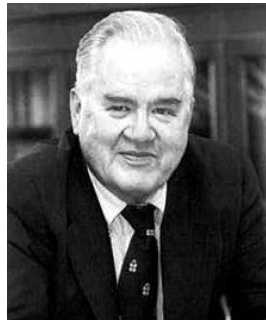


Adjusted P-values for Multiple Comparisons



Emilio Bonferroni



John Tukey



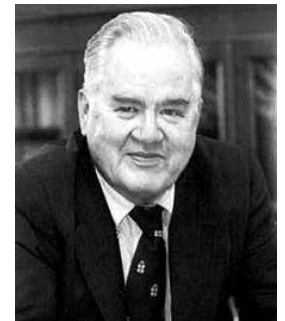
Emilio Bonferroni



Henry Scheffé



Henry Scheffé



John Tukey

Reminder: Pairwise Comparisons

```
with(fish, pairwise.t.test(Percentage, Pair, p.adj = "none"))
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Percentage and Pair  
##  
##      Pair1 Pair2 Pair3 Pair4 Pair5  
## Pair2 0.431 -      -      -      -  
## Pair3 0.267 0.783 -      -      -  
## Pair4 0.065 0.299 0.415 -      -  
## Pair5 0.229 0.616 0.781 0.674 -  
## Pair6 0.224 0.676 0.871 0.532 0.895  
##  
## P value adjustment method: none
```

0.431: the p-value of the Pair1-Pair2 comparison

=> Cannot reject the hypothesis that the means for Pair1 and Pair2 are the same at 95% confidence since $0.431 > 0.05$

Reminder: Bonferroni Correction

- $P\left(\bigcup_{i=1}^n \left(p_i \leq \frac{\alpha}{n}\right)\right) \leq \sum_{i=1}^n P\left(p_i \leq \frac{\alpha}{n}\right) = \frac{n\alpha}{n} = \alpha$
- If we want the *familywise* p-value threshold to be α , make the individual p-value threshold be $\frac{\alpha}{n}$, where n is the number of comparisons

Bonferroni Correction Output

```
with(fish, pairwise.t.test(Percentage, Pair, p.adj = "bonf"))
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Percentage and Pair  
##  
##      Pair1 Pair2 Pair3 Pair4 Pair5  
## Pair2 1.00  -    -    -    -  
## Pair3 1.00  1.00 -    -    -  
## Pair4 0.97  1.00 1.00 -    -  
## Pair5 1.00  1.00 1.00 1.00 -  
## Pair6 1.00  1.00 1.00 1.00 1.00  
##  
## P value adjustment method: bonferroni
```

Bonferroni Adjustment

- $P \left(\bigcup_{i=1 \dots n}^n \left(p_i \leq \frac{\alpha}{n} \right) \right) < \alpha$
- $P \left(\bigcup_{i=1 \dots 15}^{15} \left(p_i \leq \frac{0.05}{15} \right) \right) < 0.05 \quad 15 = \binom{6}{2}$
- $P \left(\bigcup_{i=1 \dots 15}^{15} (15p_i \leq 0.05) \right) < 0.05$
- => Need to multiply all p-values by 15 in order to be able to say that a difference is significant using Bonferroni correction

```
with(fish, pairwise.t.test(Percentage, Pair, p.adj = "bonf"))
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Percentage and Pair  
##  
##      Pair1 Pair2 Pair3 Pair4 Pair5  
## Pair2 1.00 - - - -  
## Pair3 1.00 1.00 - - -  
## Pair4 0.97 1.00 1.00 - -  
## Pair5 1.00 1.00 1.00 1.00 -  
## Pair6 1.00 1.00 1.00 1.00 1.00  
##  
## P value adjustment method: bonferroni
```

$$0.97 \approx 0.065 \times 15$$

$$1 < 0.431 \times 15$$

```
with(fish, pairwise.t.test(Percentage, Pair, p.adj = "none"))
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Percentage and Pair  
##  
##      Pair1 Pair2 Pair3 Pair4 Pair5  
## Pair2 0.431 - - - -  
## Pair3 0.267 0.783 - - -  
## Pair4 0.065 0.299 0.415 - -  
## Pair5 0.229 0.616 0.781 0.674 -  
## Pair6 0.224 0.676 0.871 0.532 0.895  
##  
## P value adjustment method: none
```

Probability of a false positive

- For an individual t-test, assuming the null hypothesis is true:
 - $P(p_i < 0.05) = ?$

Probability of a false positive

- For an individual t-test, assuming the null hypothesis is true:
 - $P(p_i < 0.05) = ?$
- Suggestion from Piazza, for multiple t-tests:
 - $P(\cup_{i=1\dots n} (p_i \leq 0.05)) =$
 $1 - P(\cap_{i=1\dots n} (p_i > 0.05)) = 1 - 0.95^n$
- *That would only work if $(p_1 \leq 0.05), (p_2 \leq$*

Non-Independence of pairwise t-Tests

- Assume $\mu_{Pair1} = \mu_{Pair2} = \mu_{Pair3} = \mu_{Pair4} = \dots$
- If the sample mean for Pair1 is unusually large, that will influence (at least) all the comparisons that involve Pair1
- So the p-values in the multiple comparisons table are not independent
- **Problem:** If the p-value for the Pair1-Pair2 comparison is small, is the p-value for the Pair1-Pair3 comparison likely to be large or small?
 - Assume the true means are all equal

```
with(fish, pairwise.t.test(Percentage, Pair, p.adj = "none"))
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Percentage and Pair  
##  
##      Pair1 Pair2 Pair3 Pair4 Pair5  
## Pair2 0.431 - - - -  
## Pair3 0.267 0.783 - - -  
## Pair4 0.065 0.299 0.415 - -  
## Pair5 0.229 0.616 0.781 0.674 -  
## Pair6 0.224 0.676 0.871 0.532 0.895  
##  
## P value adjustment method: none
```

Reminder

- The Problem with Multiple Comparisons:
 - Looking at multiple p-values and reporting the results when you see a small p-value increases the probability of rejecting *some* null hypothesis even if all the null hypotheses are true
 - True for any kind of set of p-values, even though we were looking specifically at pairwise comparisons of means
- Not a problem if all the comparisons are *pre-planned*
 - But then you have to report that you were planning on performing all the comparisons
 - The reader of your study can then decide that your study is likely wrong in some respect, though, since in effect you're performing multiple studies (one per hypothesis)
- The F-test in ANOVA is a single test that tests the hypothesis that *all the means are the same*
 - Rejecting the hypothesis means that there is at least one difference, but you don't know which
 - Can follow up to find which differences are significant using Tukey's HSD adjustment