Unit 3: Inference for Categorical and Numerical Data

3. Difference of many means

(Chapter 4.4)

3/2/2020

Recap

1. We can use the t-distribution to estimate the probability of a difference between unpaired values.

2. Degrees of freedom depends on the size of both samples

3. The right test depends on where you think variance comes from

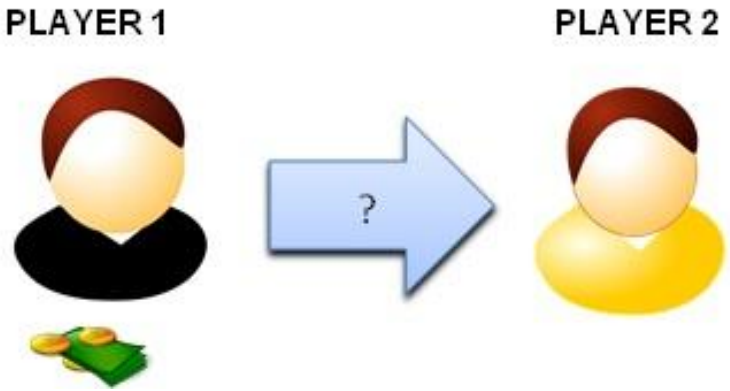
Key ideas

1. If you have multiple groups, you don’t want to just use multiple t-tests.

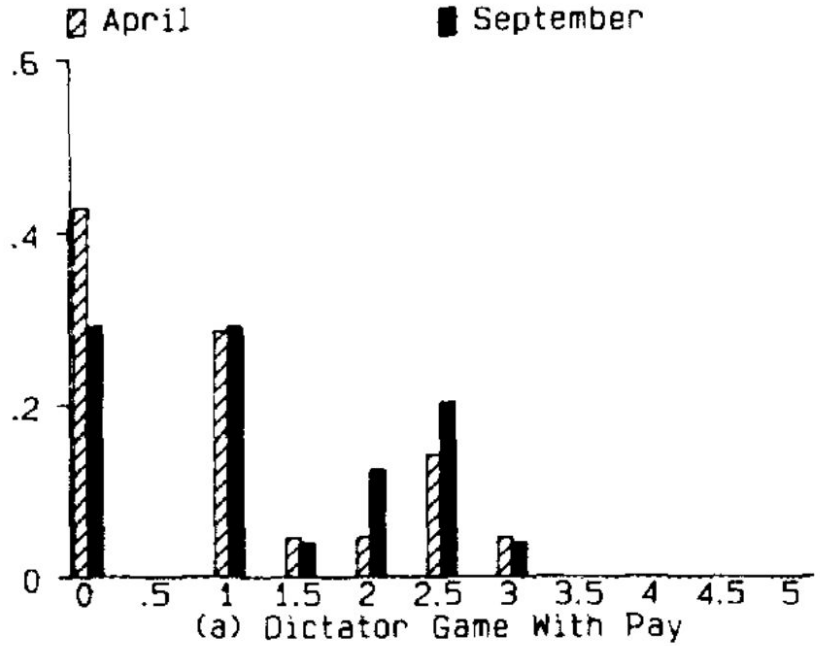
2. Analysis of variance is a method for comparing many means

3. If you want to compare specific groups, you can use corrections that control for false alarm rates

The Dictator Game (Forsyth et al., 1998)

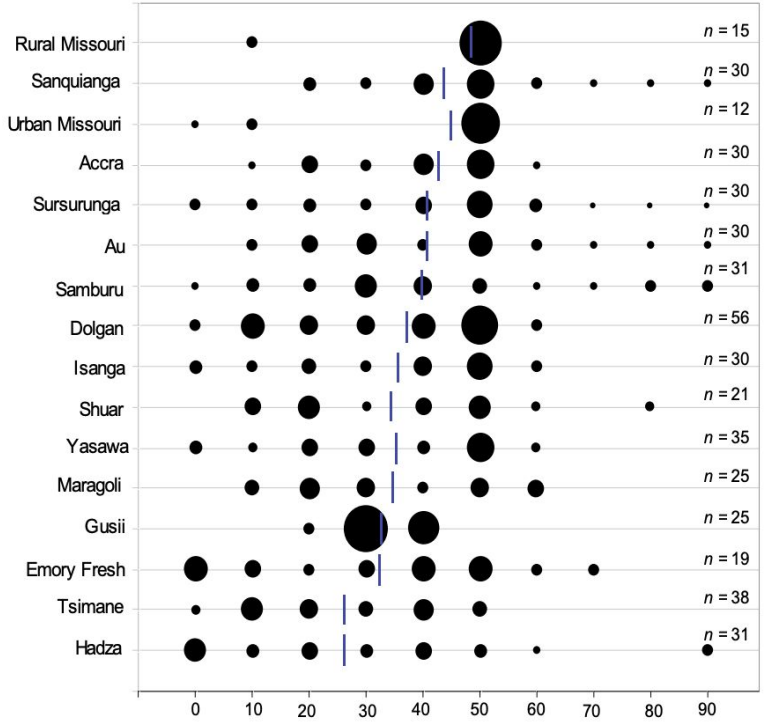


How much of the $10 would you give to Player 2?



https://en.wikibooks.org/wiki/Bestiary_of_Behavioral_Economics/Dictator_Game

Does giving vary across cultures?



Henrich et al. (2006)

Practice question 1

Suppose α = 0.05. What is the probability of making a Type 1 error and rejecting a null hypothesis like

$$H_0: \mu_{\text{rural Missouri}} - \mu_{\text{Sanquianga}} = 0$$

when it is actually true?

a) 1%

b) 5%

c) 36%

d) 64%

e) 95%

f) >99%

Practice question 1

Suppose α = 0.05. What is the probability of making a Type 1 error and rejecting a null hypothesis like

$$H_0: \mu_{\text{rural Missouri}} - \mu_{\text{Sanquianga}} = 0$$

when it is actually true?

a) 1%

b) 5%

c) 36%

d) 64%

e) 95%

f) >99%

Practice question 2

Suppose we want to test all of these 16 different cultures against each-other to see if any are different

$$H_0: \mu_{\text{rural Missouri}} - \mu_{\text{Sanquianga}} = 0$$

$$H_0: \mu_{\text{Accra}} - \mu_{\text{Sursurunga}} = 0$$

$$H_0: \mu_{\text{Isanga}} - \mu_{\text{Maragoli}} = 0$$

...

What is the probability of making at least 1 type 1 Error?

a) 1%

b) 5%

c) 36%

c) 64%

d) 95%

e) >99%

Practice question 2

Suppose we want to test all of these 16 different cultures against each-other to see if any are different

$$H_0: \mu_{\text{rural Missouri}} - \mu_{\text{Sanquianga}} = 0$$

$$H_0: \mu_{\text{Accra}} - \mu_{\text{Sursurunga}} = 0$$

$$H_0: \mu_{\text{Isanga}} - \mu_{\text{Maragoli}} = 0$$

...

What is the probability of making at least 1 type 1 Error?

a) 1%

b) 5%

c) 36%

c) 64%

d) 95%

e) >99%

Analysis of Variance (ANOVA)

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category *i*

H_A : At least one mean is different than others

Conditions for Analysis of Variance

Independence within groups

The people in each society were samples independently

Independence between groups

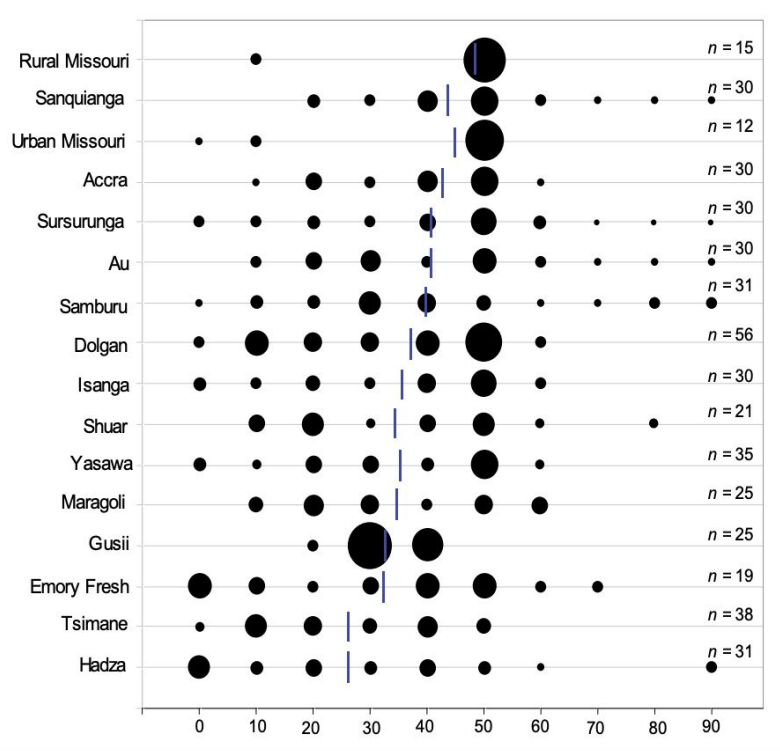
No one was in more than one society

Samples should be nearly normal

A little bit questionable (see e.g. Rural MI)

Groups should similar variance

A little bit questionable (see e.g. Rural MI)



z/t vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

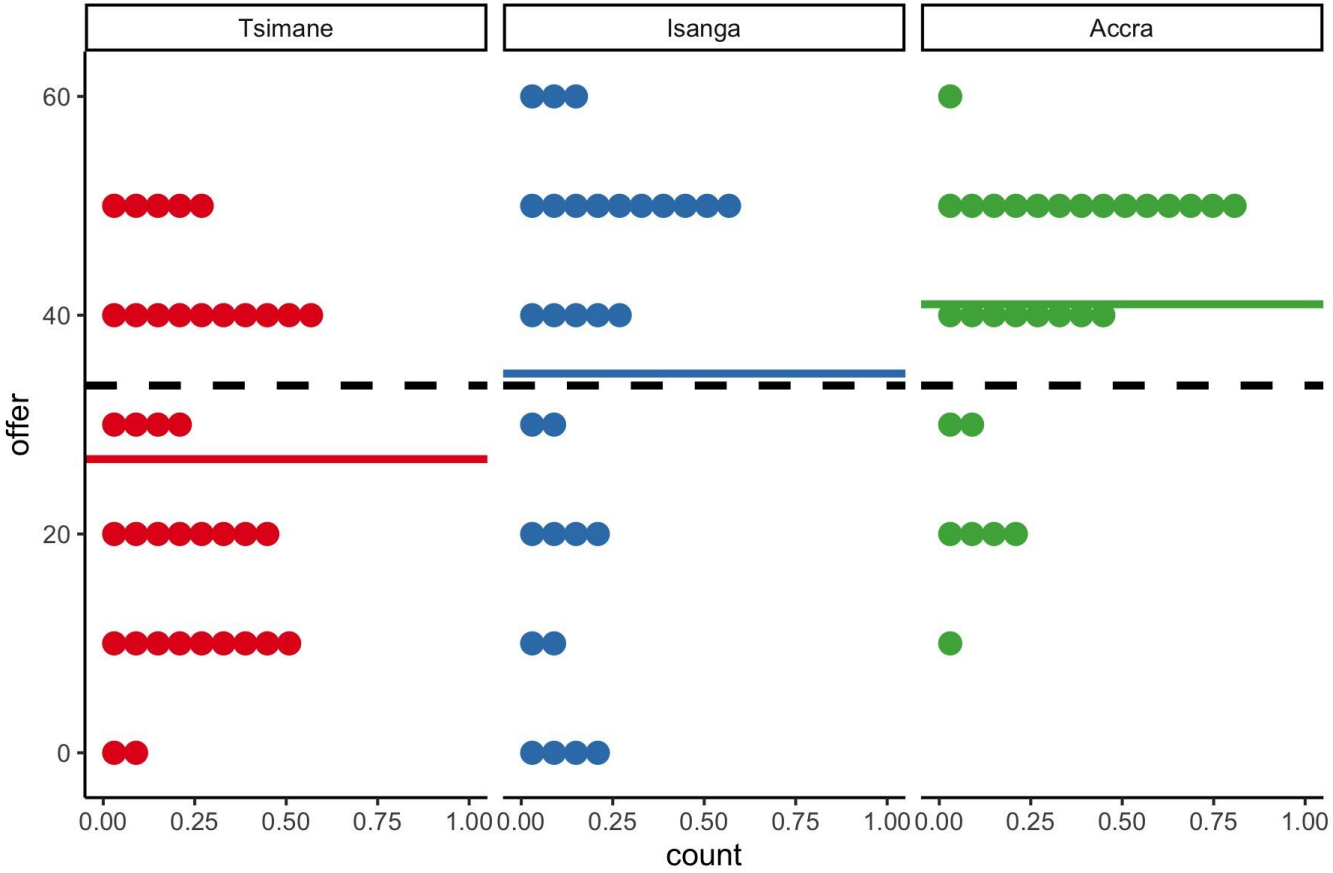
Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

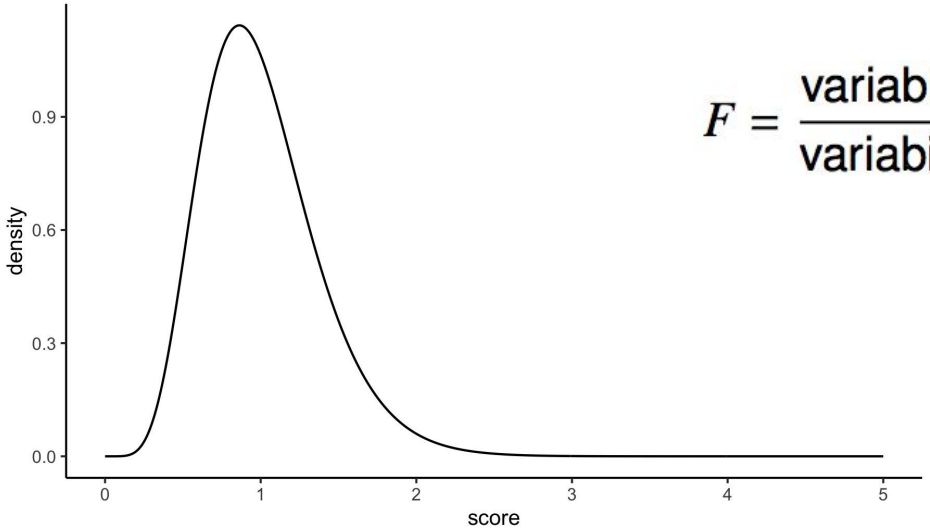
Large test statistics lead to small p-values.

If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

Within and between group variance



F-distribution and p-values



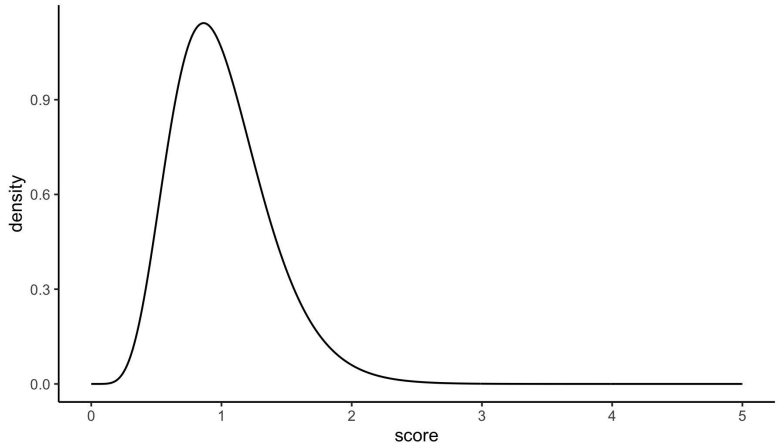
$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

The F-distribution gives the probability that between-group variability will be high while within-group variability will be low if H_0 is true

Where is the peak of the distribution?

F-distribution and p-values

The F-distribution depends on two factors:
(1) The number of categories ***k***
(2) number of data points **n**



F-has two parameters:
$df_1 = k - 1$,
$df_2 = n - k - 1$

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

ANOVA in R

```
> culture_anova <- aov(offer ~ culture, data = tidy_data)
> summary(culture_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

ANOVA output: Degrees of freedom

```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

Degrees of freedom associated with ANOVA

- Groups: $df_G = k - 1$, where k is the number of groups
- Total: $df_T = n - 1$, where n is the total sample size
- Error: $df_E = df_T - df_G$

- $df_G = k - 1 = 16 - 1 = 15$
- $df_T = n - 1 = 475 - 1 = 474$
- $df_E = 474 - 15 = 459$

ANOVA output: Sum of Squares

```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

Sum of Squares between groups (SSG) measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group,
$\bar{x}$ is the overall (grand) mean.

SSG = 15 x (47.3 - 36.02)^2 +
30 x (46.3 - 36.02)^2 +
12 x (43.3 - 36.02)^2 + ...

	mean	n
rural MI	47.3	15
Sanquianga	46.3	30
Urban MI	43.3	12
overall	36.02	475

ANOVA output: Sum of Squares

```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

Sum of Squares between groups (SST)
measures the variability across all observations

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

SST = $(50 - 36.02)^2 + (10 - 36.02)^2 + (30 - 36.02)^2 + (50 - 36.02)^2 + \ldots$

Sum of Squares error (SSE)
measures the variability within groups

$$SSE = SST - SSG$$

ANOVA output: Mean squared error

```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

Mean Square Error (MSE)

Calculated as sum of squares divided by the degrees of freedom.

$MSG = SSG / DF_g = 21283/15 = 1418.9$

$MSE = SSE / DF_E = 142697/459 = 310.9$

ANOVA output: F-value

```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

Test statistic - F

The ratio between within group variability and between group variability

$$F = \frac{MSG}{MSE}$$

$$F = \frac{1418.9}{310.9} = 4.564$$

ANOVA output: p-value

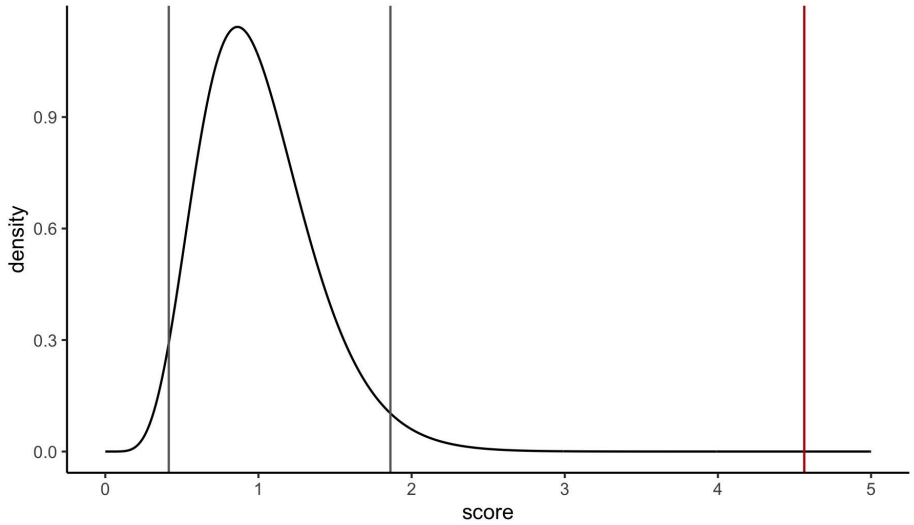
```
> summary(culture_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08 ***
Residuals   459 142697  310.9
```

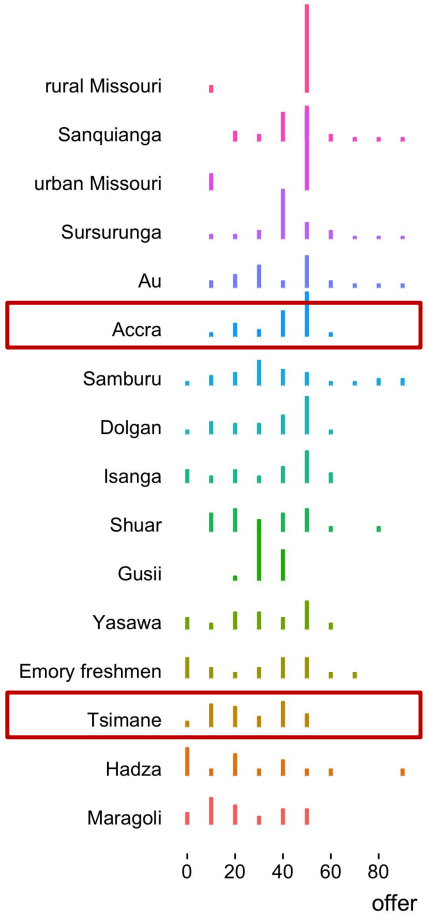
p-value

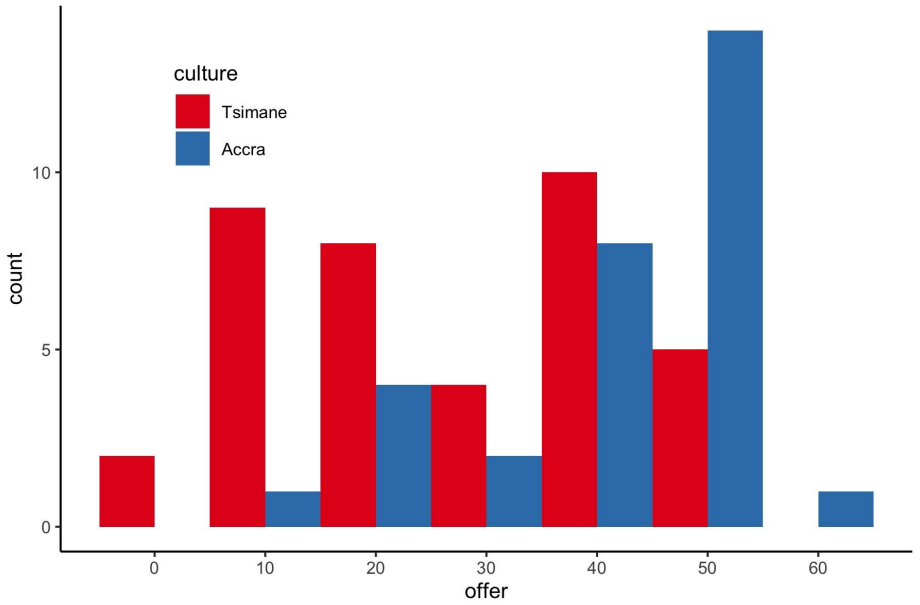
probability of at least as large a ratio between the “between group” and “within group” variability, if the means of all groups are equal.

It's calculated the same was as with the Normal and t-distributions, but with the F-distribution instead



But which groups are different?





Using corrected t-tests: Bonferonni's correction

If the ANOVA yields a significant results, next natural question is:
"Which means are different?"

Use t-tests comparing each pair of means to each other,

- with a common variance (*MSE* from the ANOVA table) instead of each group's variances in the calculation of the standard error,
- and with a common degrees of freedom (df_E from the ANOVA table)

Compare resulting p-values to a modified significance level

$$\alpha^* = \frac{\alpha}{K}$$

where K is the total number of pairwise tests

Post-hoc tests

If we *knew* we wanted to test only Tsimane vs. Accra, we're only doing one test. But then why did we gather all of this other data?

If we are doing our analyses post-hoc, we are implicitly saying something like "I want to compare the groups that look most different", which is like doing all of those other tests and then rejecting them.

In that case, we are actually doing $\frac{K(K-1)}{2}$ tests.

So our $\alpha^* = \frac{.05}{(16 \cdot 15)/2} = 0.0004$

Comparing Tsimane and Accra

```
> summary(culture_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08
Residuals   459 142697  310.9
```

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

$$T_{df_E} = \frac{(\bar{x}_{Accra} - \bar{x}_{Tsimane})}{\sqrt{\frac{MSE}{n_{Accra}} + \frac{MSE}{n_{Tsimane}}}}$$

$$T_{459} = \frac{(41 - 26.8)}{\sqrt{\frac{311}{30} + \frac{311}{38}}} = \frac{14.2}{4.31} \sim 3.3$$

```
> qt(.975, 459) = 1.97
```

Should I reject the null hypothesis?

No! That's the wrong critical value

Comparing Tsimane and Accra

```
> summary(culture_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
culture      15  21283  1418.9   4.564 3.86e-08
Residuals   459 142697  310.9
```

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

$$T_{df_E} = \frac{(\bar{x}_{Accra} - \bar{x}_{Tsimane})}{\sqrt{\frac{MSE}{n_{Accra}} + \frac{MSE}{n_{Tsimane}}}}$$

$$T_{459} = \frac{(41 - 26.8)}{\sqrt{\frac{311}{30} + \frac{311}{38}}} = \frac{14.2}{4.31} \sim 3.3$$

```
> qt(.9998, 459) = 3.57
```

Should I reject the null hypothesis?

No. After the correction, this is not significantly different from chance

Key ideas

1. If you have multiple groups, you don’t want to just use multiple t-tests.

2. Analysis of variance is a method for comparing many means

3. If you want to compare specific groups, you can use corrections that control for false alarm rates