Multiple Comparisons & Simultaneous Inference

Methods and Advice

Dan Hall, Director of the SCC

Department of Statistics
Franklin College of Arts and Sciences
Statistical Consulting Center
UNIVERSITY OF GEORGIA

Table of Contents

Related Resources

- A companion video for this talk can be found here kaltura.uga.edu/media/t/1_l5t793g7.
- An accompanying R script, multCompExams.R, is available as an attachment to the video linked above. Follow the link and click on attachments.
- The video shows how to implement some well-known multiple comparison methods in R and uses simulated data to illustrate various different error rates and the methods that do (and don't) control them.

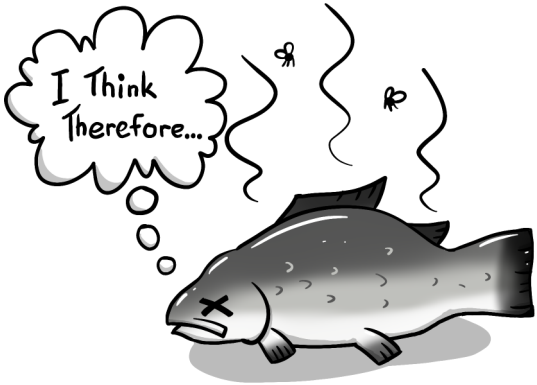
Related Resources

- A companion video for this talk can be found here kaltura.uga.edu/media/t/1_l5t793g7.
- An accompanying R script, multCompExams.R, is available as an attachment to the video linked above. Follow the link and click on attachments.
- The video shows how to implement some well-known multiple comparison methods in R and uses simulated data to illustrate various different error rates and the methods that do (and don't) control them.

Related Resources

- A companion video for this talk can be found here kaltura.uga.edu/media/t/1_l5t793g7.
- An accompanying R script, multCompExams.R, is available as an attachment to the video linked above. Follow the link and click on attachments.
- The video shows how to implement some well-known multiple comparison methods in R and uses simulated data to illustrate various different error rates and the methods that do (and don't) control them.

The Problem



Bennet Et Al. (2010, *Journal of Serendipitous and Unexpected Results*)

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Hypothesis testing:
  - Assume a null hypothesis H_0 is true,
  - gather evidence (data),
  - summarize evidence against H_0 (test statistic),
  - quantify strength of the evidence (p-value), and
  - make a decision (reject, fail to reject).
- Possible outcomes:

	The Truth	
Decision	H_0 is true	H_0 is false
Don't reject	Correct	Type II Error
Reject	Type I Error	Correct

$$\alpha = \Pr(\text{Type I Error})$$
$$\beta = \Pr(\text{Type II Error}) = 1 - \text{Power}$$

- α and β are negatively related to one another.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Tests and Confidence Intervals

- Typically, we fix α (the significance level) at 0.05, say, to make a Type I error unlikely.
- Smaller α typically means larger β. Rather acquit a guilty defendant than convict an innocent one.
  - Small α is "safe" strategy when we wish to be cautious about rejecting H_0 (sometimes safer to reject, though, as in a model diagnostic test).
- Confidence intervals and tests are flip-sides of same coin.
  - $\alpha = 0.05$-level test corresponds to 95% confidence interval.
  - 5% Type I error for a test corresponds to 95% coverage probability for an interval.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- Tests and CIs are designed to have a given Type I error rate or coverage probability.
- Those rates apply to one inference (test or interval) at a time.
- If we conduct two tests, each at level 0.05, the probability that at least one is falsely significant is > 0.05.
  - That is, the combined Type I error (probability of at least one Type I error) for multiple tests is greater than that of a single test.
- Similarly, one interval may have coverage probability 95%. But the probability that *two* intervals *both* cover their respective parameters (the simultaneous coverage probability) will be < 0.95.
- Simple principle: the more chances you have to make a mistake, the more likely you will eventually make one.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Simultaneity

- This issue is known as
  - the problem of simultaneous inference, or *simultaneity*,
  - AKA *multiplicity* or, in the context comparing means (e.g., treatment means in a designed experiment), the *multiple comparisons problem.*
- Most analyses involve conducting more than one inference, so simultaneity arises all the time.
- *Should we be concerned about it? Should we adjust for it?*
- When analyzing designed experiments with ANOVA models, the consensus is **yes**, and methods are well-developed.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Multiple Comparison Methods

Consider a one-way ANOVA with $a = 5$ treatments (5 levels of treatment factor A).

- The ANOVA yields an F test of

$$H_0 : \mu_1 = \cdots = \mu_5 \quad \text{vs.} \quad H_A : \{\text{not } H_0\}$$

- This test of the *main effects of A* doesn't tell us very much.
- If we reject H_0, this does not mean $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$! Still must determine which means differ.
- If fail to reject H_0, this does not mean H_0 is true!
  - Still possible to find one or more significant differences among the μ_i's.
- So testing more specific hypotheses about differences among the treatment means is usually well-motivated.

Demo

```
set.seed(20923); library(emmeans)
a <- 5; n <- 6 # 5 trts, 6 reps/trt
N <- a*n     # sample size
err <- rnorm(N,0,1)
trtMeans <- c(8.8,11.2,10,10,10); trt <- 1:a; repl <- 1:n
nullData <- within(expand.grid(rep=repl,trt=trt),{
  trtFac <- factor(trt); y <- trtMeans[trt] + err
})
m1 <- aov(y~trtFac,data=nullData); anova(m1)[1,]
```

```
Analysis of Variance Table

Response: y
       Df Sum Sq Mean Sq F value Pr(>F)
trtFac  4 6.2257  1.5564  1.5297  0.224
```

```
contrast(emmeans(m1,specs=~trtFac),method=list(trt1.Vs.trt2=c(1,-1,0,0,0)))
```

```
 contrast     estimate    SE df t.ratio p.value
 trt1.Vs.trt2    -1.26 0.582 25  -2.166  0.0401
```

Contrasts

Comparisons among means are done via **contrasts**.

- Examples:
  - Pairwise contrasts: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, etc.
  - Contrasts need not be pairwise. E.g., suppose treatments 1 & 2 are Drug I delivered via pill and capsule, and treatments 3, 4, 5 are drug II via pill, capsule, and oral suspension (liquid). Might want to compare Drug I vs Drug II via

$$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4 + \mu_5}{3}.$$

Contrasts

Comparisons among means are done via **contrasts**.

- Examples:
  - Pairwise contrasts: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, etc.
  - Contrasts need not be pairwise. E.g., suppose treatments 1 & 2 are Drug I delivered via pill and capsule, and treatments 3, 4, 5 are drug II via pill, capsule, and oral suspension (liquid). Might want to compare Drug I vs Drug II via

$$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4 + \mu_5}{3}.$$

Contrasts

Comparisons among means are done via **contrasts**.

- Examples:
  - Pairwise contrasts: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, etc.
  - Contrasts need not be pairwise. E.g., suppose treatments 1 & 2 are Drug I delivered via pill and capsule, and treatments 3, 4, 5 are drug II via pill, capsule, and oral suspension (liquid). Might want to compare Drug I vs Drug II via

$$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4 + \mu_5}{3}.$$

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Error Rates

- Comparison-wise Error Rate: the usual, one-at-a-time rate.
- Familywise Error Rate: the probability of at least one Type I error in a collection (or family) of tests.
  - (Weak) FWER: assumes all null hypotheses are true.
  - Strong FWER: does not assume all null hypotheses are true.
- False Discovery Rate: a rejection of H_0 is a *discovery*. FDR is the expected false discovery fraction, which is the proportion of discoveries that are mistakes.
- Simultaneous Coverage Probability: the probability that all intervals cover their respective parameters.

Which Error Rate?

- *Which error rate should we control?*
- *How do we define the Family?*

No easy answers. But we should take into account:

$$\underset{\text{more powerful}}{\overset{\text{less stringent,}}{\left(\right.}} \qquad \qquad \underset{\text{less powerful}}{\overset{\text{more stringent,}}{\left(\right.}}$$

$$\text{CWER} \leq \text{FWER} \leq \text{FDR} \leq \text{SFWER} \leq \text{SCP}$$

Which Error Rate?

- *Which error rate should we control?*
- *How do we define the Family?*

No easy answers. But we should take into account:

$$\underset{\text{more powerful}}{\overset{\text{less stringent,}}{\left(\right.}} \qquad \qquad \underset{\text{less powerful}}{\overset{\text{more stringent,}}{\left(\right.}}$$

$$\text{CWER} \leq \text{FWER} \leq \text{FDR} \leq \text{SFWER} \leq \text{SCP}$$

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Which Error Rate?

Answers also depend on

- Tradition and convention!
- Personal risk tolerance.
- Exploratory or confirmatory?
- Observational or experimental?
- Size of the family?
- Degree of dependence/redundancy among the inferences.
- Consequences of Type I vs Type II error.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, especially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, especially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, expecially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, expecially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, expecially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

Tradition says, if main effects are statistically significant, do *mean separation* adjusting for multiplicity.
Lots of methods. But for some types of contrasts and error rates, there are recommended approaches.

- To control FWER, use Fisher's "protected" LSD method.
  - Simple method: Test all planned comparisons **without** multiplicity adjustment, but **only if** main effect test is significant.
- To control the SFWER when making all pairwise comparisons, use Tukey's Honest Significant Difference (HSD) method.
  - Based on distribution of the studentized range of a set of sample means from the same population.
  - Looking at all pairwise comparisons is often a "fishing expedition" approach that's best avoided, expecially if the number of means is large.
  - In that case, the number of mean pairs is very large, making it very difficult to detect significant differences under any valid MCP.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

Multiple Comparison Procedures in ANOVA Models

- Often better to make all pairwise comparisons with a reference treatment (the control, or best, or worst treatment). In this case, Dunnett's method, which controls the SFWER, is recommended.
  - If 30 treatments, there are 29 pairwise comparisons with the best treatment, but $\binom{30}{2} = 435$ pairwise comparisons overall.
  - Good choice in "pick the winner" contexts.
  - Should use a one-sided alternative if reference is best or worst treatment.
- Letting data suggest the comparison to test is data-snooping (bad!).
  - Poses a severe multiplicity problem even if you do just one test because, informally, you did many tests.
  - Best: Don't do it. But if you do do it, use Scheffé's method to control SFWER.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Bonferroni Method
  - Simple, widely applicable approach to multiplicity in almost any context.
  - If we have a family of K inferences, divide the overall α Type I error rate evenly between them.
    - E.g., conduct each of K tests in your family at level α/K.
    - Or construct $100(1-\alpha/K)\%$ confidence intervals for each of K parameters in your family.
  - Simple, and applicable to contexts where another, more powerful approach may not be available.
  - Can be overly conservative, sacrificing power, especially for large K.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

General Methods: Bonferroni, Holm, & Benjamini-Hochberg

- Holm & Benjamini-Hochberg (aka FDR) methods.
  - Closely related to Bonferroni.
  - In each method we put the p-values of our K tests in ascending order. Then, significance threshold differs as we go sequentially through the list.
  - Holm: go up the list and compare the jth smallest p-value to $\frac{\alpha}{K-j+1}$ stopping at the first non-signicant test.
  - B-H: go down the list and compare the jth largest p-value to $\frac{j\alpha}{K}$. If significant, the jth test and all those with smaller p-values are significant.
- For tests, Holm controls SFWER and should *always* be used in place of Bonferroni.
- B-H method controls the FDR *if the tests are independent* and under many dependence scenarios.

Demo

```
options(width=100)
set.seed(9293); pVec <- sort(c(runif(5,0,.1),runif(5,0,.01))); alpha <- 0.05
# Adjusted alpha values (compare p to adjAlpha)
rbind(pVals=pVec,
      alpha=rep(alpha,10),
      Bon.alpha=alpha*pVec/p.adjust(pVec,method="bonferroni"),
      Holm.alpha=alpha*pVec/p.adjust(pVec,method="holm"),
      BH.alpha=alpha*pVec/p.adjust(pVec,method="BH")) %>% round(4)
```

```
             [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
pVals      0.0003 0.0017 0.0040 0.0051 0.0088 0.0306 0.0407 0.0661 0.0739 0.0864
alpha      0.0500 0.0500 0.0500 0.0500 0.0500 0.0500 0.0500 0.0500 0.0500 0.0500
Bon.alpha  0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050
Holm.alpha 0.0050 0.0056 0.0063 0.0071 0.0083 0.0100 0.0125 0.0167 0.0186 0.0218
BH.alpha   0.0050 0.0100 0.0157 0.0200 0.0250 0.0300 0.0350 0.0403 0.0450 0.0500
```

```
# Adjusted p-values (compare adjP to 0.05):
rbind(pVals=pVec,
      Bon.adjP =p.adjust(pVec,method="bonferroni"),
      Holm.adjP=p.adjust(pVec,method="holm"),
      BH.adjP  =p.adjust(pVec,method="BH")) %>% round(4)
```

```
            [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
pVals     0.0003 0.0017 0.0040 0.0051 0.0088 0.0306 0.0407 0.0661 0.0739 0.0864
Bon.adjP  0.0030 0.0169 0.0403 0.0514 0.0882 0.3058 0.4069 0.6610 0.7390 0.8641
Holm.adjP 0.0030 0.0152 0.0322 0.0360 0.0529 0.1529 0.1628 0.1983 0.1983 0.1983
BH.adjP   0.0030 0.0085 0.0129 0.0129 0.0176 0.0510 0.0581 0.0821 0.0821 0.0864
```

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

More on the FDR

The B-H method is chosen to control the FDR often in exploratory settings, especially when the number of tests to be conducted is very large.

- E.g., it is used in genomics problems, neuro-imaging studies, and other settings where test statistics and p-values are generated at 1000's of genes, 10,000's of voxels, etc.
- In such settings,
  - discoveries are almost certain;
  - methods to control the SFWER have very low power;
  - some false discoveries are tolerable in order to find real effects, as long as they are a small fraction of the total number of discoveries; and
  - replication is relied upon to separate the false positives from the true ones.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Regression and ANCOVA

Regression has multiple, distinct purposes.

- Prediction
- Estimation and inference on effects and associations.
  - Assessing effects of a single predictor.
  - Assessing effects of multiple predictors.

Variable selection presents a major multiplicity problem. Automatic variable selection methods are typically inadequately adjusted for multiplicity, and typically yield models with

- many spurious predictors (Type I errors),
- inflated regression coefficients among the predictors that are subject to selection.

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

Some say no.

- Rothman (1990) and others argue that multiplicity adjustment is misguided because we pay with Type II errors, and we will overlook interesting and potentially important findings.
  - This argument would be stronger if non-reproducibility would sort things out. But replication studies are hard to publish and discouraged relative to original research, so many exploratory results are more likely to get cited than re-tested.
- Another argument against adjustment is the arbitrariness of the exercise.
  - Adjustment in ANOVA, but not in equivalent regression model.
  - In a 4-way ANOVA, there are 15 tests of main effects and interactions. Do we adjust for multiplicity in this family? No one does, but they do when comparing means across levels of each factor.
  - If analyses of same intervention on several outcomes are published in different papers, or even by different authors, do the tests of the intervention have to be adjusted across all papers?

Should We Control for Multiplicity at All?

- Others object to the fact that multiplicity adjustment penalizes effects in large families more than in small families.
- A further objection goes like this: suppose we have two outcomes, and treatment effect is significant with $p = .047$ and $p = .032$. After Bonferroni adjustment, neither are significant.
  - Yes, but suppose they have p-values 0.047 and 0.86. Now do we object to the adjustment?
  - And if we have consistent findings, it is time to start estimating treatment effects rather than focusing on significance tests.

Should We Control for Multiplicity at All?

- Others object to the fact that multiplicity adjustment penalizes effects in large families more than in small families.
- A further objection goes like this: suppose we have two outcomes, and treatment effect is significant with $p = .047$ and $p = .032$. After Bonferroni adjustment, neither are significant.
  - Yes, but suppose they have p-values 0.047 and 0.86. Now do we object to the adjustment?
  - And if we have consistent findings, it is time to start estimating treatment effects rather than focusing on significance tests.

Should We Control for Multiplicity at All?

- Others object to the fact that multiplicity adjustment penalizes effects in large families more than in small families.
- A further objection goes like this: suppose we have two outcomes, and treatment effect is significant with $p = .047$ and $p = .032$. After Bonferroni adjustment, neither are significant.
  - Yes, but suppose they have p-values 0.047 and 0.86. Now do we object to the adjustment?
  - And if we have consistent findings, it is time to start estimating treatment effects rather than focusing on significance tests.

Should We Control for Multiplicity at All?

- Others object to the fact that multiplicity adjustment penalizes effects in large families more than in small families.
- A further objection goes like this: suppose we have two outcomes, and treatment effect is significant with $p = .047$ and $p = .032$. After Bonferroni adjustment, neither are significant.
  - Yes, but suppose they have p-values 0.047 and 0.86. Now do we object to the adjustment?
  - And if we have consistent findings, it is time to start estimating treatment effects rather than focusing on significance tests.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice

- Avoid the problem whenever possible.
  - Do not measure and analyze everything. Choose a limited set of outcomes carefully.
  - Select a primary outcome whenever possible and distinguish primary from secondary from exploratory.
  - Replace multiple outcomes by indices and/or use other methods to reduce the dimension of the set of outcomes.
  - In randomized experiments, do not test for covariate differences across groups in order to select covariates to include in the analyses.
  - "Statistical significance" is always arbitrary. Judge effects in light of theory, literature, whether results on multiple outcomes are consistent, etc.
- Plan ahead.
  - Plan the analyses, define families for which protection is sensible, and choose methods of adjustment.
  - Check the norms and requirements in your field, and in targeted journals.
  - Identify families based on distinct research questions.
  - Consider preregistration.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).
- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).
- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).
- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).
- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).
- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).

- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).

- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

Advice for Regression

- In predictive regression, focus isn't on significance of predictors; multiplicity not a major issue. Cross-validation can help justify the model.
- When focusing on effect of one or very few predictors, Type II errors more consequential. Accepting the null is the bigger problem.
  - Unless set of control variables and higher-order effects is very large, base inference on a maximal model without variable selection, it is unnecessary.
  - Don't re-do the analysis to re-test effect of predictor of interest under many different covariate sets.
  - Consider propensity score matching or weighting to adjust for covariates instead of or in addition to including covariates in the model.
  - Best not to even report estimates and tests on control variables, as they can be biased and/or misleading (Weistreich & Greenland, 2013; Hünermund & Louw, 2022).

- If want to find and quantify effects of important predictors, FDR or no adjustment makes more sense.
  - LASSO and related modern variable selection methods can diminish the multiplicity problem in this setting.

More Advice

- Be clear about what was done and report adjusted *and* un-adjusted results. Reader can decide. Hang your hat on the adjusted ones, though.
- Subgroup analysis is fishing. Adjust for multiplicity and report significant as exploratory and tentative.
- Multiplicity adjustment is not just for tests. CIs needs adjustment too.
- Be honest with yourself and your audience about what is confirmatory and, especially, what is exploratory. But there's a big grey zone.
- But... the real world is very different from the methodologically ideal one!

More Advice

- Be clear about what was done and report adjusted *and* un-adjusted results. Reader can decide. Hang your hat on the adjusted ones, though.
- Subgroup analysis is fishing. Adjust for multiplicity and report significant as exploratory and tentative.
- Multiplicity adjustment is not just for tests. CIs needs adjustment too.
- Be honest with yourself and your audience about what is confirmatory and, especially, what is exploratory. But there's a big grey zone.
- But... the real world is very different from the methodologically ideal one!

More Advice

- Be clear about what was done and report adjusted *and* un-adjusted results. Reader can decide. Hang your hat on the adjusted ones, though.
- Subgroup analysis is fishing. Adjust for multiplicity and report significant as exploratory and tentative.
- Multiplicity adjustment is not just for tests. CIs needs adjustment too.
- Be honest with yourself and your audience about what is confirmatory and, especially, what is exploratory. But there's a big grey zone.
- But... the real world is very different from the methodologically ideal one!

More Advice

- Be clear about what was done and report adjusted *and* un-adjusted results. Reader can decide. Hang your hat on the adjusted ones, though.
- Subgroup analysis is fishing. Adjust for multiplicity and report significant as exploratory and tentative.
- Multiplicity adjustment is not just for tests. CIs needs adjustment too.
- Be honest with yourself and your audience about what is confirmatory and, especially, what is exploratory. But there's a big grey zone.
- But. . . the real world is very different from the methodologically ideal one!

More Advice

- Be clear about what was done and report adjusted *and* un-adjusted results. Reader can decide. Hang your hat on the adjusted ones, though.
- Subgroup analysis is fishing. Adjust for multiplicity and report significant as exploratory and tentative.
- Multiplicity adjustment is not just for tests. CIs needs adjustment too.
- Be honest with yourself and your audience about what is confirmatory and, especially, what is exploratory. But there's a big grey zone.
- But... the real world is very different from the methodologically ideal one!

References & Resources

The Dead Salmon Study:

- Bennett et al. (2010). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction, &Journal of Serendipitous and Unexpected Results,* 1(1):1-5. Download article here, download poster here.

Nice blog post on multiple comparisons in neuroimaging:

- http://jpeelle.net/mri/statistics/multiple_comparisons.html

Guidelines on multiple comparisons procedures from the *British Journal of Dermatology*:

- Hollestein, L., Lo, S., Leonardi-Bee, J., Rosset, S., Shomron, N., Couturier, D.-L. and Gran, S. (2021), MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *Br J Dermatol*, 185: 1081-1083.

References & Resources

The Dead Salmon Study:

- Bennett et al. (2010). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction, &Journal of Serendipitous and Unexpected Results,* 1(1):1-5. Download article here, download poster here.

Nice blog post on multiple comparisons in neuroimaging:

- http://jpeelle.net/mri/statistics/multiple_comparisons.html

Guidelines on multiple comparisons procedures from the *British Journal of Dermatology*:

- Hollestein, L., Lo, S., Leonardi-Bee, J., Rosset, S., Shomron, N., Couturier, D.-L. and Gran, S. (2021), MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *Br J Dermatol*, 185: 1081-1083.

References & Resources

The Dead Salmon Study:

- Bennett et al. (2010). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction, &Journal of Serendipitous and Unexpected Results,* 1(1):1-5. Download article here, download poster here.

Nice blog post on multiple comparisons in neuroimaging:

- http://jpeelle.net/mri/statistics/multiple_comparisons.html

Guidelines on multiple comparisons procedures from the *British Journal of Dermatology*:

- Hollestein, L., Lo, S., Leonardi-Bee, J., Rosset, S., Shomron, N., Couturier, D.-L. and Gran, S. (2021), MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *Br J Dermatol*, 185: 1081-1083.

References & Resources

Guidelines on multiple comparisons procedures from the Office of Evaluation Sciences, General Services Administration of the US federal government:

- https://oes.gsa.gov/assets/files/multiple-comparison-adjustment.pdf

Guidelines on Multiplicity Issues in Clinical Trials from the European Medicines Agency:

- https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

Guidelines for Multiple Testing in Impact Evaluations from the National Center for Education Evaluation and Regional Assistance:

- https://ies.ed.gov/ncee/pubs/20084018/chapter_2.asp

An excellent introduction to multiple comparisons from EGAP:

- Coppock, A. 10 Things to Know About Multiple Comparisons. https://egap.org/resource/10-things-to-know-about-multiple-comparisons/

References & Resources

Guidelines on multiple comparisons procedures from the Office of Evaluation Sciences, General Services Administration of the US federal government:

- https://oes.gsa.gov/assets/files/multiple-comparison-adjustment.pdf

Guidelines on Multiplicity Issues in Clinical Trials from the European Medicines Agency:

- https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

Guidelines for Multiple Testing in Impact Evaluations from the National Center for Education Evaluation and Regional Assistance:

- https://ies.ed.gov/ncee/pubs/20084018/chapter_2.asp

An excellent introduction to multiple comparisons from EGAP:

- Coppock, A. 10 Things to Know About Multiple Comparisons. https://egap.org/resource/10-things-to-know-about-multiple-comparisons/

References & Resources

Guidelines on multiple comparisons procedures from the Office of Evaluation Sciences, General Services Administration of the US federal government:

- https://oes.gsa.gov/assets/files/multiple-comparison-adjustment.pdf

Guidelines on Multiplicity Issues in Clinical Trials from the European Medicines Agency:

- https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

Guidelines for Multiple Testing in Impact Evaluations from the National Center for Education Evaluation and Regional Assistance:

- https://ies.ed.gov/ncee/pubs/20084018/chapter_2.asp

An excellent introduction to multiple comparisons from EGAP:

- Coppock, A. 10 Things to Know About Multiple Comparisons. https://egap.org/resource/10-things-to-know-about-multiple-comparisons/

References & Resources

Guidelines on multiple comparisons procedures from the Office of Evaluation Sciences, General Services Administration of the US federal government:

- https://oes.gsa.gov/assets/files/multiple-comparison-adjustment.pdf

Guidelines on Multiplicity Issues in Clinical Trials from the European Medicines Agency:

- https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

Guidelines for Multiple Testing in Impact Evaluations from the National Center for Education Evaluation and Regional Assistance:

- https://ies.ed.gov/ncee/pubs/20084018/chapter_2.asp

An excellent introduction to multiple comparisons from EGAP:

- Coppock, A. 10 Things to Know About Multiple Comparisons. https://egap.org/resource/10-things-to-know-about-multiple-comparisons/

References & Resources

The argument for not controlling for multiplicity:

- Rothman, Kenneth J. (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology,* 1(1): 43-46.
- Althouse, A.D. (2016). Adjust for Multiple Comparisons? It's Not That Simple. *The Annals of Thoracic Surgery,* 101(5): 1644–1645.

G. van Belle's Statistical Rule of Thumb on Multiple Comparisons:

- http://www.vanbelle.org/rom/ROM_2002_06.pdf

References & Resources

The argument for not controlling for multiplicity:

- Rothman, Kenneth J. (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology,* 1(1): 43-46.
- Althouse, A.D. (2016). Adjust for Multiple Comparisons? It's Not That Simple. *The Annals of Thoracic Surgery,* 101(5): 1644–1645.

G. van Belle's Statistical Rule of Thumb on Multiple Comparisons:

- http://www.vanbelle.org/rom/ROM_2002_06.pdf

References & Resources

The argument for not controlling for multiplicity:

- Rothman, Kenneth J. (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology,* 1(1): 43-46.
- Althouse, A.D. (2016). Adjust for Multiple Comparisons? It's Not That Simple. *The Annals of Thoracic Surgery,* 101(5): 1644–1645.

G. van Belle's Statistical Rule of Thumb on Multiple Comparisons:

- http://www.vanbelle.org/rom/ROM_2002_06.pdf

References & Resources

A nice article on multiple comparisons in regression:

- Anderson, S. F. (2022, May 19). Multiplicity in Multiple Regression: Defining the Issue, Evaluating Solutions, and Integrating Perspectives. *Psychological Methods.*

An overview of multiple comparisons with advice:

- Streiner, DL (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests, *Am J Clin Nutr,* 102:721–8

References on not reporting/interpretting control variables:

- Hünermund, P. and Louw, B. (2022). On the nuisance of control variables in regression.https://arxiv.org/pdf/2005.10314.pdf
- Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology,* 177(4):292–298.

References & Resources

A nice article on multiple comparisons in regression:

- Anderson, S. F. (2022, May 19). Multiplicity in Multiple Regression: Defining the Issue, Evaluating Solutions, and Integrating Perspectives. *Psychological Methods.*

An overview of multiple comparisons with advice:

- Streiner, DL (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests, *Am J Clin Nutr,* 102:721–8

References on not reporting/interpretting control variables:

- Hünermund, P. and Louw, B. (2022). On the nuisance of control variables in regression.https://arxiv.org/pdf/2005.10314.pdf
- Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology,* 177(4):292–298.

References & Resources

A nice article on multiple comparisons in regression:

- Anderson, S. F. (2022, May 19). Multiplicity in Multiple Regression: Defining the Issue, Evaluating Solutions, and Integrating Perspectives. *Psychological Methods.*

An overview of multiple comparisons with advice:

- Streiner, DL (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests, *Am J Clin Nutr,* 102:721–8

References on not reporting/interpretting control variables:

- Hünermund, P. and Louw, B. (2022). On the nuisance of control variables in regression.https://arxiv.org/pdf/2005.10314.pdf
- Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology,* 177(4):292–298.

References & Resources

A nice article on multiple comparisons in regression:

- Anderson, S. F. (2022, May 19). Multiplicity in Multiple Regression: Defining the Issue, Evaluating Solutions, and Integrating Perspectives. *Psychological Methods.*

An overview of multiple comparisons with advice:

- Streiner, DL (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests, *Am J Clin Nutr,* 102:721–8

References on not reporting/interpretting control variables:

- Hünermund, P. and Louw, B. (2022). On the nuisance of control variables in regression.https://arxiv.org/pdf/2005.10314.pdf
- Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology,* 177(4):292–298.

Thanks

- If you need assistance with multiple comparisons or with any statistical design or analysis task, please contact the SCC.
  - www.stat.uga/consulting
- We can help!

Thank you!

Thanks

- If you need assistance with multiple comparisons or with any statistical design or analysis task, please contact the SCC.
  - www.stat.uga/consulting
- We can help!

Thank you!

Thanks

- If you need assistance with multiple comparisons or with any statistical design or analysis task, please contact the SCC.
  - www.stat.uga/consulting
- We can help!

Thank you!

Questions?

Questions?