

Multiple Comparisons

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

P310, 2011

Multiple Comparisons

1 Introduction

2 Two Major Problems

- Multiple Testing and the Proliferation of Type I Error
- The Problem of Post-Hoc Inference

3 The “Big Three” Multiple Comparison Procedures

- Planned Contrasts
- The Sheffe Test
- The Tukey Test

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - How many groups are significantly different?
- Some key questions are:
 - How can we do these tests?
 - What are the answers?

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:
 - How can we do these tests?
 - What are the consequences?

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:
 - How can we do these tests?
 - What can go wrong?

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:
 - How can we do these tests?
 - What can go wrong?

Multiple Testing and Multiple Comparisons

A Standard Situation

- Suppose you perform a 1-Way Analysis of Variance (ANOVA) on 5 groups. Group 1 is a control group, and Groups 2–5 are experimental groups.
- The standard F -test is significant. At that point, you have several questions. For example,
 - Which pairs of groups are significantly different?
 - Which experimental groups are significantly different from the controls?
- Some key questions are:
 - How can we do these tests?
 - What can go wrong?

Multiple Testing and Multiple Comparisons

A Standard Situation

- Now, your first reaction might be to simply do a bunch of two-sample t -tests.
- To speed things up, you might look at two means that are really far apart, and do a t -test to see if they are significantly different.
- The problem is — that would be wrong, in several ways, for several reasons.
- Let’s look at the ways performing these t -tests might be wrong.

Multiple Testing and Multiple Comparisons

A Standard Situation

- Now, your first reaction might be to simply do a bunch of two-sample t -tests.
- To speed things up, you might look at two means that are really far apart, and do a t -test to see if they are significantly different.
- The problem is — that would be wrong, in several ways, for several reasons.
- Let’s look at the ways performing these t -tests might be wrong.

Multiple Testing and Multiple Comparisons

A Standard Situation

- Now, your first reaction might be to simply do a bunch of two-sample t -tests.
- To speed things up, you might look at two means that are really far apart, and do a t -test to see if they are significantly different.
- The problem is — that would be wrong, in several ways, for several reasons.
- Let's look at the ways performing these t -tests might be wrong.

Multiple Testing and Multiple Comparisons

A Standard Situation

- Now, your first reaction might be to simply do a bunch of two-sample t -tests.
- To speed things up, you might look at two means that are really far apart, and do a t -test to see if they are significantly different.
- The problem is — that would be wrong, in several ways, for several reasons.
- Let's look at the ways performing these t -tests might be wrong.

Two Major Problems

- You recall that I suggested two approaches to questions of which pairs of means are significantly different.
- These approaches are wrong, because they ignore two fundamental problems that occur in the situation we described, but also occur very widely throughout applications of statistics in science.
- These problems are:

Two Major Problems

- You recall that I suggested two approaches to questions of which pairs of means are significantly different.
- These approaches are wrong, because they ignore two fundamental problems that occur in the situation we described, but also occur very widely throughout applications of statistics in science.
- These problems are:

- *Multiple testing and the proliferation of Type I errors.* When you do a lot of tests, the probabilities “catch up to you.”

Two Major Problems

- You recall that I suggested two approaches to questions of which pairs of means are significantly different.
- These approaches are wrong, because they ignore two fundamental problems that occur in the situation we described, but also occur very widely throughout applications of statistics in science.
- These problems are:
 - *Multiple testing and the proliferation of Type I errors.* When you do a lot of tests, the probabilities “catch up to you.”
 - *Post-hoc inference.* Once you have looked at your data and sorted through the data to find impressive-looking things to analyze, the probability model has changed, and you are vulnerable to statistical self-delusion.

Two Major Problems

- You recall that I suggested two approaches to questions of which pairs of means are significantly different.
- These approaches are wrong, because they ignore two fundamental problems that occur in the situation we described, but also occur very widely throughout applications of statistics in science.
- These problems are:
 - *Multiple testing and the proliferation of Type I errors.* When you do a lot of tests, the probabilities “catch up to you.”
 - *Post-hoc inference.* Once you have looked at your data and sorted through the data to find impressive-looking things to analyze, the probability model has changed, and you are vulnerable to statistical self-delusion.

Two Major Problems

- You recall that I suggested two approaches to questions of which pairs of means are significantly different.
- These approaches are wrong, because they ignore two fundamental problems that occur in the situation we described, but also occur very widely throughout applications of statistics in science.
- These problems are:
 - *Multiple testing and the proliferation of Type I errors.* When you do a lot of tests, the probabilities “catch up to you.”
 - *Post-hoc inference.* Once you have looked at your data and sorted through the data to find impressive-looking things to analyze, the probability model has changed, and you are vulnerable to statistical self-delusion.

Multiple Testing and the Proliferation of Type I Error

- In our example, we have five groups. If we did all possible two-sample independent sample t -tests, we would be doing 10 t – tests.
- Suppose that groups 1 and 2 are actually different, but all the other groups have means equal to Group 2. So the null hypothesis of equality is false for 5 of the comparisons, and true for 5. What is the probability that *at least one of the t -tests* will be falsely significant?
- We can get a handle on the problem by imagining that the 5 significance tests are all done at the .05 level, i.e., with $\alpha = .05$, and also that all the 5 tests are independent.
- In this case, the probability of all 5 tests *not* rejecting is $(1 - \alpha)^5 = .95^5$. So the probability of the complementary event, that at least one of the tests rejects incorrectly, is $1 - .95^5 = 0.226$. This is known as the *familywise error rate* for the family of tests. (It is also referred to as *Error Rate Experimentwise* in earlier literature.)
- In general, if you perform k significance tests in a situation where all null hypotheses are true and the tests are independent, the familywise error rate is $1 - (1 - \alpha)^k$.

Multiple Testing and the Proliferation of Type I Error

- In our example, we have five groups. If we did all possible two-sample independent sample t -tests, we would be doing 10 t – tests.
- Suppose that groups 1 and 2 are actually different, but all the other groups have means equal to Group 2. So the null hypothesis of equality is false for 5 of the comparisons, and true for 5. What is the probability that *at least one of the t -tests* will be falsely significant?
- We can get a handle on the problem by imagining that the 5 significance tests are all done at the .05 level, i.e., with $\alpha = .05$, and also that all the 5 tests are independent.
- In this case, the probability of all 5 tests *not* rejecting is $(1 - \alpha)^5 = .95^5$. So the probability of the complementary event, that at least one of the tests rejects incorrectly, is $1 - .95^5 = 0.226$. This is known as the *familywise error rate* for the family of tests. (It is also referred to as *Error Rate Experimentwise* in earlier literature.)
- In general, if you perform k significance tests in a situation where all null hypotheses are true and the tests are independent, the familywise error rate is $1 - (1 - \alpha)^k$.

Multiple Testing and the Proliferation of Type I Error

- In our example, we have five groups. If we did all possible two-sample independent sample t -tests, we would be doing 10 t – tests.
- Suppose that groups 1 and 2 are actually different, but all the other groups have means equal to Group 2. So the null hypothesis of equality is false for 5 of the comparisons, and true for 5. What is the probability that *at least one of the t -tests* will be falsely significant?
- We can get a handle on the problem by imagining that the 5 significance tests are all done at the .05 level, i.e., with $\alpha = .05$, and also that all the 5 tests are independent.
- In this case, the probability of all 5 tests *not* rejecting is $(1 - \alpha)^5 = .95^5$. So the probability of the complementary event, that at least one of the tests rejects incorrectly, is $1 - .95^5 = 0.226$. This is known as the *familywise error rate* for the family of tests. (It is also referred to as *Error Rate Experimentwise* in earlier literature.)
- In general, if you perform k significance tests in a situation where all null hypotheses are true and the tests are independent, the familywise error rate is $1 - (1 - \alpha)^k$.

Multiple Testing and the Proliferation of Type I Error

- In our example, we have five groups. If we did all possible two-sample independent sample t -tests, we would be doing 10 t - tests.
- Suppose that groups 1 and 2 are actually different, but all the other groups have means equal to Group 2. So the null hypothesis of equality is false for 5 of the comparisons, and true for 5. What is the probability that *at least one of the t -tests* will be falsely significant?
- We can get a handle on the problem by imagining that the 5 significance tests are all done at the .05 level, i.e., with $\alpha = .05$, and also that all the 5 tests are independent.
- In this case, the probability of all 5 tests *not* rejecting is $(1 - \alpha)^5 = .95^5$. So the probability of the complementary event, that at least one of the tests rejects incorrectly, is $1 - .95^5 = 0.226$. This is known as the *familywise error rate* for the family of tests. (It is also referred to as *Error Rate Experimentwise* in earlier literature.)
- In general, if you perform k significance tests in a situation where all null hypotheses are true and the tests are independent, the familywise error rate is $1 - (1 - \alpha)^k$.

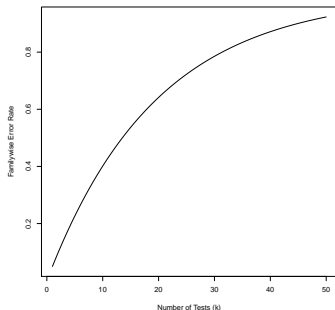
Multiple Testing and the Proliferation of Type I Error

- In our example, we have five groups. If we did all possible two-sample independent sample t -tests, we would be doing 10 t – tests.
- Suppose that groups 1 and 2 are actually different, but all the other groups have means equal to Group 2. So the null hypothesis of equality is false for 5 of the comparisons, and true for 5. What is the probability that *at least one of the t -tests* will be falsely significant?
- We can get a handle on the problem by imagining that the 5 significance tests are all done at the .05 level, i.e., with $\alpha = .05$, and also that all the 5 tests are independent.
- In this case, the probability of all 5 tests *not* rejecting is $(1 - \alpha)^5 = .95^5$. So the probability of the complementary event, that at least one of the tests rejects incorrectly, is $1 - .95^5 = 0.226$. This is known as the *familywise error rate* for the family of tests. (It is also referred to as *Error Rate Experimentwise* in earlier literature.)
- In general, if you perform k significance tests in a situation where all null hypotheses are true and the tests are independent, the familywise error rate is $1 - (1 - \alpha)^k$.

Multiple Testing and the Proliferation of Type I Error

- Below is a plot of familywise error rate as a function of k , the number of tests.
- You can see that, if you do a lot of tests, the probability of at least one false positive becomes non-negligible.

```
> curve(1-.95^x,1,50,  
+ xlab="Number of Tests (k)",ylab="Familywise Error Rate")
```



Multiple Testing and the Proliferation of Type I Error

What to do?

- If all the significance tests you do are independent, then it is easy to compute the familywise error rate.
- You can solve inversely for an α significance level that you can use for each individual test, so that the familywise error rate is controlled at some desirable level FWE .
- Specifically,

$$\alpha = 1 - (1 - FWE)^{(1/k)} \quad (1)$$

- But what if the tests are *not* independent?

Multiple Testing and the Proliferation of Type I Error

What to do?

- If all the significance tests you do are independent, then it is easy to compute the familywise error rate.
- You can solve inversely for an α significance level that you can use for each individual test, so that the familywise error rate is controlled at some desirable level FWE .

- Specifically,

$$\alpha = 1 - (1 - FWE)^{(1/k)} \quad (1)$$

- But what if the tests are *not* independent?

Multiple Testing and the Proliferation of Type I Error

What to do?

- If all the significance tests you do are independent, then it is easy to compute the familywise error rate.
- You can solve inversely for an α significance level that you can use for each individual test, so that the familywise error rate is controlled at some desirable level FWE .

- Specifically,

$$\alpha = 1 - (1 - FWE)^{(1/k)} \quad (1)$$

- But what if the tests are *not* independent?

Multiple Testing and the Proliferation of Type I Error

What to do?

- If all the significance tests you do are independent, then it is easy to compute the familywise error rate.
- You can solve inversely for an α significance level that you can use for each individual test, so that the familywise error rate is controlled at some desirable level FWE .

- Specifically,

$$\alpha = 1 - (1 - FWE)^{(1/k)} \quad (1)$$

- But what if the tests are *not* independent?

Multiple Testing and the Proliferation of Type I Error

Doing a “Bonferroni Split”

- One of the inequalities due to Bonferroni is that, for a set of events E_i ,

$$\Pr\left(\bigcup_{i=1}^k E_i\right) \leq \sum_{i=1}^k \Pr(E_i) \quad (2)$$

- So if all k tests are done at the α level, then the FWE must be less than or equal to $k\alpha$.
- This in turn implies that a failsafe way of controlling FWE *at or below* α is to do each significance test at the α/k significance level.
- When I was a young graduate student, we sometimes referred to this very general approach to controlling familywise error rate as “doing a Bonferroni split.”

Multiple Testing and the Proliferation of Type I Error

Doing a “Bonferroni Split”

- One of the inequalities due to Bonferroni is that, for a set of events E_i ,

$$\Pr\left(\bigcup_{i=1}^k E_i\right) \leq \sum_{i=1}^k \Pr(E_i) \quad (2)$$

- So if all k tests are done at the α level, then the FWE must be less than or equal to $k\alpha$.
- This in turn implies that a failsafe way of controlling FWE *at or below* α is to do each significance test at the α/k significance level.
- When I was a young graduate student, we sometimes referred to this very general approach to controlling familywise error rate as “doing a Bonferroni split.”

Multiple Testing and the Proliferation of Type I Error

Doing a “Bonferroni Split”

- One of the inequalities due to Bonferroni is that, for a set of events E_i ,

$$\Pr\left(\bigcup_{i=1}^k E_i\right) \leq \sum_{i=1}^k \Pr(E_i) \quad (2)$$

- So if all k tests are done at the α level, then the FWE must be less than or equal to $k\alpha$.
- This in turn implies that a failsafe way of controlling FWE *at or below* α is to do each significance test at the α/k significance level.
- When I was a young graduate student, we sometimes referred to this very general approach to controlling familywise error rate as “doing a Bonferroni split.”

Multiple Testing and the Proliferation of Type I Error

Doing a “Bonferroni Split”

- One of the inequalities due to Bonferroni is that, for a set of events E_i ,

$$\Pr\left(\bigcup_{i=1}^k E_i\right) \leq \sum_{i=1}^k \Pr(E_i) \quad (2)$$

- So if all k tests are done at the α level, then the FWE must be less than or equal to $k\alpha$.
- This in turn implies that a failsafe way of controlling FWE *at or below* α is to do each significance test at the α/k significance level.
- When I was a young graduate student, we sometimes referred to this very general approach to controlling familywise error rate as “doing a Bonferroni split.”

Improving on Bonferroni

- The Bonferroni method is straightforward, but it comes at a cost.
- For example, if you have 10 tests, and you want to control FWE at .05, you have to do the individual tests at the .005 level.
- Of course, this costs you power and raises the possibility of Type II errors.
- Statisticians learned this fairly early in the game, and began looking for ways to improve power while still controlling FWE.

Improving on Bonferroni

- The Bonferroni method is straightforward, but it comes at a cost.
- For example, if you have 10 tests, and you want to control FWE at .05, you have to do the individual tests at the .005 level.
- Of course, this costs you power and raises the possibility of Type II errors.
- Statisticians learned this fairly early in the game, and began looking for ways to improve power while still controlling FWE.

Improving on Bonferroni

- The Bonferroni method is straightforward, but it comes at a cost.
- For example, if you have 10 tests, and you want to control FWE at .05, you have to do the individual tests at the .005 level.
- Of course, this costs you power and raises the possibility of Type II errors.
- Statisticians learned this fairly early in the game, and began looking for ways to improve power while still controlling FWE.

Improving on Bonferroni

- The Bonferroni method is straightforward, but it comes at a cost.
- For example, if you have 10 tests, and you want to control FWE at .05, you have to do the individual tests at the .005 level.
- Of course, this costs you power and raises the possibility of Type II errors.
- Statisticians learned this fairly early in the game, and began looking for ways to improve power while still controlling FWE.

Improving on Bonferroni

The Method of Closure

- The method of closure assumes that the tests are mathematically disconnected in the sense that the falsity of one hypothesis is not precluded by the status of others.
- In general, this method works well even when the tests are not perfectly disconnected, and controls FWE at or below α .
- In this method, you order the significance levels (p -values) of your k tests from lowest to highest. Then you look at the lowest p -value. If it is less than α/k , you reject the null hypothesis for that test and move on to the next test. If not, you stop.
- For each successive test, you reduce the divisor by one. For example, the second test is performed by finding the second smallest p -value, and seeing if it is less than $\frac{\alpha}{k-1}$. You keep performing tests until the first one fails to reject, and then you must stop.
- The method of closure has more power than the Bonferroni method, and has been applied to testing large numbers of correlations.

Improving on Bonferroni

The Method of Closure

- The method of closure assumes that the tests are mathematically disconnected in the sense that the falsity of one hypothesis is not precluded by the status of others.
- In general, this method works well even when the tests are not perfectly disconnected, and controls FWE at or below α .
- In this method, you order the significance levels (p -values) of your k tests from lowest to highest. Then you look at the lowest p -value. If it is less than α/k , you reject the null hypothesis for that test and move on to the next test. If not, you stop.
- For each successive test, you reduce the divisor by one. For example, the second test is performed by finding the second smallest p -value, and seeing if it is less than $\frac{\alpha}{k-1}$. You keep performing tests until the first one fails to reject, and then you must stop.
- The method of closure has more power than the Bonferroni method, and has been applied to testing large numbers of correlations.

Improving on Bonferroni

The Method of Closure

- The method of closure assumes that the tests are mathematically disconnected in the sense that the falsity of one hypothesis is not precluded by the status of others.
- In general, this method works well even when the tests are not perfectly disconnected, and controls FWE at or below α .
- In this method, you order the significance levels (p -values) of your k tests from lowest to highest. Then you look at the lowest p -value. If it is less than α/k , you reject the null hypothesis for that test and move on to the next test. If not, you stop.
- For each successive test, you reduce the divisor by one. For example, the second test is performed by finding the second smallest p -value, and seeing if it is less than $\frac{\alpha}{k-1}$. You keep performing tests until the first one fails to reject, and then you must stop.
- The method of closure has more power than the Bonferroni method, and has been applied to testing large numbers of correlations.

Improving on Bonferroni

The Method of Closure

- The method of closure assumes that the tests are mathematically disconnected in the sense that the falsity of one hypothesis is not precluded by the status of others.
- In general, this method works well even when the tests are not perfectly disconnected, and controls FWE at or below α .
- In this method, you order the significance levels (p -values) of your k tests from lowest to highest. Then you look at the lowest p -value. If it is less than α/k , you reject the null hypothesis for that test and move on to the next test. If not, you stop.
- For each successive test, you reduce the divisor by one. For example, the second test is performed by finding the second smallest p -value, and seeing if it is less than $\frac{\alpha}{k-1}$. You keep performing tests until the first one fails to reject, and then you must stop.
- The method of closure has more power than the Bonferroni method, and has been applied to testing large numbers of correlations.

Improving on Bonferroni

The Method of Closure

- The method of closure assumes that the tests are mathematically disconnected in the sense that the falsity of one hypothesis is not precluded by the status of others.
- In general, this method works well even when the tests are not perfectly disconnected, and controls FWE at or below α .
- In this method, you order the significance levels (p -values) of your k tests from lowest to highest. Then you look at the lowest p -value. If it is less than α/k , you reject the null hypothesis for that test and move on to the next test. If not, you stop.
- For each successive test, you reduce the divisor by one. For example, the second test is performed by finding the second smallest p -value, and seeing if it is less than $\frac{\alpha}{k-1}$. You keep performing tests until the first one fails to reject, and then you must stop.
- The method of closure has more power than the Bonferroni method, and has been applied to testing large numbers of correlations.

The Problem of Post-Hoc Inference

- We saw early in Psychology 310 that once we examine (and order) a set of observations, they become, in a sense, *order statistics*.
- Order statistics do not, in general, have the same distribution of the observations on which they are based.
- For example, suppose we take a sample of $n = 100$ observations from a $N(0, 1)$ distribution, then order them from lowest to highest.
- The n th lowest observation is known as the n th order statistic.
- The distribution of the n th order statistic is quite different from a $N(0, 1)$ distribution, because we have gone through the data to find it.

The Problem of Post-Hoc Inference

- We saw early in Psychology 310 that once we examine (and order) a set of observations, they become, in a sense, *order statistics*.
- Order statistics do not, in general, have the same distribution of the observations on which they are based.
- For example, suppose we take a sample of $n = 100$ observations from a $N(0, 1)$ distribution, then order them from lowest to highest.
- The n th lowest observation is known as the n th order statistic.
- The distribution of the n th order statistic is quite different from a $N(0, 1)$ distribution, because we have gone through the data to find it.

The Problem of Post-Hoc Inference

- We saw early in Psychology 310 that once we examine (and order) a set of observations, they become, in a sense, *order statistics*.
- Order statistics do not, in general, have the same distribution of the observations on which they are based.
- For example, suppose we take a sample of $n = 100$ observations from a $N(0, 1)$ distribution, then order them from lowest to highest.
- The n th lowest observation is known as the n th order statistic.
- The distribution of the n th order statistic is quite different from a $N(0, 1)$ distribution, because we have gone through the data to find it.

The Problem of Post-Hoc Inference

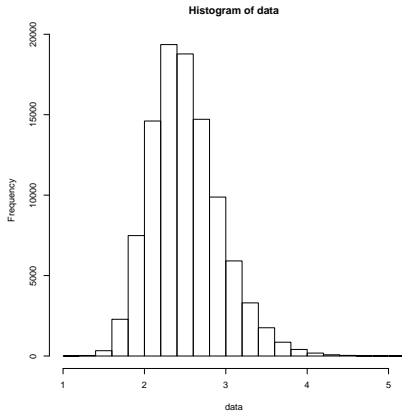
- We saw early in Psychology 310 that once we examine (and order) a set of observations, they become, in a sense, *order statistics*.
- Order statistics do not, in general, have the same distribution of the observations on which they are based.
- For example, suppose we take a sample of $n = 100$ observations from a $N(0, 1)$ distribution, then order them from lowest to highest.
- The n th lowest observation is known as the n th order statistic.
- The distribution of the n th order statistic is quite different from a $N(0, 1)$ distribution, because we have gone through the data to find it.

The Problem of Post-Hoc Inference

- We saw early in Psychology 310 that once we examine (and order) a set of observations, they become, in a sense, *order statistics*.
- Order statistics do not, in general, have the same distribution of the observations on which they are based.
- For example, suppose we take a sample of $n = 100$ observations from a $N(0, 1)$ distribution, then order them from lowest to highest.
- The n th lowest observation is known as the n th order statistic.
- The distribution of the n th order statistic is quite different from a $N(0, 1)$ distribution, because we have gone through the data to find it.

The Problem of Post-Hoc Inference

```
> data <- replicate(100000,max(rnorm(100,0,1)))  
> hist(data)
```



The Problem of Post-Hoc Inference

- So, of course, if we sample the data, page through it, and “cherry-pick” the most interesting and outstanding results, we can no longer apply the probability model that would be relevant had we not gone through the data.
- This point is fundamental, yet it is still violated by many researchers, aided and abetted by software such as SPSS.
- When researchers violate this principle, I like to say they are falling victim to “the Fallacy of the Order Statistic.”
- In order to decide whether a “cherry-picked” observation (often, a really low p -value) is “really” significant, we need to take into account the fact that it is an order statistic.

The Big Three Multiple Comparison Procedures in ANOVA

- In the context of ANOVA, we traditionally begin the discussion of multiple testing issues by discussing three procedures:
 - ① Planned Orthogonal Contrasts
 - ② The Sheffe post-hoc procedure
 - ③ The Tukey test

The Big Three Multiple Comparison Procedures in ANOVA

- In the context of ANOVA, we traditionally begin the discussion of multiple testing issues by discussing three procedures:
 - ① Planned Orthogonal Contrasts
 - ② The Sheffe post-hoc procedure
 - ③ The Tukey test

The Big Three Multiple Comparison Procedures in ANOVA

- In the context of ANOVA, we traditionally begin the discussion of multiple testing issues by discussing three procedures:
 - 1 Planned Orthogonal Contrasts
 - 2 The Sheffe post-hoc procedure
 - 3 The Tukey test

The Big Three Multiple Comparison Procedures in ANOVA

- In the context of ANOVA, we traditionally begin the discussion of multiple testing issues by discussing three procedures:
 - 1 Planned Orthogonal Contrasts
 - 2 The Sheffe post-hoc procedure
 - 3 The Tukey test

Planned Contrasts

- In some cases, you are less interested in the overall F test than you are in testing highly specific linear combination hypotheses.
- Moreover, in some cases these hypotheses “break down” the information in a set of means into non-overlapping components that are statistically uncorrelated.
- “Planned Orthogonal Contrasts” are employed in such a situation.

Planned Contrasts

- In some cases, you are less interested in the overall F test than you are in testing highly specific linear combination hypotheses.
- Moreover, in some cases these hypotheses “break down” the information in a set of means into non-overlapping components that are statistically uncorrelated.
- “Planned Orthogonal Contrasts” are employed in such a situation.

Planned Contrasts

- In some cases, you are less interested in the overall F test than you are in testing highly specific linear combination hypotheses.
- Moreover, in some cases these hypotheses “break down” the information in a set of means into non-overlapping components that are statistically uncorrelated.
- “Planned Orthogonal Contrasts” are employed in such a situation.

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- A linear combination hypothesis is a contrast if and only if the linear weights sum to zero and $\alpha = 0$, i.e., it is of the form

$$H_0 : \sum_{j=1}^J c_j \mu_j = 0, \text{ with } \sum_{j=1}^J c_j = 0 \quad (3)$$

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- A linear combination hypothesis is a *contrast* if and only if the linear weights sum to zero and $a = 0$, i.e., it is of the form

$$H_0 : \sum_j c_j \mu_j = 0, \quad \text{with} \quad \sum_j c_j = 0 \quad (3)$$

- Consider two contrasts A and B defined by linear weights $c_{j,a}$ and $c(j,b)$, for $j = 1, J$. These contrast are said to be orthogonal if and only if $\sum_{j=1}^J c_{j,a} c(j,b) = 0$, that is, the sum of cross-products of linear weights must be zero.
- For $\hat{\sigma}^2$, you use the data from *all the groups*, regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination.
- Generally, you do a Bonferroni split on the significance level.

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- 1 A linear combination hypothesis is a *contrast* if and only if the linear weights sum to zero and $a = 0$, i.e., it is of the form

$$H_0 : \sum_j c_j \mu_j = 0, \quad \text{with} \quad \sum_j c_j = 0 \quad (3)$$

- 2 Consider two contrasts A and B defined by linear weights $c_{j,a}$ and $c_{j,b}$, for $j = 1, J$. These contrast are said to be orthogonal if and only if $\sum_{j=1}^J c_{j,a} c_{j,b} = 0$, that is, the sum of cross-products of linear weights must be zero.
- 3 For $\hat{\sigma}^2$, you use the data from *all the groups*, regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination.
- 4 Generally, you do a Bonferroni split on the significance level.

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- A linear combination hypothesis is a *contrast* if and only if the linear weights sum to zero and $a = 0$, i.e., it is of the form

$$H_0 : \sum_j c_j \mu_j = 0, \quad \text{with} \quad \sum_j c_j = 0 \quad (3)$$

- Consider two contrasts A and B defined by linear weights $c_{j,a}$ and $c(j,b)$, for $j = 1, J$. These contrast are said to be orthogonal if and only if $\sum_{j=1}^J c_{j,a} c(j,b) = 0$, that is, the sum of cross-products of linear weights must be zero.
- For $\hat{\sigma}^2$, you use the data from *all the groups*, regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination.
- Generally, you do a Bonferroni split on the significance level.

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- 1 A linear combination hypothesis is a *contrast* if and only if the linear weights sum to zero and $a = 0$, i.e., it is of the form

$$H_0 : \sum_j c_j \mu_j = 0, \quad \text{with} \quad \sum_j c_j = 0 \quad (3)$$

- 2 Consider two contrasts A and B defined by linear weights $c_{j,a}$ and $c(j,b)$, for $j = 1, J$. These contrast are said to be orthogonal if and only if $\sum_{j=1}^J c_{j,a} c(j,b) = 0$, that is, the sum of cross-products of linear weights must be zero.
- 3 For $\hat{\sigma}^2$, you use the data from *all the groups*, regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination.
- 4 Generally, you do a Bonferroni split on the significance level.

Planned Contrasts

- Planned Contrasts are performed just like the generalized t statistic we discussed earlier in the course, with a few additional restrictions.
- To begin with, recall that, once we have our data from J independent groups, any hypothesis is defined by a set of linear weights.
- To perform a set of planned orthogonal contrasts (POC), we generally add the following features:

- A linear combination hypothesis is a *contrast* if and only if the linear weights sum to zero and $a = 0$, i.e., it is of the form

$$H_0 : \sum_j c_j \mu_j = 0, \quad \text{with} \quad \sum_j c_j = 0 \quad (3)$$

- Consider two contrasts A and B defined by linear weights $c_{j,a}$ and $c(j,b)$, for $j = 1, J$. These contrast are said to be orthogonal if and only if $\sum_{j=1}^J c_{j,a} c(j,b) = 0$, that is, the sum of cross-products of linear weights must be zero.
- For $\hat{\sigma}^2$, you use the data from *all the groups*, regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination.
- Generally, you do a Bonferroni split on the significance level.

the Scheffe Test

- The Scheffe test allows you to perform *any* contrast hypothesis test *after having viewed the data*.
- It provides FWE protection, and yet also protects against post-hoc cherry-picking of the data.
- This is a lot of protection, but it comes at a cost.
- In general, the larger the family for which “protection is sought,” the more protection is needed and the lower the power of the test will be.

the Scheffe Test

- The Scheffe test allows you to perform *any* contrast hypothesis test *after having viewed the data*.
- It provides FWE protection, and yet also protects against post-hoc cherry-picking of the data.
- This is a lot of protection, but it comes at a cost.
- In general, the larger the family for which “protection is sought,” the more protection is needed and the lower the power of the test will be.

the Scheffe Test

- The Scheffe test allows you to perform *any* contrast hypothesis test *after having viewed the data*.
- It provides FWE protection, and yet also protects against post-hoc cherry-picking of the data.
- This is a lot of protection, but it comes at a cost.
- In general, the larger the family for which “protection is sought,” the more protection is needed and the lower the power of the test will be.

the Scheffe Test

- The Scheffe test allows you to perform *any* contrast hypothesis test *after having viewed the data*.
- It provides FWE protection, and yet also protects against post-hoc cherry-picking of the data.
- This is a lot of protection, but it comes at a cost.
- In general, the larger the family for which “protection is sought,” the more protection is needed and the lower the power of the test will be.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_* - J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_* - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_*-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_* - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

the Scheffe Test

- To perform a Scheffe test, simply compute the generalized t -statistic exactly the same way we first computed it.
- Then, you have to compare the resulting statistic to a new critical value.
- This critical value is

$$S = \sqrt{(J - 1)F_{1-\alpha, J-1, n_{\bullet}-J}} \quad (4)$$

- When computing confidence intervals on a contrast, you also use the S value instead of t .
- **Note:** The J in the above formula is the total number of independent groups you had before concentrating on any contrast hypothesis.
- Some textbooks will square the t statistic to get an F statistic. In general, a squared t statistic has an F distribution with $df = 1, n_{\bullet} - J$.
- So an alternative procedure is to square the t , divide it by $J - 1$, and compare to the standard F critical value.
- The problem with this latter approach is that it does not transform directly into a confidence interval procedure.

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.

1. Look up the critical value $q_{\alpha, m-1, \nu}$ of the *Studentized Range Statistic*.

2. Compare the observed range $r = \max(\bar{y}_i) - \min(\bar{y}_i)$ to $q_{\alpha, m-1, \nu}$.

3. If $r > q_{\alpha, m-1, \nu}$, then reject H_0 and conclude that at least one pair of group means differs.

4. If $r \leq q_{\alpha, m-1, \nu}$, then do not reject H_0 and conclude that all group means are equal.

5. If $r > q_{\alpha, m-1, \nu}$, then compare each pair of group means \bar{y}_i and \bar{y}_j to $q_{\alpha, m-1, \nu}$.

6. If $|\bar{y}_i - \bar{y}_j| > q_{\alpha, m-1, \nu}$, then reject H_0 and conclude that $\mu_i \neq \mu_j$.

7. If $|\bar{y}_i - \bar{y}_j| \leq q_{\alpha, m-1, \nu}$, then do not reject H_0 and conclude that $\mu_i = \mu_j$.

8. Repeat steps 5-7 for all possible pairs of group means.

9. If at least one pair of group means is rejected, then reject H_0 and conclude that at least one pair of group means differs.

10. If no pair of group means is rejected, then do not reject H_0 and conclude that all group means are equal.

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - ① Look up the critical value $q_{J, n_* - J}$ of the *Studentized Range Statistic*.
 - ② Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- ③ Any two means that are farther apart than HSD are declared significant.
- ④ Note that the degrees of freedom for q are J and $n_* - J$, *not the same* as in ANOVA!
- ⑤ Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- ⑥ The Tukey procedure is automated in P.

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - 1 Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - 2 Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- 3 Any two means that are farther apart than HSD are declared significant.
- 4 Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- 5 Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- 6 The Tukey procedure is automated in P

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - 1 Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - 2 Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- 1 Any two means that are farther apart than HSD are declared significant.
- 1 Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- 1 Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- 1 The Tukey procedure is automated in P

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - ① Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - ② Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- ③ Any two means that are farther apart than HSD are declared significant.
- ④ Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- ⑤ Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- ⑥ The Tukey procedure is automated in P

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - ① Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - ② Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- ③ Any two means that are farther apart than HSD are declared significant.
- ④ Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- ⑤ Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- ⑥ The Tukey procedure is automated in R.

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - ① Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - ② Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- ③ Any two means that are farther apart than HSD are declared significant.
- ④ Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- ⑤ Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- ⑥ The Tukey procedure is automated in R

- The Tukey procedure is defined to allow testing in a very specific situation: You are interested in performing all pairwise comparisons between pairs of group means.
- By concentrating on this reduced “family,” you can get full protection at lower cost.
- As originally designed, the Tukey test is performed as follows.
 - ① Look up the critical value $q_{J, n_{\bullet} - J}$ of the *Studentized Range Statistic*.
 - ② Compute the “honestly significant difference” HSD as

$$HSD = q \sqrt{\frac{MSE}{n}} \quad (5)$$

where $MSE = \hat{\sigma}^2$, and n is the sample size per group.

- ③ Any two means that are farther apart than HSD are declared significant.
- ④ Note that the degrees of freedom for q are J and $n_{\bullet} - J$, *not the same* as in ANOVA!
- ⑤ Note also that it is possible to have a significant result from the Tukey test without having a significant overall ANOVA F . (Why? C.P.)
- ⑥ The Tukey procedure is automated in R.