Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 2, 2022

Topics

Objectives

Objectives:

- Distinguish between pairwise and familywise Type I error probabilities, and distinguish between pairwise and familywise levels of confidence.
- Carry out a Bonferroni multiple comparison procedure, and interpret the results.
- Carry out a Tukey multiple comparison procedure, and interpret the results.

One-Factor ANOVA for Population Means $\mu_1, \mu_2, \ldots, \mu_I$ (Cont'd)

Multiple Comparison Tests

- After rejecting H_0 in an ANOVA F test, we can determine **which** means differ from each other using a ***multiple comparison*** procedure.

One-Factor ANOVA for Population Means $\mu_1, \mu_2, \ldots, \mu_I$ (Cont'd)

Multiple Comparison Tests

- After rejecting H_0 in an ANOVA F test, we can determine **which** means differ from each other using a ***multiple comparison*** procedure.
- The total number of *pairwise* comparisons of means is

$$\binom{I}{2} = \frac{I!}{2!(I-2)!} = \frac{I(I-1)}{2}.$$

Example

For the lead measurements made at 5 labs, if we want to know *which* labs differ from each other, we'd need to make

$$\frac{I(I-1)}{2} = \frac{5(5-1)}{2} = 10$$

comparisons, namely

Lab1 vs Lab2
Lab1 vs Lab3
Lab1 vs Lab4
Lab1 vs Lab5
Lab2 vs Lab3
Lab2 vs Lab4
Lab2 vs Lab5
Lab3 vs Lab4
Lab3 vs Lab5
Lab4 vs Lab5

- It's ***not*** **appropriate** to carry out *multiple* two-sample t tests, each at level $\alpha = 0.05$, say.

- It's ***not*** **appropriate** to carry out *multiple* two-sample t tests, each at level $\alpha = 0.05$, say.

  Although the **Type I error probability** would be **0.05** on any ***particular*** t test, ...

- It's ***not*** **appropriate** to carry out *multiple* two-sample t tests, each at level $\alpha = 0.05$, say.

  Although the **Type I error probability** would be **0.05** on any ***particular*** t test, ...

  the **probability** of making ***at least one*** **Type I error** among the ***family*** of t tests would be substantially **greater than 0.05**.

Example

For the five labs, suppose the null hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

was true, and that **ten separate** two-sample t tests are performed, each at level $\alpha = 0.05$.

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$P(\text{At least one Type I error}) = 1 - P(\text{No Type I errors})$$

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{No Type I errors}) \\ &= 1 - P(\text{All 10 tests fail to reject } H_0) \end{aligned}$$

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{No Type I errors}) \\ &= 1 - P(\text{All 10 tests fail to reject } H_0) \\ &= 1 - (1 - 0.05)^{10} \end{aligned}$$

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{No Type I errors}) \\ &= 1 - P(\text{All 10 tests fail to reject } H_0) \\ &= 1 - (1 - 0.05)^{10} \\ &= 0.40, \end{aligned}$$

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{No Type I errors}) \\ &= 1 - P(\text{All 10 tests fail to reject } H_0) \\ &= 1 - (1 - 0.05)^{10} \\ &= 0.40, \end{aligned}$$

which is unacceptable.

If the outcomes of the t tests were *independent* of each other*, the probability of making **at least one** Type I error would be

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{No Type I errors}) \\ &= 1 - P(\text{All 10 tests fail to reject } H_0) \\ &= 1 - (1 - 0.05)^{10} \\ &= 0.40, \end{aligned}$$

which is unacceptable.

* In reality, the t tests *aren't* independent of each other because each sample is used in several of the tests. Thus the probability 0.40 above is only an approximation.

- In general, if m *independent** two-sample t tests were performed, each at level α, the **probability** that ***at least one*** of them would result in a **Type I error** would be

$$P(\text{at least one Type I error}) = 1 - (1-\alpha)^m.$$

- In general, if m *independent** two-sample t tests were performed, each at level α, the **probability** that ***at least one*** of them would result in a **Type I error** would be

$$P(\text{at least one Type I error}) = 1 - (1-\alpha)^m.$$

- Similarly, if we compute m *independent** two-sample t CIs for $\boldsymbol{\mu_i - \mu_j}$, each with confidence level **95%**, say, and check which ones contain **zero**.

- Similarly, if we compute m *independent** two-sample t CIs for $\mu_i - \mu_j$, each with confidence level **95%**, say, and check which ones contain **zero**.

  If all $\mu_i - \mu_j$'s were in reality **zero**, then although the **probability** of any ***particular*** CI containing **zero** would be **0.95**, the probability of ***all*** of them containing **zero** would only be 0.95^m.

- Similarly, if we compute m *independent** two-sample t CIs for $\mu_i - \mu_j$, each with confidence level **95%**, say, and check which ones contain **zero**.

  If all $\mu_i - \mu_j$'s were in reality **zero**, then although the **probability** of any ***particular*** CI containing **zero** would be **0.95**, the probability of ***all*** of them containing **zero** would only be $\mathbf{0.95^m}$.

***** In reality, the t tests and CIs *aren't* independent of each other because each sample is used in several of the tests or CIs. Thus the probabilities $1 - (1 - \alpha)^m$ and 0.95^m above are only approximations.

Pairwise and Familywise Type I Error Rates

- Suppose I population means are being tested for differences $\mu_i - \mu_j$ one pair at a time.

Pairwise and Familywise Type I Error Rates

- Suppose I population means are being tested for differences $\mu_i - \mu_j$ one pair at a time.

  The ***pairwise Type I error rate*** is the **probability** that any ***particular*** pairwise test will result in a **Type I error**.

Pairwise and Familywise Type I Error Rates

- Suppose I population means are being tested for differences $\mu_i - \mu_j$ one pair at a time.

  The ***pairwise Type I error rate*** is the **probability** that any ***particular*** pairwise test will result in a **Type I error.**

  The ***overall*** (or ***familywise***) ***Type I error rate*** is the **probability** that ***at least one*** of the tests will result in a **Type I error.**

- Likewise, if CIs are being constructed for the differences $\mu_i - \mu_j$ one pair at a time, the ***pairwise level of confidence*** is the **probability** that any ***particular*** CI will contain the true difference.

- Likewise, if CIs are being constructed for the differences $\mu_i - \mu_j$ one pair at a time, the ***pairwise level of confidence*** is the **probability** that any ***particular*** CI will contain the true difference.

  The ***overall*** (or ***familywise***) ***level*** is the **probability** that ***all*** of them will contain their true difference.

- We'll denote the **overall** (**familywise**) **Type I error rate** by α_f and the **pairwise Type I error rate** by α_p.

- We'll denote the **overall** (**familywise**) **Type I error rate** by α_f and the **pairwise Type I error rate** by α_p.
- The goal in a ***multiple comparison procedure*** is to hold the **familywise Type I error rate** at a fixed level, say $\alpha_f = 0.05$, or equivalently to control the **familywise confidence level** at, say, **95%.**

The Bonferroni Procedure

- The **Bonferroni procedure** holds the **familywise Type I error rate** at a fixed level (usually $\alpha_f = 0.05$) by using a sufficiently small level of significance α_p for each pairwise test of hypotheses

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned}$$

- More specifically, it divides the **familywise Type I** error rate equally among the pairwise tests.

- More specifically, it divides the **familywise Type I** error rate equally among the pairwise tests.

  Thus, for example, to perform the **10** pairwise tests comparing the **five labs**, we'd use level of significance

  $$\alpha_p = \frac{0.05}{10} = 0.005$$

  for each test.

Bonferroni Procedure After an ANOVA F Test

- The next slide gives the **Bonferroni procedure** after the null hypothesis is rejected in an **ANOVA F test.**

Bonferroni Procedure After an ANOVA F Test

- The next slide gives the **Bonferroni procedure** after the null hypothesis is rejected in an **ANOVA F test.**

  It merely involves doing **multiple two-sample t tests**, but with two adjustments:

Bonferroni Procedure After an ANOVA F Test

- The next slide gives the **Bonferroni procedure** after the null hypothesis is rejected in an **ANOVA F test.**

  It merely involves doing **multiple two-sample t tests**, but with two adjustments:

  1. We use the **Bonferroni-corrected** level of significance on each test.

Bonferroni Procedure After an ANOVA F Test

- The next slide gives the **Bonferroni procedure** after the null hypothesis is rejected in an **ANOVA F test.**

  It merely involves doing **multiple two-sample t tests**, but with two adjustments:

  1. We use the **Bonferroni-corrected** level of significance on each test.
  2. We use the **square root** of the **MSE** in place of S_i and S_j in the t **test statistics.**

Bonferroni Multiple Comparison Procedure After One-Factor ANOVA: To decide which pairs of means differ while controlling the familywise Type I error rate at α_f, for each pair of means μ_i and μ_j, test the hypotheses

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned}$$

using the ***Bonferroni pairwise t test statistic***

$$T = \frac{\bar{Y}_i - \bar{Y}_j - 0}{\sqrt{\frac{\text{MSE}}{n} + \frac{\text{MSE}}{n}}} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\frac{2 \cdot \text{MSE}}{n}}}$$

and decision rule

$$\begin{array}{l} \text{Reject } H_0 \text{ if p-value } < \alpha_p \\ \text{Fail to reject } H_0 \text{ if p-value } \geq \alpha_p\,, \end{array}$$

where

$$\alpha_p = \frac{\alpha_f}{(I(I-1)/2)}.$$

When the corresponding H_0 is true, the test statistic T follows a $t(I(J-1))$ distribution, from which the p-value for that test is obtained.

Example

For the study of lead measurements at five labs, we'll use the **Bonferroni procedure** to decide ***which*** labs' means differ from each other, while controlling the **familywise Type I error rate** at $\alpha_f = 0.05$.

Example

For the study of lead measurements at five labs, we'll use the **Bonferroni procedure** to decide ***which*** labs' means differ from each other, while controlling the **familywise Type I error rate** at $\alpha_f = 0.05$.

We need to test **10** sets of hypotheses of the form

$$\begin{aligned} H_0 : \mu_i - \mu_j &= 0 \\ H_a : \mu_i - \mu_j &\neq 0 \end{aligned}$$

Because $I = 5$, the **Bonferroni-corrected level of significance** to use for each **pairwise test** is

$$\boldsymbol{\alpha_p} = \frac{0.05}{5(5-1)/2} = \mathbf{0.005},$$

Because $I = 5$, the **Bonferroni-corrected level of significance** to use for each **pairwise test** is

$$\boldsymbol{\alpha_p} = \frac{0.05}{5(5-1)/2} = \mathbf{0.005},$$

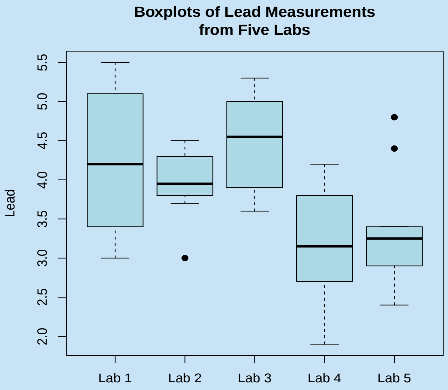
and so the decision rule is

Reject H_0 if p-value < 0.005
Fail to reject H_0 if p-value ≥ 0.005

Statistical software reports the results of **all 10 pairwise tests.** Statistically significant differences (at the Bonferroni-corrected significance level $\alpha_p = 0.005$) are marked with an asterisk.

Pair of Means	t	P-value
Lab1 vs Lab2	1.03	0.3070
Lab1 vs Lab3	-0.50	0.6188
Lab1 vs Lab4	3.69	0.0006*
Lab1 vs Lab5	3.01	0.0043*
Lab2 vs Lab3	-1.53	0.1320
Lab2 vs Lab4	2.66	0.0107
Lab2 vs Lab5	1.97	0.0547
Lab3 vs Lab4	4.20	0.0001*
Lab3 vs Lab5	3.51	0.0010*
Lab4 vs Lab5	-0.69	0.4945

We conclude that **Labs 1** and **4** differ, **Labs 1** and **5** differ, **Labs 3** and **4** differ, and **Labs 3** and **5** differ.


Boxplots of Lead Measurements
from Five Labs
Lead
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Lab 1
Lab 2
Lab 3
Lab 4
Lab 5

Tukey's Multiple Comparison Procedure

- In ***Tukey's multiple comparison procedure***, we construct CIs for all pairwise differences $\mu_i - \mu_j$ in such a way that the ***familywise level of confidence*** is $100(1 - \alpha_f)\%$ (where usually $\alpha_f = 0.05$).

Tukey's Multiple Comparison Procedure

- In ***Tukey's multiple comparison procedure***, we construct CIs for all pairwise differences $\mu_i - \mu_j$ in such a way that the ***familywise level of confidence*** is $100(1-\alpha_f)\%$ (where usually $\alpha_f = 0.05$).

  This says that the **probability** that ***all*** of the CIs will ***simultaneously*** contain their true $\mu_i - \mu_j$'s is $1 - \alpha_f$.

Tukey's Multiple Comparison Procedure

- In ***Tukey's multiple comparison procedure***, we construct CIs for all pairwise differences $\mu_i - \mu_j$ in such a way that the ***familywise level of confidence*** is $100(1 - \alpha_f)\%$ (where usually $\alpha_f = 0.05$).

  This says that the **probability** that ***all*** of the CIs will ***simultaneously*** contain their true $\mu_i - \mu_j$'s is $1 - \alpha_f$.

- We'll need the following fact.

Proposition

Suppose the assumptions of the ANOVA F test are met (i.e. independent samples from N(μ_i, σ) distributions), and that the samples are all of size J. Then the random variable

$$Q = \frac{\max_{i,j}\{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - (\mu_i - \mu_j)\}}{\sqrt{\frac{MSE}{J}}}$$

follows a so-called ***Studentized range distribution*** with $\boldsymbol{I}$ ***numerator degrees of freedom*** and $I(J-1)$ ***denominator degrees of freedom***, which we'll denote by $Q(\boldsymbol{I},\ \boldsymbol{I}(\boldsymbol{J}-1))$.

- Using the above fact, it can be shown that **with probability $\mathbf{1-\boldsymbol{\alpha}}$**, ***all*** of the pairwise differences $\mu_i - \mu_j$ will *simultaneously* satisfy

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - Q_{\alpha_f, I, I(J-1)}\sqrt{\frac{MSE}{J}} \;\le\; \mu_i - \mu_j \;\le\; \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + Q_{\alpha_f, I, I(J-1)}\sqrt{\frac{MSE}{J}},$$

where $\boldsymbol{Q_{\alpha_f, I, I(J-1)}}$ is the $100(1-\alpha_f)$th percentile of the $Q(I,\, I(J-1))$ distribution.

Tukey's Multiple Comparison Procedure: *After* the ANOVA F test rejects H_0:

1. Choose an **overall familywise confidence level** $100(1-\alpha_f)\%$ (usually $\alpha_f = 0.05$ for a 95% confidence level).
2. Compute the $I(I-1)/2$ **CIs**:

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \ \pm \ Q_{\alpha_f, I, I(J-1)} \sqrt{\frac{MSE}{J}}\,. \tag{1}$$

3. For any interval that **doesn't contain zero**, deem those means μ_i and μ_j to be **different.**

- In practice, **Tukey's multiple comparison procedure** is carried out using statistical software.

Example

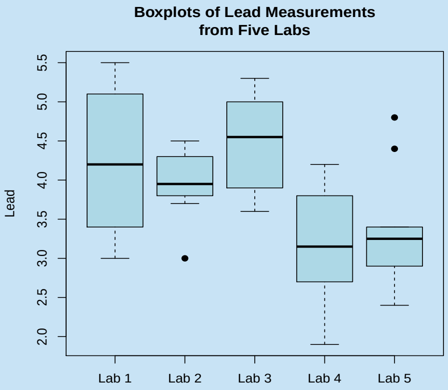
For the study comparing lead measurements at five labs, the **Tukey procedure** in R produces the following CIs:

Labs	Difference	Lower End Pt	Upper End Pt	
Lab2-Lab1	-0.33	-1.2373875	0.57738749	
Lab3-Lab1	0.16	-0.7473875	1.06738749	
Lab4-Lab1	-1.18	-2.0873875	-0.27261251	*
Lab5-Lab1	-0.96	-1.8673875	-0.05261251	*
Lab3-Lab2	0.49	-0.4173875	1.39738749	
Lab4-Lab2	-0.85	-1.7573875	0.05738749	
Lab5-Lab2	-0.63	-1.5373875	0.27738749	
Lab4-Lab3	-1.34	-2.2473875	-0.43261251	*
Lab5-Lab3	-1.12	-2.0273875	-0.21261251	*
Lab5-Lab4	0.22	-0.6873875	1.12738749	

Intervals marked with asterisks don't contain zero.

Intervals marked with asterisks don't contain zero.

We conclude that **Lab 1** differs from both **Labs 4** and **5**, and **Lab 3** differs from **Labs 4** and **5**, but no other differences exist.


Boxplots of Lead Measurements
from Five Labs
Lead
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Lab 1
Lab 2
Lab 3
Lab 4
Lab 5