Stony Brook Medicine

From P-value To FDR

Jie Yang, Ph.D.

Associate Professor
Department of Family, Population and Preventive Medicine
Director
Biostatistical Consulting Core

In collaboration with Clinical Translational Science Center (CTSC) and
the Biostatistics and Bioinformatics Shared Resource (BB-SR), Stony Brook Cancer Center (SBCC).

OUTLINE

❑ P-values

- What is p-value?
- How important is a p-value?
- Misinterpretation of p-values

❑ Multiple Testing Adjustment

- Why, How, When?
- Bonferroni: What and How?
- FDR: What and How?

STATISTICAL MODELS

- **Statistical model** is a mathematical representation of data variability, ideally catching all sources of such variability.
- All methods of statistical inference have assumptions about
  - How data were collected
  - How data were analyzed
  - How the analysis results were selected for presentation
- Assumptions are often simple to express mathematically, but difficult to satisfy and verify in practice.

- Hypothesis test is the predominant approach to statistical inference on effect sizes which describe the magnitude of a quantitative relationship between variables (such as standardized differences in means, odds ratios, correlations etc).

BASIC STEPS IN HYPOTHESIS TEST

1. State null (H_0) *and* alternative (H_1) hypotheses

2. Choose a significance level, α (usually 0.05)

3. Based on the sample, calculate the test statistic and calculate p-value based on a theoretical distribution of the test statistic

4. Compare p-value with the significance level α

5. Make a decision, and state the conclusion

HISTORY OF P-VALUES

- P-values have been in use for nearly a century.
- The *p*-value was first formally introduced by **Karl Pearson**, in his Pearson's chi-squared test and popularized by **Ronald Fisher**.
- In his influential book *Statistical Methods for Research Workers* (1925), Fisher proposed the level *p* = 0.05, or a 1 in 20 chance of being exceeded by chance, as a limit for statistical significance.

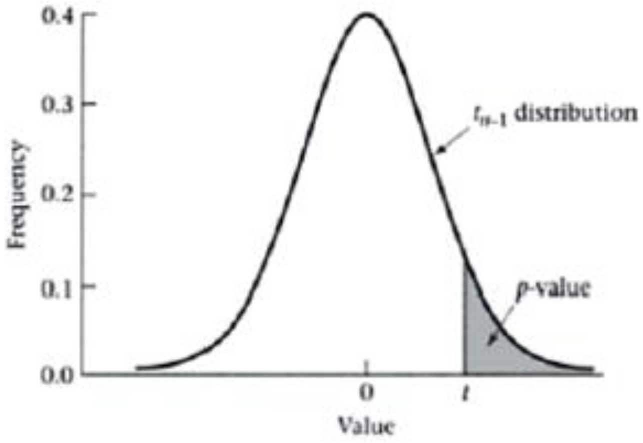
Karl Pearson, 1857-1936
English mathematician
and Statistician

Ronald A. Fisher, 1890-1962
English mathematician and
Statistician

P-VALUE

<u>Definition:</u> the **probability** of obtaining a test statistic, which measures the distance between expected and observed data patterns, as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

- Also called **observed significance level,** the α level at which we would be indifferent between failing to reject and rejecting H_0 given the sample data at hand
- It is a statistical summary of the compatibility between observed data and what we expect to see if the entire statistical model (all assumptions used to compute the p-value) were correct.



WHAT A P-VALUE CAN TELL

- It is a continuous measure with 0 for completely incompatibility between data and the model used to compute p-value and 1 for complete compatibility.

- The smaller the p-value, the more unusual the sample data would be **if every single assumption were correct**

- A reported small p-value may be because:
  1. The alternative hypothesis is true
  2. The study protocols were violated so some key assumption is wrong
  3. It is selected for presentation because it is small!

MISINTERPRETATION OF P-VALUE

1. P-value is the probability that the null hypothesis is true. If a hypothesis test has a p-value of 0.01, the null hypothesis only has 1% of chance of being true.

- The calculation of p-value is from the assumption that the null hypothesis is true. It simply indicates the degree to which the data conform to the pattern predicted by alternative hypothesis and all the other assumptions used in the test. A p-value of 0.01 only indicate that the data are not very close to what the statistical model predicated they should be.

2. A nonsignificant test result (p-value > 0.5) means that the null hypothesis is true or a large p-value is evidence in favor of the null hypothesis.

- P-value >0.05 only means that a discrepancy from the null hypothesis would be as large or larger than observed more than 5% of the time *if only* chance were creating the discrepancy.

MISINTERPRETATION OF P-VALUE

3. A p-value <0.05 indicates a scientifically or substantively important relation has been detected.

Statistical significance ≠ clinical/scientific significance

- For a large study, very minor effects or small assumption violations can lead to statistically significant tests of the null hypothesis or small p-values. Again, a small p-value simply flags the data as being unusual if all the assumptions used to compute it including null hypothesis were correct; but the way the data are unusual might be of no clinical interest.

Statistical significance ≠ clinical/scientific significance

	P-value=0.001	P-value=0.21	P-value=0.001
Statistical Perspective	WOW	~~WOW~~	WOW
Clinical Perspective	WOW	WOW	~~WOW~~
	e.g. Mean BMI dropped from 45 to 30	e.g. Mean BMI dropped from 45 to 30	e.g. Mean BMI dropped from 45 to 44.8

- One must look at confidence interval to determine which effect sizes of clinical/scientific importance are relatively compatible with data, given all other assumptions.

MISINTERPRETATION OF P-VALUE

4. A large p-value indicates that the effect size is small.

- When a study is small, even large effect sizes may be “drowned in noise” and hence fail to be detected by a hypothesis test or has a large p-value.

- Again, one must look at the confidence interval to determine if it includes the effect sizes of importance.

5. If you reject the null hypothesis because $p<0.05$, the chance you are in making a type I error is 5%.

- The chance of making a type I error is 100% if H_0 is really true. The 5% refers only to how often you would reject H_0 when H_0 is true over many uses of the tests across different studies when the test hypothesis and all other assumptions used for the test are true. It does not refer to your single use of the test.

MISINTERPRETATION OF P-VALUE

6. When the same hypothesis is tested in different studies and all studies reported large p-values, the overall evidence supports the null hypothesis.

- In practice, every study could fail to reach statistical significance and yet when combined show a statistical significance. For example, if there were 5 studies each with p-value=0.1, the combined p-value using Fisher's formula, the overall p-value would be 0.01.

7. If one observes a small p-value, there is a good chance that the next study will produce a p-value at least as small for the same hypothesis.

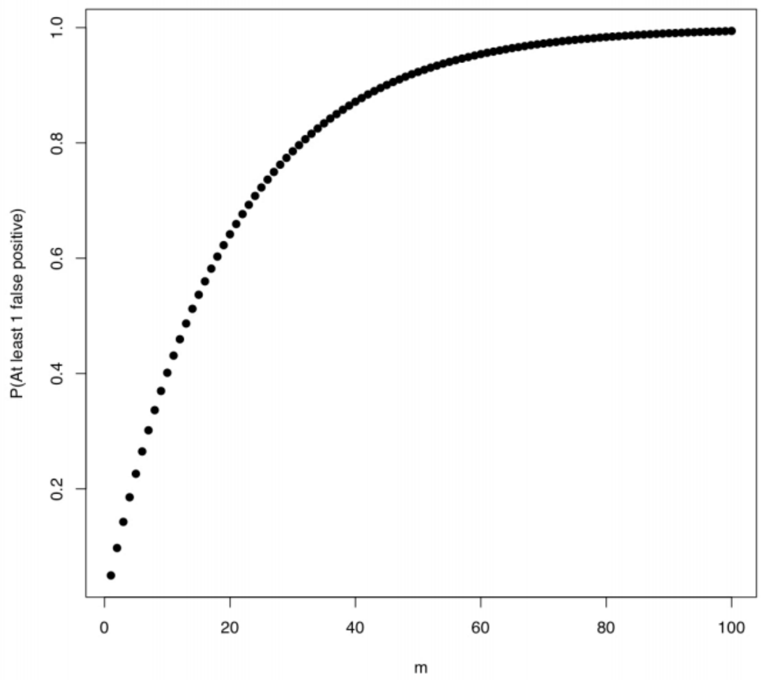
- The size of new p-value is extremely sensitive to the study size and the extent to which the null hypothesis or other assumptions are violated in the new study. It may be much smaller or much larger.

MULTIPLE TESTING ISSUES

- A typical microarray experiment might result in performing 10000 separate hypothesis tests. If a significance level is set at 0.05, 500 genes are expected to be deemed “significant” by chance.

- In general, the probability of making at least 1 false positive while performing m hypothesis test is approximated by $1-(1-\alpha)^m$

The number of hypothesis tests, m	Probability of making at least one false positive
1	0.05
2	0.0975
3	0.1426
4	0.1855
5	0.2262



WHAT DOES CORRECTING FOR MULTIPLE TETING MEAN

- “adjusting p-values for the number of hypothesis tests performed” means to control the Type I error rate.

- Very active area in statistics - many different methods have been proposed.

- Although these proposed approaches have the same overall goal, they handle the multiple testing issue in fundamentally different ways.

# OF ERROR DECISIONS

Suppose totally *m* hypotheses are tested:

	H0 is true	H1 is true	Total
Fail to reject	**U**	**T**	**m-R**
Reject	**V**	**S**	**R**
	m_0	**$m-m_0$**	**m**

❑ **m_0** = # of true null hypothesis
❑ **R** = # of rejected null hypothesis
❑ **V** = # of type I errors (false positive)

APPROACH TO MULTIPLE TESTING ADJUSTMENT

1. **Family-wise Error Rate (FWER)**: the probability of at least one Type I error

$$FWER = P(V>=1)$$

2. **False Discovery Rate (FDR)**: the expected proportion of Type I errors among the rejected hypotheses

$$FDR = E(V/R|R>0)P(R>0)$$

- **positive false discovery rate (pFDR)**: the rate that discovery are false – $pFDR=E(V/R|R>0)$

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

   - **Bonferroni Adjustment**

2. *Sequential*: adaptive adjustment made to each p-value

   - **Holm's Method**

BONFERRONI ADJUSTMENT

- Very simple method for ensuring that the overall Type I error rate of α is maintained when performing *m* (independent) hypothesis tests

- Rejects any hypothesis with **p-value ≤ α/m**. Or use
  adjusted p-value= min(m*p-value, 1)

- For example, if we want to have an experiment-wide Type I error rate of 0.05 when we perform 10,000 hypothesis tests, we'd need a p-value of 0.05/10000 = 0.000005 or smaller to declare significance

- **Note**: interpretation of finding depends on the number of other tests performed.

BONFERRONI ADJUSTMENT

- Bonferroni adjustment is conservative

When rejecting H0 when p-value < 0.0025 among all 20 tests, assuming all tests are independent of each other,

$$\begin{aligned} P(\text{at least one significant result}) &= 1- P(\text{no significant results}) \\ &= 1- (1-0.0025)^{20} \\ &\sim 0.0488 \mathbf{< 0.05} \end{aligned}$$

- In practice, tests may be correlated. Depending on the correlation structure of all tests, Bonferroni adjustment could lead to a high rate of false negatives.

HOLM'S METHOD

- Order the unadjusted p-values such that p1 ≤ p2 ≤ … ≤ pm
- For control of the FWER at level α, the step-down Holm adjusted p-values (j=1,….m) are

$$\widetilde{p_j} = \min[(m - j + 1) * p_j, 1]$$

- The point here is that we don't multiply every p_j by the same factor m!

For example, when doing 10000 hypotheses tests:

$$\widetilde{p_1} = 10000 * p_1;\ \widetilde{p_2} = 9999 * p_2; \dots ; \widetilde{p_m} = 1 * p_m$$

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*. **6** (2): 65–70.

FALSE DISCOVERY RATE (FDR)

- What if not caring about making ANY Type I errors? For example, in genomics studies, a certain number of false positives are tolerable.
- The more relevant error rate to control is the false discovery rate (FDR).
- FDR is designed to control the proportion of false positives **among the set of rejected hypotheses (R)**

	H0 is true	H1 is true	Total
Fail to reject	U	T	m-R
Reject	V	S	R
	m_0	$m-m_0$	m

FDR VS FPR

	H0 is true	H1 is true	Total
Fail to reject	U	T	m-R
Reject	V	S	R
	m_0	$m-m_0$	m

False Discovery Rate:

$$FDR = \frac{V}{R}$$

False Positive Rate (Type I error) :

$$FPR = \frac{V}{m_0}$$

BENJAMINI AND HOCHBERG (BH) FDR

To control FDR at level δ -

1. Order the unadjusted p-values:

   $p(1) \le p(2) \le \ldots \le p(m)$

2. Then find the test with the highest rank, j, for which the p value, pj,

$$p(j)<=(j/m) \times \delta$$

3. Declare the tests of rank 1, 2, …, j as significant

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 57, No. 1, pp. 289-300

B&H FDR EXAMPLE

Controlling the FDR at δ = 0.05, m=10

Rank (j)	Unadjusted P-value	(j/m)*δ	Reject H0?
1	0.0004	0.005	Yes
2	0.008	0.010	Yes
3	0.0123	0.015	Yes
4	0.1211	0.020	No
5	0.2301	0.025	No
6	0.2678	0.030	No
7	0.3455	0.035	No
8	0.4681	0.040	No
9	0.6788	0.045	No
10	0.911	0.05	No

STOREY'S POSITIVE FDR (PFDR)

$$\text{BH}: FDR = E\left[\frac{V}{R} \mid R > 0\right] P(R > 0)$$

$$\text{Storey}: \text{p}FDR = E\left[\frac{V}{R} \mid R > 0\right]$$

- Since P(R > 0) is ~ 1 in most genomics 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

Storey, John D. (2002). "A direct approach to false discovery rates" (PDF). *Journal of the Royal Statistical Society, Series B*. **64** (3): 479–498.

Q-VALUE

- **q-value** is defined as the **minimum FDR** that can be attained when calling that test significant (i.e., expected proportion of false positives incurred when calling that test 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, PNAS, 2003)

- For example, in GWAS, if gene X has a q-value of 0.013, it means that 1.3% of genes that show p-values smaller or at least as small as gene X are false positives.

Q-VALUE EXAMPLE

m=10

Rank (j)	Unadjusted P-value	(j/m)*δ	Reject H0?	Q-value*
1	0.0004	0.005	Yes	0.0019
2	0.008	0.010	Yes	0.0191
3	0.0123	0.015	Yes	0.0196
4	0.1211	0.020	No	0.1447
5	0.2301	0.025	No	0.2301
6	0.2678	0.030	No	0.2678
7	0.3455	0.035	No	0.3455
8	0.4681	0.040	No	0.6810
9	0.6788	0.045	No	0.6788
10	0.911	0.05	No	0.9110

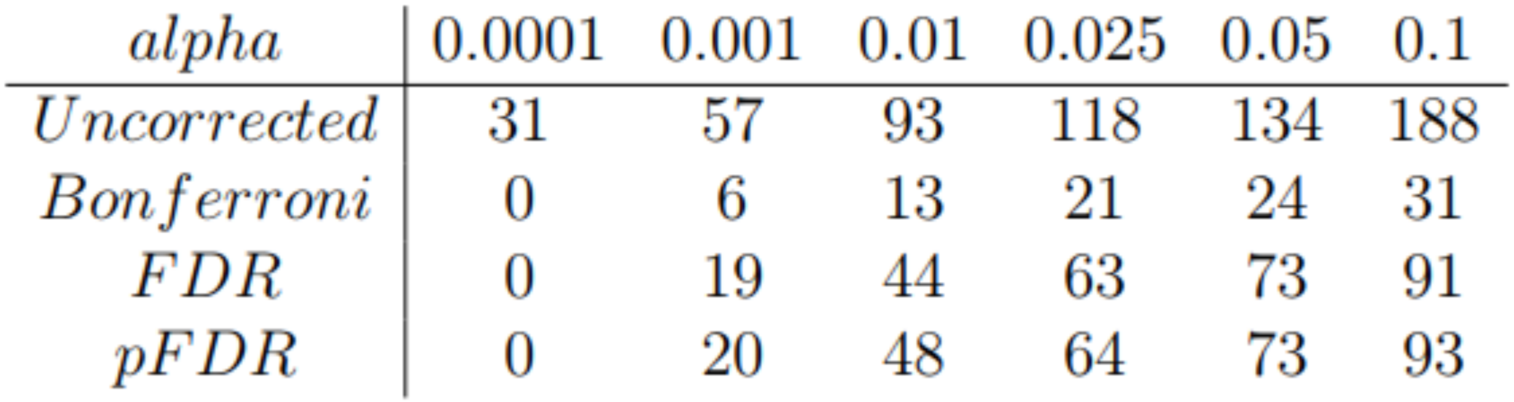
*Q-value calculated using Proc Multtest in SAS 9.4 with option pFDR.

COMPARISON OF BONFERRONI, FDR &PFDR

A simulation study to compare Bonferroni Adjustment, FDR and pFDR

- Simulate first 900 sets of data from a standard normal distribution N(0,1), the next 100 sets of data from a normal distribution with mean at 3.
- Hypothesis test: H_0: mean=0
- So out of 1000 tests, theoretically first 900 tests shouldn't reject H_0 but the rest 100 tests should reject H_0.

# of significant calls vs different alpha/FDR level

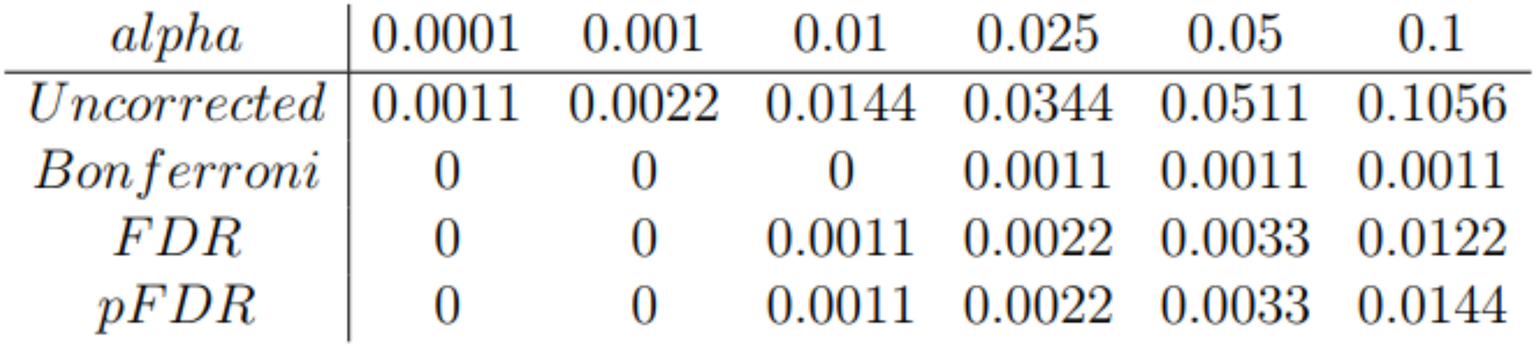


alpha	0.0001	0.001	0.01	0.025	0.05	0.1
Uncorrected	31	57	93	118	134	188
Bonferroni	0	6	13	21	24	31
FDR	0	19	44	63	73	91
pFDR	0	20	48	64	73	93

Stony Brook Medicine

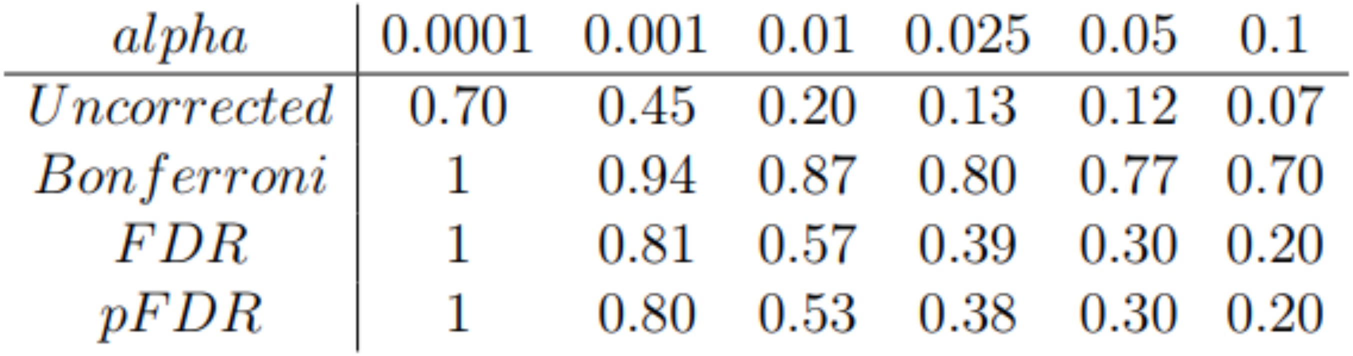
COMPARISON OF BONFERRONI, FDR &PFDR

True Type I error rate vs different alpha/FDR level



alpha	0.0001	0.001	0.01	0.025	0.05	0.1
Uncorrected	0.0011	0.0022	0.0144	0.0344	0.0511	0.1056
Bonferroni	0	0	0	0.0011	0.0011	0.0011
FDR	0	0	0.0011	0.0022	0.0033	0.0122
pFDR	0	0	0.0011	0.0022	0.0033	0.0144

True Type II error rate vs different alpha/FDR level



alpha	0.0001	0.001	0.01	0.025	0.05	0.1
Uncorrected	0.70	0.45	0.20	0.13	0.12	0.07
Bonferroni	1	0.94	0.87	0.80	0.77	0.70
FDR	1	0.81	0.57	0.39	0.30	0.20
pFDR	1	0.80	0.53	0.38	0.30	0.20

WHEN TO USE MULTIPLE TESTING ADJUSTMENT

In general,

- “Adjustment for multiple testing are **REQUIRED for confirmatory studies** whenever results from multiple tests have to be combined in one final conclusion and decision.” (Bender and Lange, 2001)

- For **exploratory** analysis, adjustments for multiple comparisons are **not strictly required** since the findings are not conclusive and mainly hypothesis-generating.

- In GWAS or large scale hypothesis testing, multiple testing adjustment is recommended.

Bender R and Lange S. (2001) “ Adjusting for Multiple Testing – When and How?”. Journal of Clinical Epidemiology. 54:343-9.

OTHER SPECIFIC MULTIPLE TESTING ADJUSTMENT

Multiple groups

Comparison of the means of several groups in analysis of variance (ANOVA) :

- ➢ Simultaneous test procedures for all pairwise comparisons: *Scheffé (unequal sample size)* and *Tukey(equal sample size)*
- ➢ Compare several groups with a single control: *Dunnett*
- ➢ Multiple stage (Stepdown) tests to give homogenous sets of treatment means but no simultaneous CIs: *Ryan-Einot-Gabriel-Welsch (REGW).*

To control FWER, REGW is recommended for a balanced design and no CIs are needed. Otherwise, Tukey's procedure is appropriate (Bender and Lange, 2001) .

OTHER SPECIFIC MULTIPLE TESTING ADJUSTMENT

Multiple endpoints

- One of most common multiplicity problems in clinical trials
- Strategies to deal with this:
  1. Specify one *single primary endpoint*
  2. Combine outcomes in *one aggregated endpoint*
  3. *Multivariate* methods [e.g. multivariate analysis of variance (MANOVA) or Hotelling’s T test] or *global test statistics*
     - Only overall assessment of effects provided through statistical significance
     - Information concerning the individual endpoints is lacking.

OTHER SPECIFIC MULTIPLE TESTING ADJUSTMENT C

Repeated Measurements

- Difficult to develop a general adjustment method for multiple comparisons occur for between-subject factors (e.g. groups), within-subject factors (e.g. time), or both because the specific correlation structure has to be taken into account.

- Strategies:
  1. Treat repeated measurements as multiple endpoints if only comparisons for between-subject factors are of interest.
  2. For longitudinal measurements, may consider use of summary measures such as area under curve to describe the response curves.

OTHER SPECIFIC MULTIPLE TESTING ADJUSTMENT

Interim Analysis

- Long term clinical trials allow for early stopping for efficacy or futility;
- Multiple testing adjustment is **required** because of possible inflated Type-I error.
- *Simple rule*: p-value < 0.01 to have early stopping for efficacy and final test if no more than 10 interim analyses are planned.
- *Another simple rule*: use p-value<0.001 for interim analysis for any number of interim analysis and final analysis at p-value<0.05.
- *O'Brien and Fleming*: use varying nominal significance level for early stopping – stringent sig. level at early interim analysis and final analysis use a sig. level close to 0.05.

IN SUMMARY

- ✓ A lower p-value provides more convincing evidence against the null hypothesis.
- ✓ P-values are often misinterpreted and provide no information on the magnitude or importance of the effect.
- ✓ Confidence intervals are superior to p-values because it shows the full range of effect sizes compatible with data.
- ✓ Multiple testing adjustment depends on which type of error rate to control. Often adequate control of Type I error is quite complex.
- ✓ Bonferroni adjustment is the simplest method to correct for multiple testing issue, but it is the most conservative.
- ✓ FDR and pFDR controls false discovery rate, not Type I error.

A video about p-value: https://www.youtube.com/watch?v=ax0tDcFkPic&t=8s
A video about p-value vs CI: https://www.youtube.com/watch?v=8-PzD26WI4g

Please check our website for future lectures
https://osa.stonybrookmedicine.edu/research-core-facilities/bcc/education

Coming ones:

- April 4, performing basic statistical tests using different software
- April 17, sample size calculation

THANK YOU!