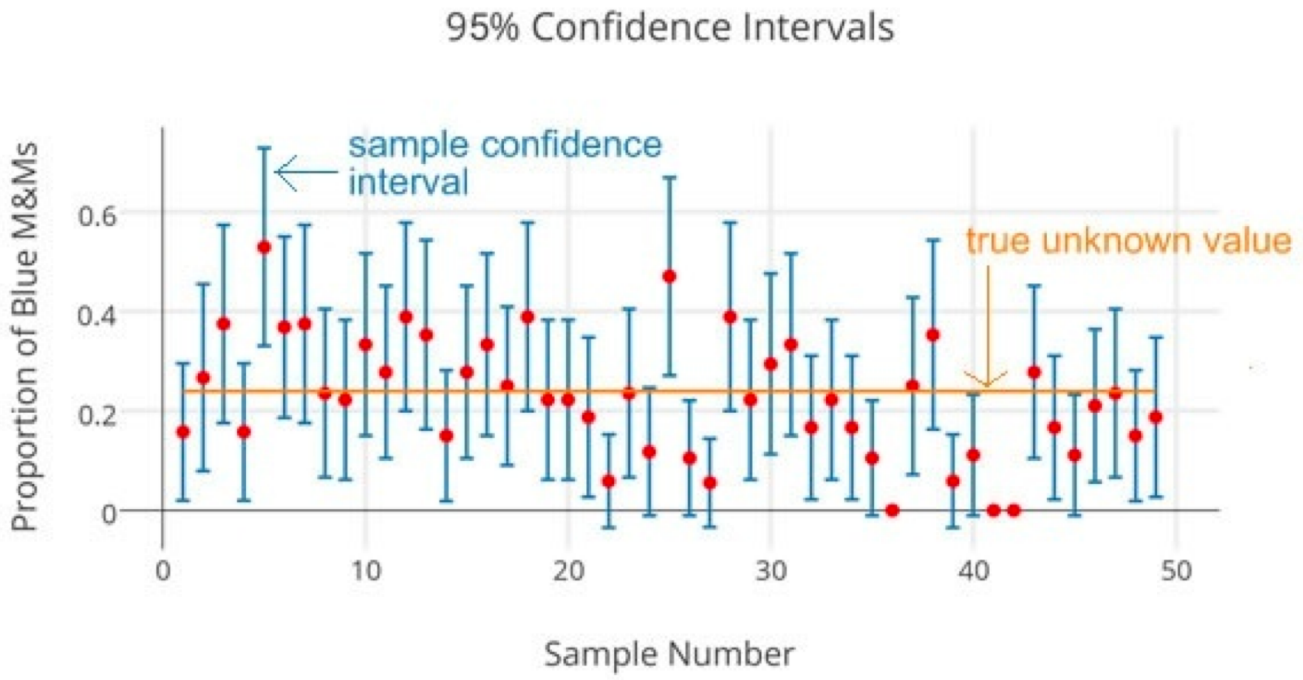
P-value

- The **p-value** is the probability of getting the observed, or more extreme, results *when the null hypothesis* (H_0) *is true* ✓



- The p-value does not provide the probability that either hypothesis is correct (a common source of confusion) ✗
- $\Pr(\text{observation} \mid \text{hypothesis}) \neq \Pr(\text{hypothesis} \mid \text{observation})$
- The term **significance level (alpha)** is used to refer to a *pre-chosen* probability and the term "p-value" is used to indicate a probability that you *calculate* after a given study
- $p = \Pr(\text{reject } H_0 \mid H_0 \text{ is true})$, to be compared to α

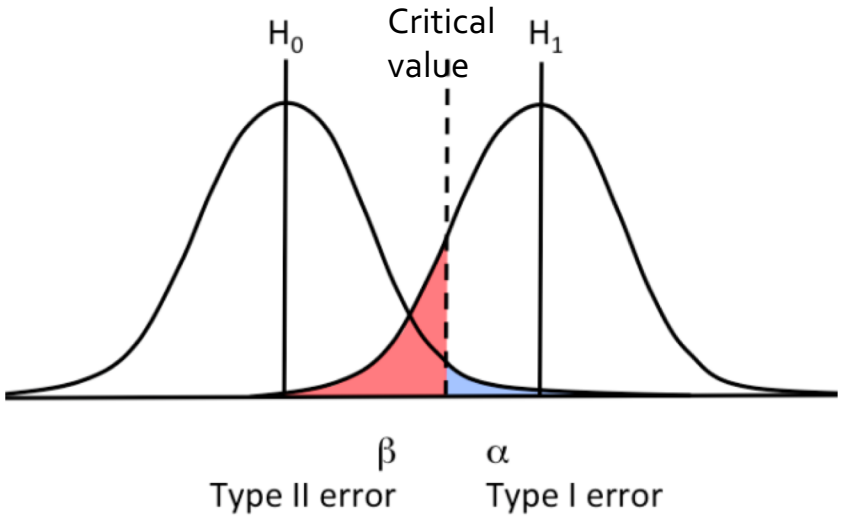
Confidence Interval



- "*Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%"* ✓
- A 95% probability that the interval covers the population parameter ✗
- A 95% confidence interval does not mean that 95% of the sample data lie within the interval ✗

Type I and Type II Errors

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Fail to reject	Correct inference (True Negative) (Probability = $1 - \alpha$)	**Type II error** (False Negative) (Probability = β)
	Reject	**Type I error** (False Positive) (Probability = α)	Correct inference (True Positive) (Probability = $1 - \beta$)



Power and Sample Size

Statistical Power

- **Power** is the probability that the test correctly rejects the null hypothesis (H_0) *when a specific alternative hypothesis (H_1) is true*
- $\text{Power} = \Pr(\text{reject } H_0 \mid H_1 \text{ is true}) = 1 - \beta$

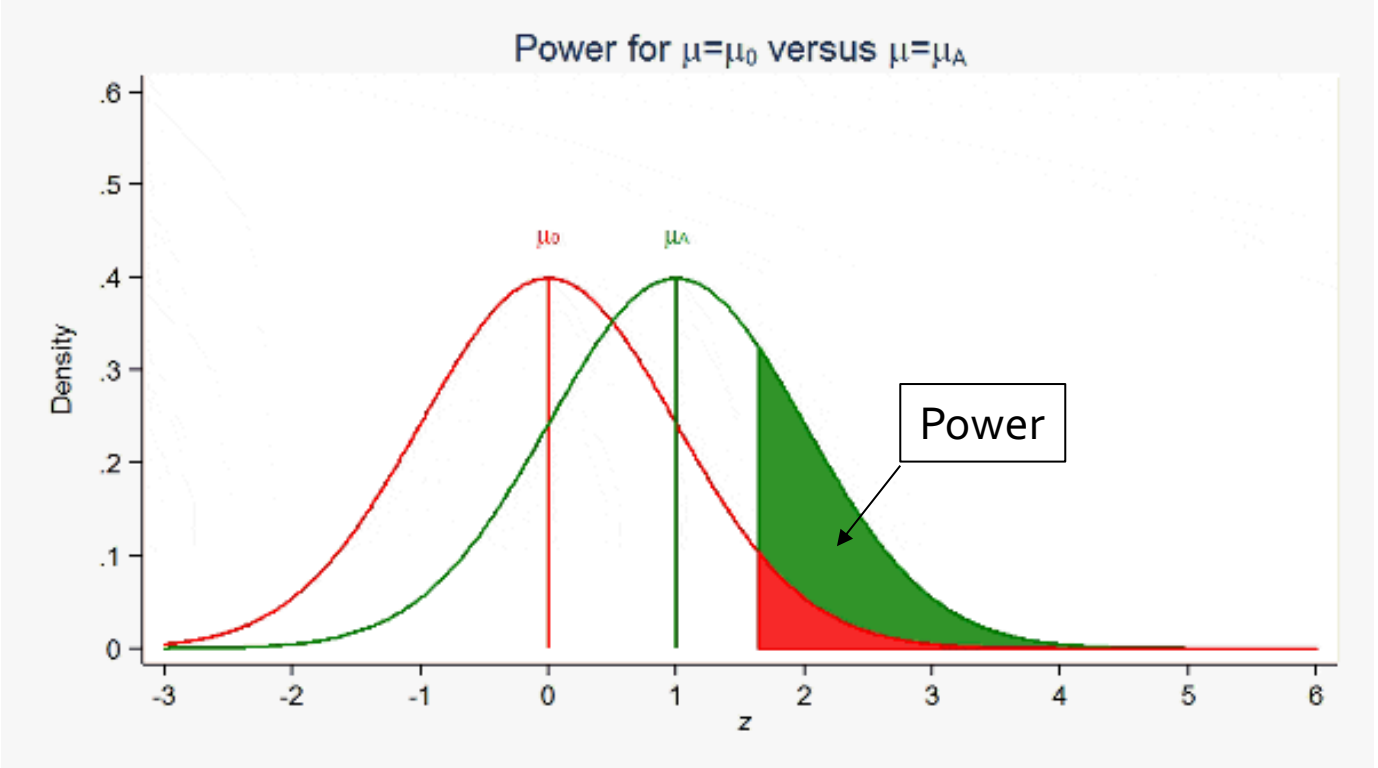
Factors influencing power
A **significance criterion** is a statement of how unlikely a positive result must be, if the null hypothesis of no effect is true, for the null hypothesis to be rejected
The **magnitude of the effect** of interest in the population can be quantified in terms of an effect size, where there is greater power to detect larger effects
The **sample size** determines the amount of sampling error inherent in a test result

- With three of the four known, calculate the unknown:

  significance level (α), effect size (d), sample size (n), power ($1 - \beta$)

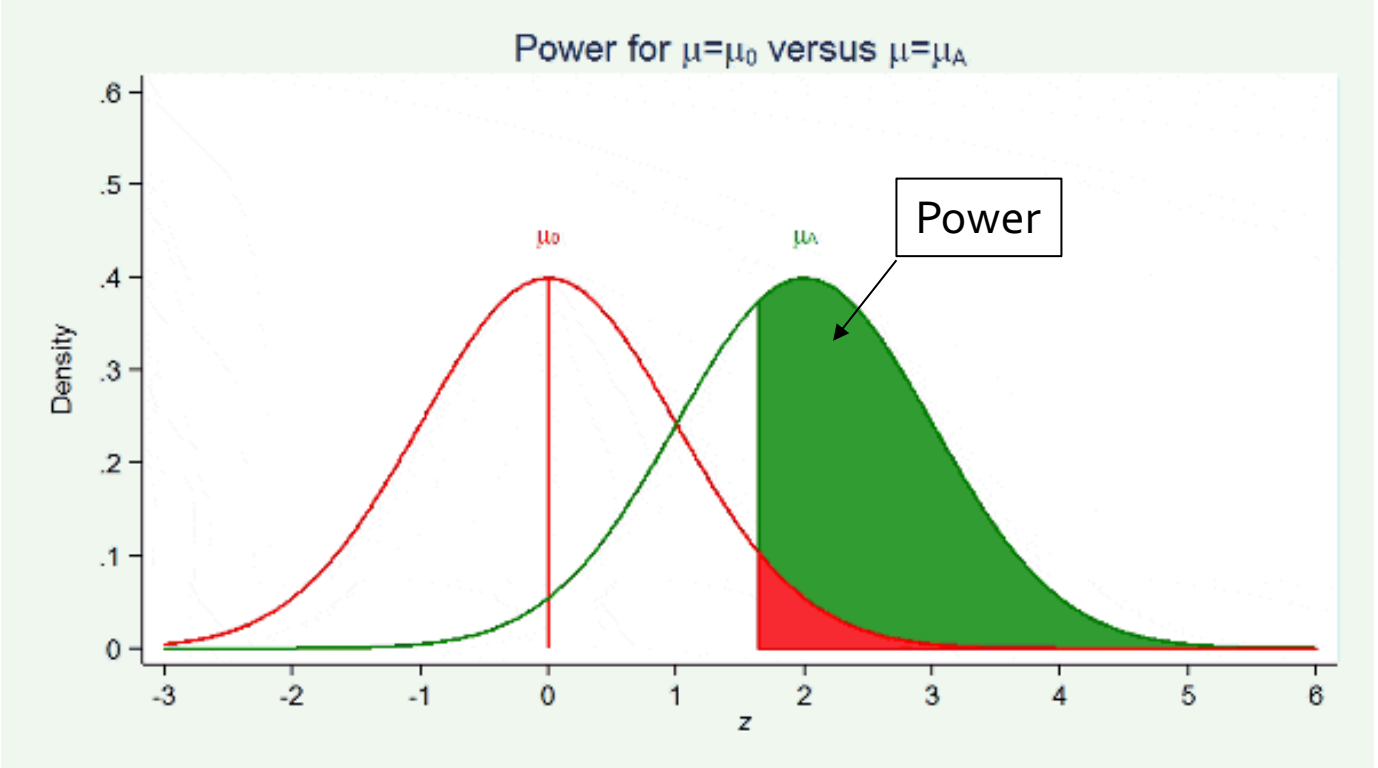
Power and Effect Size

- As effect size increases, power increases



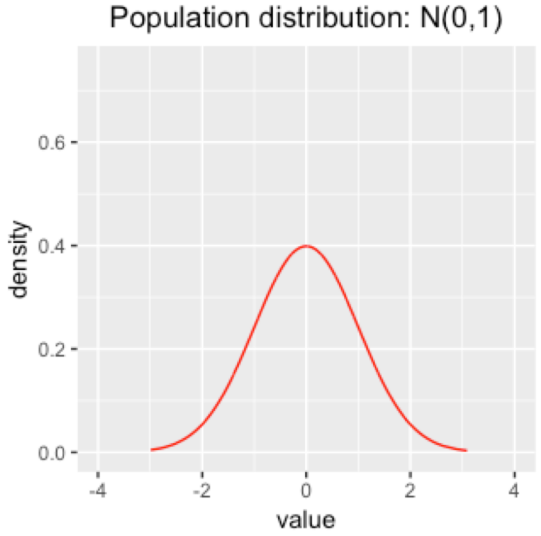
Power and Sample Size

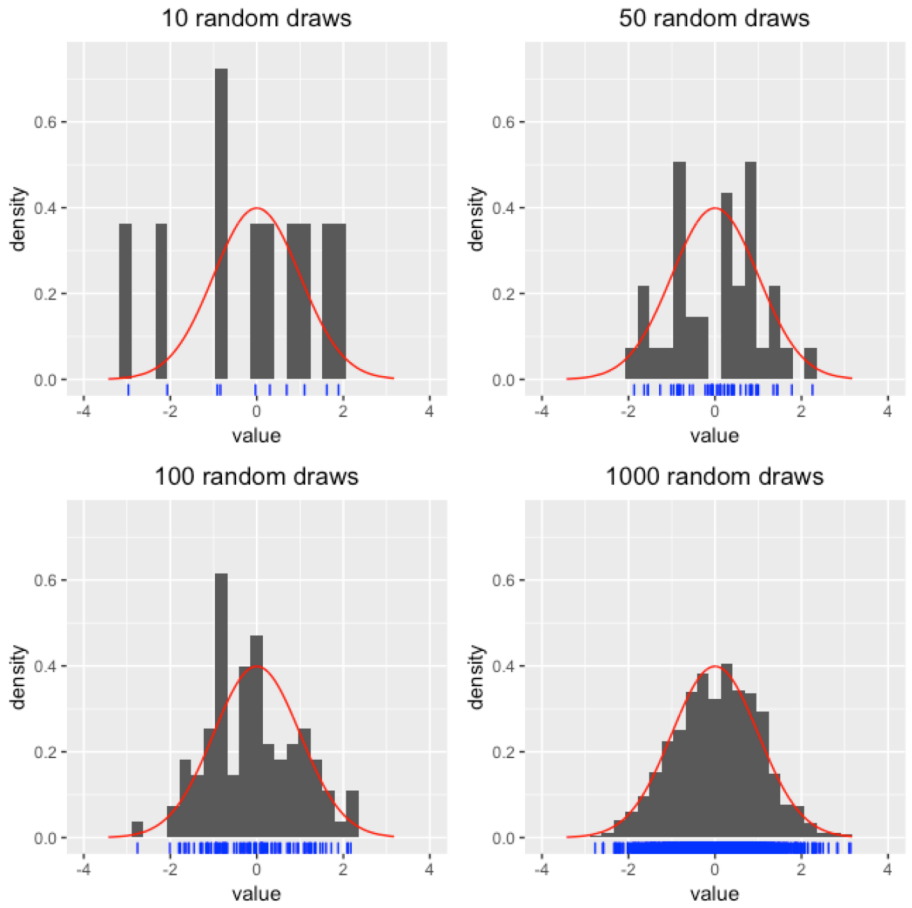
- As sample size increases, power increases



Sample Size

- A simple random sample is a subset of individuals (a sample) chosen from a larger set (a population)
- Sample sizes



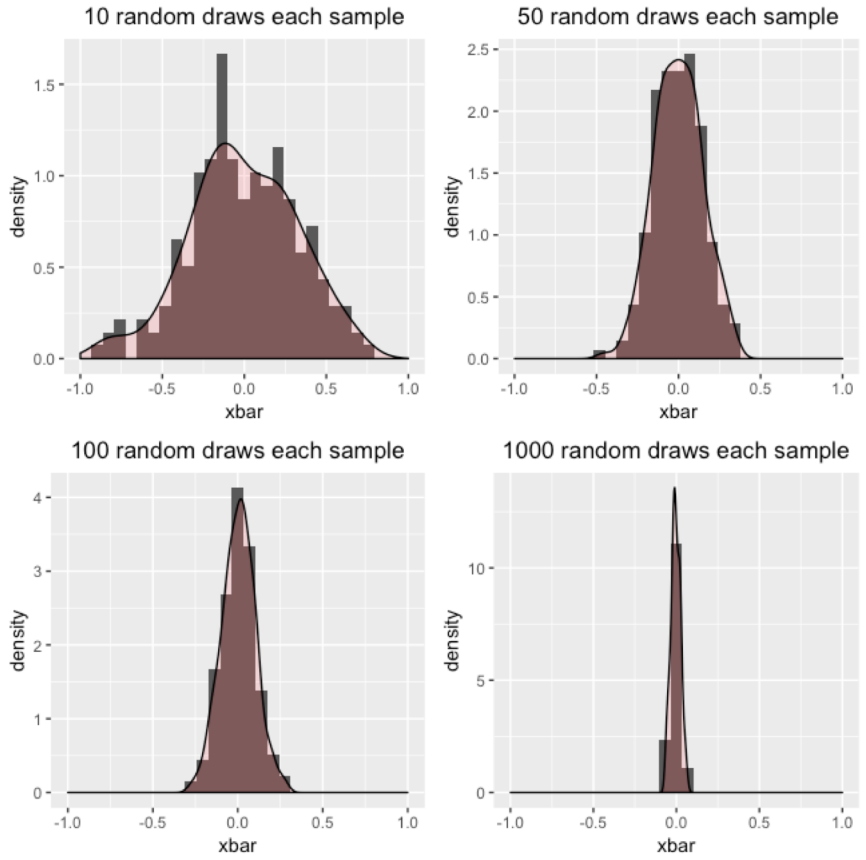


Large Sample Theory

- Central Limit Theorem

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ as } n \rightarrow \infty$$

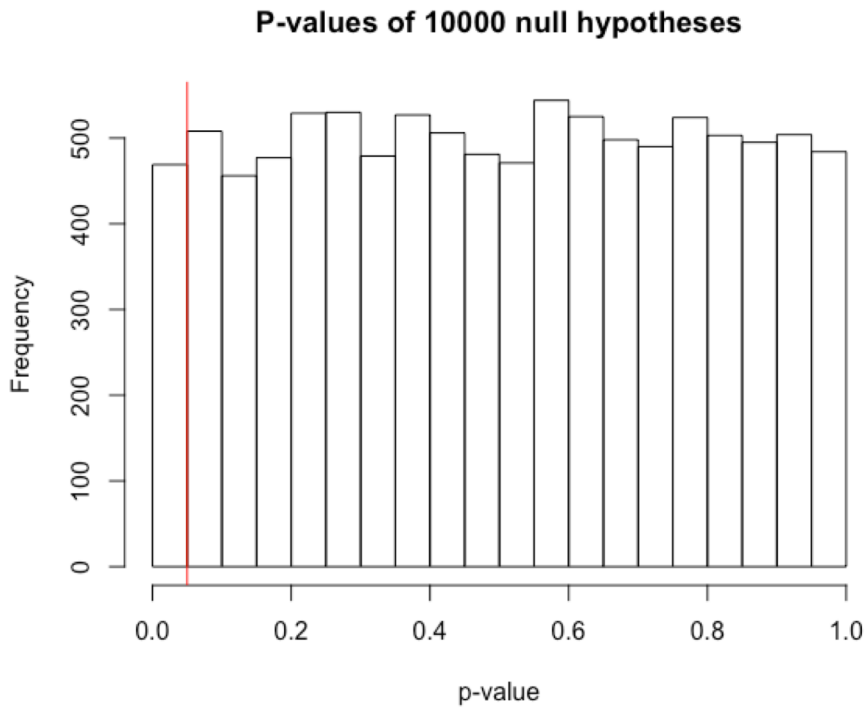
- Sample mean converges to population mean, as sample size ↑



Multiple Hypothesis Testing Correction

Intuition I

- Theoretically, p-values are **uniformly distributed** between 0 and 1 under the null hypothesis



- A typical DEG analysis might need to perform $10{,}000$ separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect 500 genes to be deemed "significant" by chance

Intuition II

- Recall: $\alpha = \Pr(\text{type I error})$
- If we perform m hypothesis tests, what is the probability of at least one false positive (an error)?

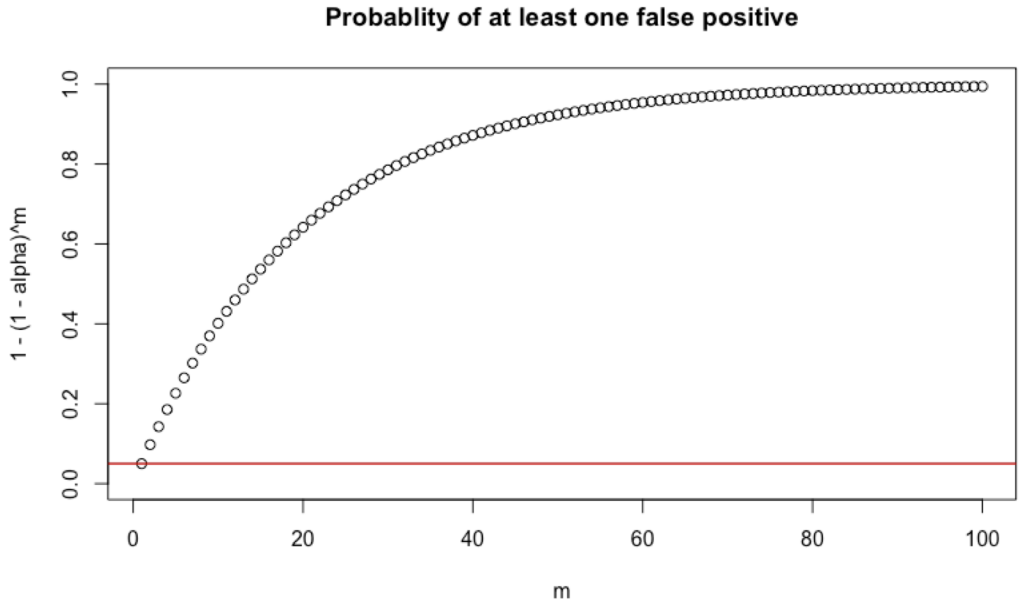
$$\Pr(\text{Making an error}) = \alpha$$

$$\Pr(\text{Not making an error}) = 1 - \alpha$$

$$\Pr(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$$

$$\Pr(\text{Making at least one error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

- If $m = 10000$ and $\alpha = 0.05$, then $1 - 0.95^{10000} = 1$



Counting Errors

- Assume we are performing m hypothesis tests: H^1 , H^2 , …, H^m
- m_0 = # of true hypotheses, R = # of rejected hypotheses

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_1)	Total
Test is declared non-significant	U	T	$m - R$
Test is declared significant	V	S	R
Total	m_0	$m - m_0$	m

- V = # of Type I error (also called "false discoveries")
- When people say "adjusting p-values for the number of hypothesis tests performed" what they mean is ***controlling the Type I error rate***

Approaches To Control Type I Errors

- **Per comparison error rate (PCER)**: the expected value of the number of Type I errors over the number of hypotheses

$$\text{PCER} = \mathrm{E}(V)/m$$

- **Per-family error rate (PFER)**: the expected number of Type I errors

$$\text{PFE} = \mathrm{E}(V)$$

- **Family-wise error rate (FWER)**: the probability of at least one type I error

$$\text{FWER} = \mathrm{P}(V \geq 1)$$

- **False discovery rate (FDR)**: the expected proportion of Type I errors among the rejected hypotheses

$$\text{FDR} = \mathrm{E}(V/R \mid R > 0)\mathrm{P}(R > 0)$$

- **Positive false discovery rate (pFDR)**: the rate that discoveries are false

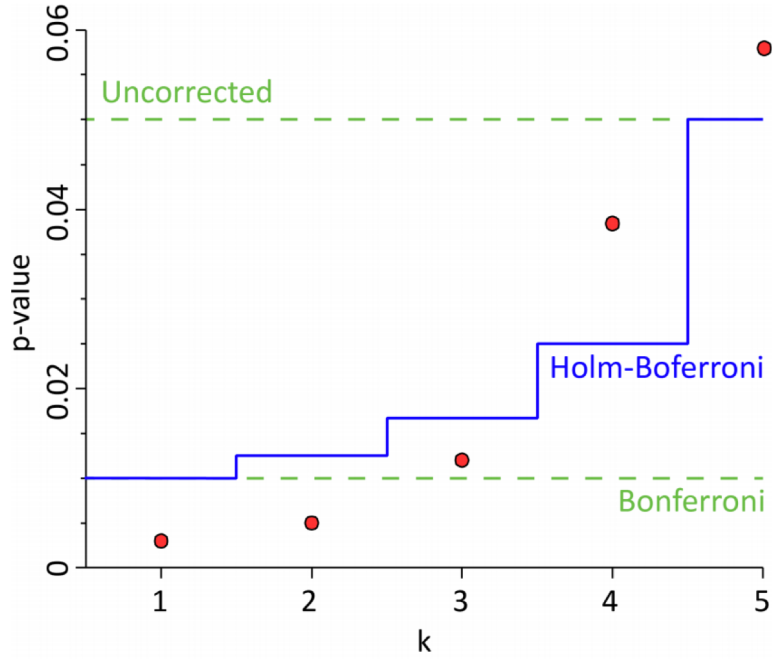
$$\text{pFDR} = \mathrm{E}(V/R \mid R > 0)$$

Controlling FWER

- FWER is appropriate when you want to guard against ANY false positives
- Two general types of FWER corrections:
  1. **Single step**: equivalent adjustments made to each p-value
  2. **Sequential**: adaptive adjustment made to each p-value
- **Bonferroni** (single step) rejects null hypothesis for each $p_i \leq \frac{\alpha}{m}$
- **Holm's** (sequential), also called Bonferroni-Holm
  - Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
  - Find the largest index k such that $p_{(k)} \leq \frac{\alpha}{m-k+1}$
  - Reject null hypotheses for $p_{(1)}, \cdots, p_{(k)}$
- Adjusted p-value:
  - Bonferroni: $p_i' = m \cdot p_i$
  - Holm's: $p'_{(1)} = m \cdot p_{(1)},\ p'_{(2)} = (m-1) \cdot p_{(2)},\ \cdots,\ p'_{(m)} = 1 \cdot p_{(m)}$
- High probability of type II errors, i.e. of not rejecting the general null hypothesis when important effects exist

Example

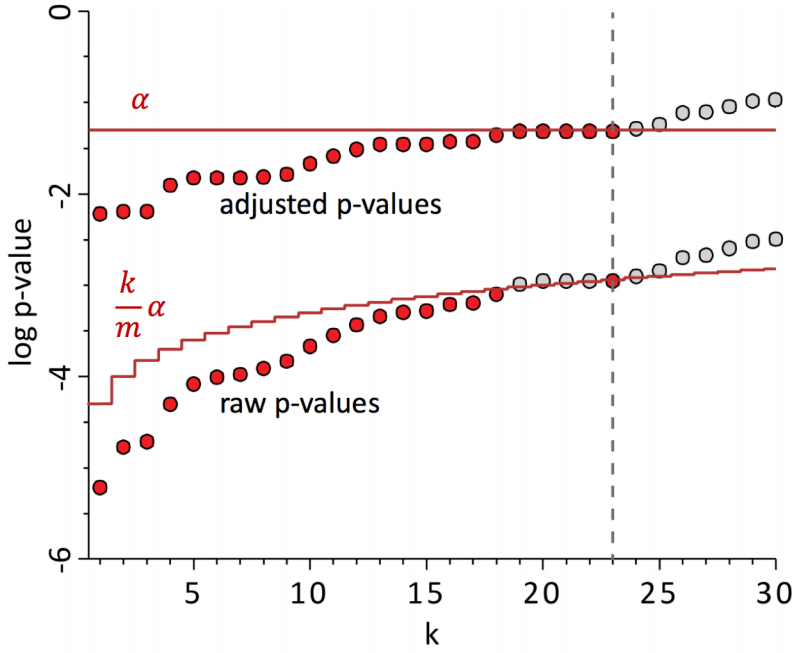
- Using raw p-value and correction threshold



k	p	α	$\frac{\alpha}{m}$	$\frac{\alpha}{m-k+1}$
1	0.003	0.05	0.01	0.01
2	0.005	0.05	0.01	0.0125
3	0.012	0.05	0.01	0.017
4	0.04	0.05	0.01	0.025
5	0.058	0.05	0.01	0.05

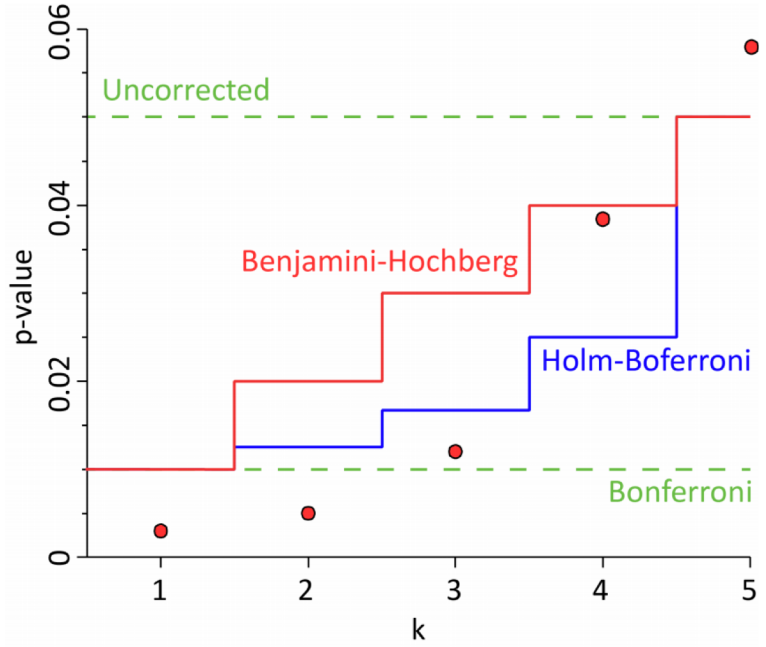
Controlling FDR

- False discovery rate (FDR) is designed to control the proportion of false positives ***among the set of rejected hypotheses*** (R)
- **Benjamini-Hochberg**
  - Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
  - Find the largest index k such that $p_{(k)} \leq \frac{k}{m}\alpha$
  - Reject null hypotheses for $p_{(1)}, \cdots, p_{(k)}$
- Adjusted p-value: $p'_{(i)} = \frac{m}{k} p_{(i)}$



Example

- Using raw p-value and correction threshold



k	p	α	$\frac{\alpha}{m}$	$\frac{\alpha}{m-k+1}$	$\frac{k}{m}\alpha$
1	0.003	0.05	0.01	0.01	0.01
2	0.005	0.05	0.01	0.0125	0.02
3	0.012	0.05	0.01	0.017	0.03
4	0.038	0.05	0.01	0.025	0.04
5	0.058	0.05	0.01	0.05	0.05

Storey's positive FDR (pFDR)

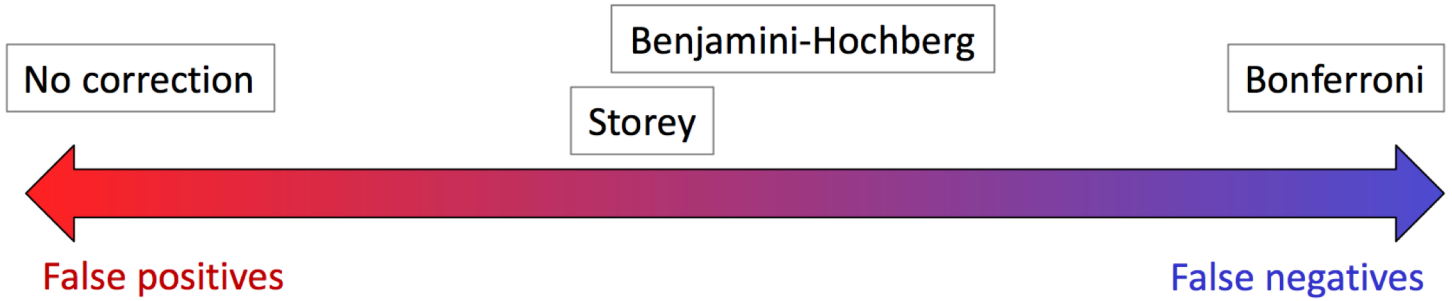
- Benjamini-Hochberg: FDR $= \mathrm{E}(V/R \mid R > 0)\mathrm{P}(R > 0)$
- Storey: pFDR $= \mathrm{E}(V/R \mid R > 0)$
- Since $P(R > 0) \approx 1$ in most high-throughput experiments, FDR and pFDR are very similar
- Omitting $P(R > 0)$ facilitated development of a measure of significance in terms of the FDR for each hypothesis

Estimating FDR: $\widehat{FDR}(p_i)$

- Q-value is defined as the minimum FDR that can be attained when calling that "feature" significant (i.e., expected proportion of false positives incurred when calling that feature significant)
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)
- Thus, in a study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% *of genes that show p-values at least as small as gene X are false positives* (Nice interpretation!)

Summary

Which multiple-test correction should I use?



False positive	False negative
"Discover" effect where there is no effect	Missed discovery
Can be tested in follow-up experiments	Once you've missed it, it's gone
Not hugely important in small samples	
Impossible to manage in large samples	

http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/p-values1/p-values8.pdf

Summary

Multiple test procedures: summary

Method	Controls	Advantages	Disadvantages	Recommendation
No correction	FPR	False negatives not inflated	Can result in $FP \gg TP$	Small samples, when the cost of FN is high
Bonferroni	FWER	None	Lots of false negatives	Do not use
Holm-Bonferroni	FWER	Slightly better than Bonferroni	Lots of false negatives	Appropriate only when you want to guard against any false positives
Benjamini-Hochberg	FDR	Good trade-off between false positives and negatives	On average, α of your positives will be false	Better in large samples
Storey	--	More powerful than BH, in particular for small $\hat{\pi}_0$	Depends on a good estimate of $\hat{\pi}_0$	The best method, gives more insight into FDR

http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/p-values1/p-values8.pdf