# Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

*ngrevsta@msudenver.edu*

October 1, 2019

# Topics

# Objectives

**Objectives**:

- Distinguish between pairwise and familywise Type I error probabilities, and distinguish between pairwise and familywise levels of confidence.
- Carry out Tukey's 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 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 $\alpha = \mathbf{0.05}$.

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

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

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 $\alpha$, 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 $\alpha$, the **probability** that ***at least one*** of them would result in a **Type I error** would be

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

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

- The issue persists if we compute $m$ **two-sample $t$ CIs** for $\mu_i - \mu_j$, each with level of confidence **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 $0.95^m$.

- The issue persists if we compute $m$ **two-sample $t$ CIs** for $\mu_i - \mu_j$, each with level of confidence **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 intervals. 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.

- The goal in a ***multiple comparison procedure*** is to hold the **familywise Type I error rate** at a fixed level, say **0.05**, or alternatively to control the **familywise confidence level** at, say, **95%.**

## 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 $100(1-\alpha)\%$.

## 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 $100(1-\alpha)\%$.

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

## 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 $100(1-\alpha)\%$.

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

- We'll need the following fact.

## Proposition

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

$$Q_{I,I(J-1)} = \frac{\max_{i,j}\{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - (\mu_i - \mu_j)\}}{\sqrt{\frac{MSE}{J}}}$$

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

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

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

where $Q_{\alpha,I,I(J-1)}$ is the $100(1-\alpha)$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** $100(1-\alpha)\%$ (usually $\alpha = 0.05$ for a 95% confidence level).
2. Compute the $I(I-1)/2$ **confidence intervals**:

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

3. For any interval that **doesn't contain zero**, deem those means $\mu_i$ and $\mu_j$ 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.