Computing ANOVA

DR KYAW OO

Focus

Dr Kyaw Oo

In the one-way analysis of variance table for the comparison of three groups:

- (a) the group mean square + the error mean square = the total mean square;
- (b) there are two degrees of freedom for groups;
- (c) the group sum of squares + the error sum of squares = the total sum of squares;
- (d) the numbers in each group must be equal;
- (e) the group degrees of freedom + the error degrees of freedom = the total degrees of freedom.

k_score	group
Xi	i
5	A
6	A
6	A
7	A
7	A
8	A
9	A
10	A
7	B
7	B
8	B
9	B
9	B
10	B
10	B
11	B
7	C
9	C
9	C
10	C
10	C
10	C
11	C
12	C
13	C

How many subjects?

How many variables?

How many groups?

Dr Kyaw Oo

An example ANOVA situation

Subjects: 25 respondents with knowledge score
Group: Group A, Group B, Group C
Measurement: # of days until blisters heal

Data [and means]:

- A: 5,6,6,7,7,8,9,10 [7.25]
- B: 7,7,8,9,9,10,10,11 [8.875]
- C: 7,9,9,10,10,10,11,12,13 [10.11]

Are these differences significant?

Informal Investigation

Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes

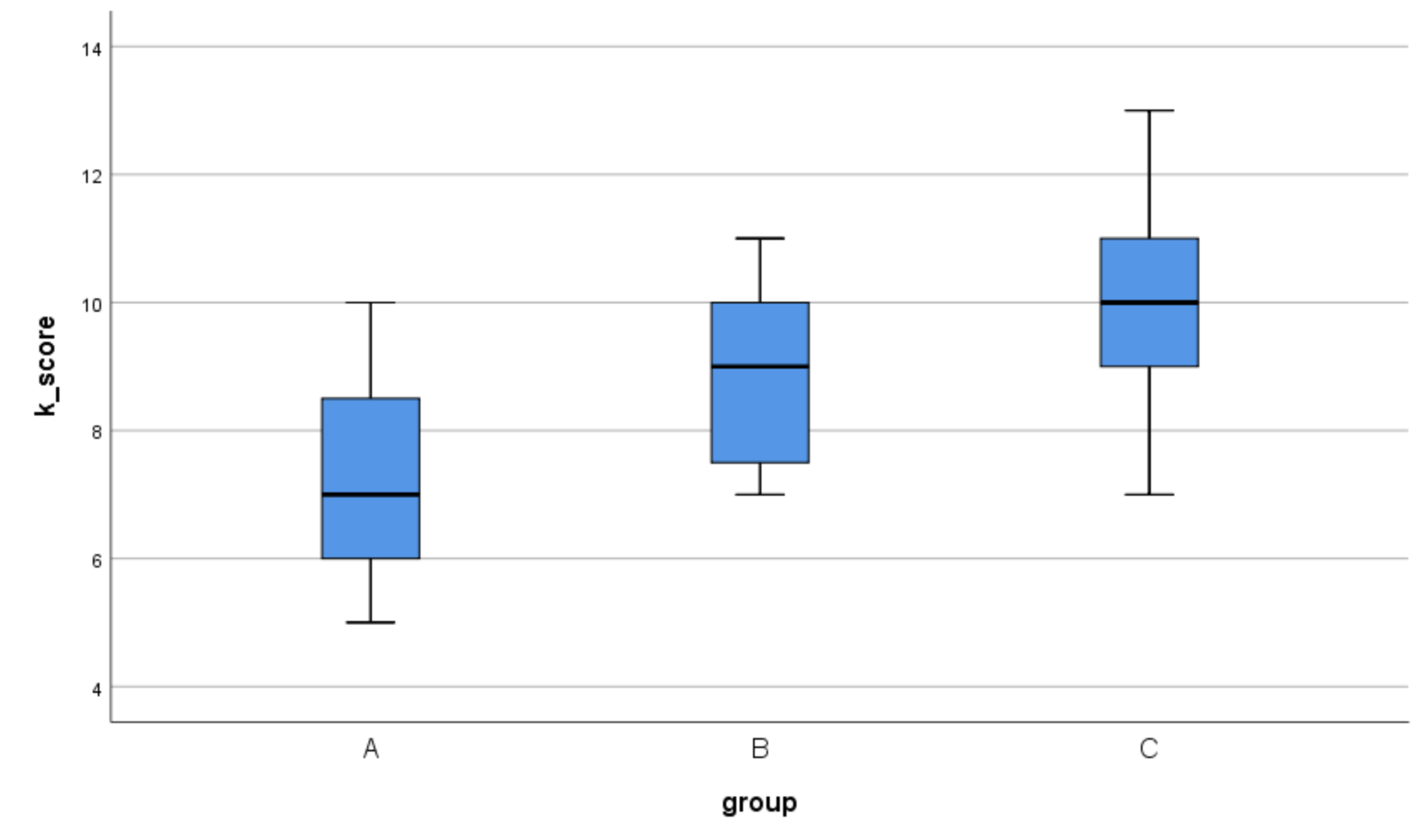
ANOVA determines P-value from the F statistic

Informal Investigation

Graphical investigation:

- side-by-side box plots
- multiple histograms

Side by Side Boxplots



Dr Kyaw Oo

Data [and means]:

- A: 5,6,6,7,7,8,9,10 [7.25]
- B: 7,7,8,9,9,10,10,11 [8.875]
- C: 7,9,9,10,10,10,11,12,13 [10.11]

Are these differences significant?

What does ANOVA do?

At its simplest ANOVA tests the following hypotheses:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

k_score	group
Xi	i
5	A
6	A
6	A
7	A
7	A
8	A
9	A
10	A
7	B
7	B
8	B
9	B
9	B
10	B
10	B
11	B
7	C
9	C
9	C
10	C
10	C
10	C
11	C
12	C
13	C

Assumptions of ANOVA

Dr Kyaw Oo

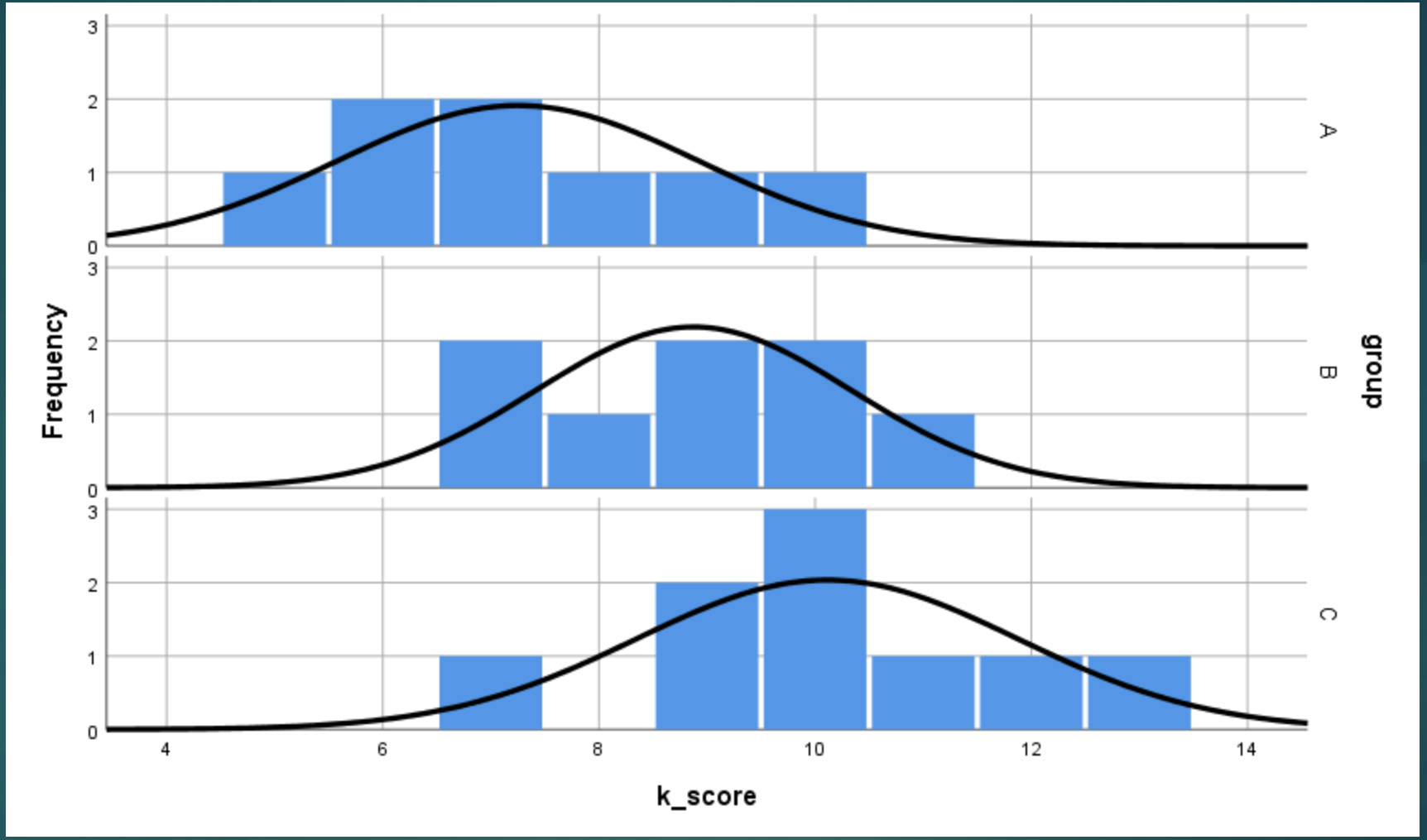
- each group is approximately normal
  - check this by looking at histograms and/or normal quantile plots, or use assumptions
  - can handle some non-normality, but not severe outliers

Normality Check

We should check for normality using:

- assumptions about population
- histograms for each group
- normal quantile plot for each group

With such small data sets, there really isn't a really good way to check normality from data, but we make the common assumption that physical measurements of people tend to be normally distributed.


Frequency
3
2
1
0
A
B
C
group
4
6
8
10
12
14
k_score

Assumptions of ANOVA

Dr Kyaw Oo

- standard deviations of each group are approximately equal
  - rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

Standard Deviation Check

Report

k_score

group	N	Mean	Std. Deviation
A	8	7.25	1.669
B	8	8.88	1.458
C	9	10.11	1.764
Total	25	8.80	1.979

Compare largest and smallest standard deviations:

- largest: 1.764
- smallest: 1.458
- 1.458 x 2 = 2.916 > 1.764

Note: variance ratio of 4:1 is equivalent.

Notation for ANOVA

Dr Kyaw Oo

- n = number of individuals all together
- I = number of groups
- $\overline{X}$ = mean for entire data set is

Group i has

- n_i = # of individuals in group i
- x_{ij} = value for individual j in group i
- $\overline{x}_i$ = mean for group i
- s_i = standard deviation for group i

How ANOVA works (outline)

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

$$\left(\overline{x}_i - \overline{x}\right)^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$\left(x_{ij} - \overline{x}_i\right)^2$$

The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

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

A large F is evidence *against* H_0, since it indicates that there is more difference between groups than within groups.

SPSS ANOVA Output

Descriptives

k_score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
A	8	7.25	1.669	.590	5.85	8.65	5	10
B	8	8.88	1.458	.515	7.66	10.09	7	11
C	9	10.11	1.764	.588	8.76	11.47	7	13
Total	25	8.80	1.979	.396	7.98	9.62	5	13

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

How are these computations made?

We want to measure the amount of variation due to BETWEEN group variation and WITHIN group variation

For each data value, we calculate its contribution to:

- BETWEEN group variation: $\left(\bar{x}_i - \bar{\bar{x}}\right)^2$
- WITHIN group variation: $\left(x_{ij} - \bar{x}_i\right)^2$

SPSS ANOVA Output

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

How do you calculate these numbers?

SPSS ANOVA Output

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

1 less than number of groups

1 less than number of individuals (just like other situations)

number of data values - number of groups (equals df for each group added together)

k_score	group	k_score_gr_mean	diff_within	squ_w_diff	k_score_mean_all	diff_groupmean	squ_g_diff	squ_all
Xi	j	Mean(Xij)	Xi-mXij	(Xi-mXij)^2	Mean Xi	Xi-mXij	(Xi-mXij)^2	(Xi-mXi)^2
5	A	7.25	-2.25	5.06	8.8	1.55	2.4	14.44
6	A	7.25	-1.25	1.56	8.8	1.55	2.4	7.84
6	A	7.25	-1.25	1.56	8.8	1.55	2.4	7.84
7	A	7.25	-0.25	0.06	8.8	1.55	2.4	3.24
7	A	7.25	-0.25	0.06	8.8	1.55	2.4	3.24
8	A	7.25	0.75	0.56	8.8	1.55	2.4	0.64
9	A	7.25	1.75	3.06	8.8	1.55	2.4	0.04
10	A	7.25	2.75	7.56	8.8	1.55	2.4	1.44
7	B	8.88	-1.88	3.52	8.8	-0.07	0.01	3.24
7	B	8.88	-1.88	3.52	8.8	-0.07	0.01	3.24
8	B	8.88	-0.88	0.77	8.8	-0.07	0.01	0.64
9	B	8.88	0.13	0.02	8.8	-0.07	0.01	0.04
9	B	8.88	0.13	0.02	8.8	-0.07	0.01	0.04
10	B	8.88	1.13	1.27	8.8	-0.07	0.01	1.44
10	B	8.88	1.13	1.27	8.8	-0.07	0.01	1.44
11	B	8.88	2.13	4.52	8.8	-0.07	0.01	4.84
7	C	10.11	-3.11	9.68	8.8	-1.31	1.72	3.24
9	C	10.11	-1.11	1.23	8.8	-1.31	1.72	0.04
9	C	10.11	-1.11	1.23	8.8	-1.31	1.72	0.04
10	C	10.11	-0.11	0.01	8.8	-1.31	1.72	1.44
10	C	10.11	-0.11	0.01	8.8	-1.31	1.72	1.44
10	C	10.11	-0.11	0.01	8.8	-1.31	1.72	1.44
11	C	10.11	0.89	0.79	8.8	-1.31	1.72	4.84
12	C	10.11	1.89	3.57	8.8	-1.31	1.72	10.24
13	C	10.11	2.89	8.35	8.8	-1.31	1.72	17.64
				59.27			34.76	94
				SSW			SSA	SST

SPSS ANOVA Output

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

1 less than number of groups

1 less than number of individuals (just like other situations)

number of data values - number of groups (equals df for each group added together)

SPSS ANOVA Output

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

$$\sum_{obs}(x_{ij}-\bar{x}_i)^2$$

$$\sum_{obs}(x_{ij}-\bar{\bar{x}})^2$$

$$\sum_{obs}(\bar{x}_i-\bar{\bar{x}})^2$$

SS stands for sum of squares

- ANOVA splits this into 3 parts

ANOVA Output

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in Table E)

Dr Kyaw Oo

SPSS ANOVA Output

ANOVA

k_score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	34.736	2	17.368	6.447	.006
Within Groups	59.264	22	2.694		
Total	94.000	24			

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in Table E)

So How big is F?

Since F is

Mean Square Between / Mean Square Within

= MSG / MSE

A large value of F indicates relatively more difference between groups than within groups (evidence against H_0)

To get the P-value, we compare to *F(I-1,n-I)*-distribution

- *I-1* degrees of freedom in numerator (# groups -1)
- *n - I* degrees of freedom in denominator (rest of df)

Connections between SST, MST, and standard deviation

If ignore the groups for a moment and just compute the standard deviation of the entire data set, we see

$$s^2 = \frac{\sum \left(x_{ij} - \bar{\bar{x}} \right)^2}{n-1} = \frac{SST}{DFT} = MST$$

So $SST = (n-1)\, s^2$, and $MST = s^2$. That is, *SST* and *MST* measure the TOTAL variation in the data set.

Connections between SSE, MSE, and standard deviation

Remember: $s_i^2 = \frac{\sum (x_{ij} - \bar{x}_i)^2}{n_i - 1} = \frac{SS[\text{Within Group } i]}{df_i}$

So $SS[\text{Within Group } i] = (s_i^2)\,(df_i)$

This means that we can compute SSE from the standard deviations and sizes (df) of each group:

$$SSE = SS[\text{Within}] = \sum SS[\text{Within Group } i]$$

$$= \sum s_i^2 (n_i - 1) = \sum s_i^2 (df_i)$$

Pooled estimate for st. dev

One of the ANOVA assumptions is that all groups have the same standard deviation. We can estimate this with a weighted average:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_I - 1)s_I^2}{n - I}$$

$$s_p^2 = \frac{(df_1)s_1^2 + (df_2)s_2^2 + \ldots + (df_I)s_I^2}{df_1 + df_2 + \ldots + df_I}$$

$$s_p^2 = \frac{SSE}{DFE} = MSE$$

so MSE is the pooled estimate of variance

Dr Kyaw Oo

In Summary

$$SST = \sum_{obs} (x_{ij} - \bar{\bar{x}})^2 = s^2(DFT)$$

$$SSE = \sum_{obs} (x_{ij} - \bar{x}_i)^2 = \sum_{groups} s_i^{\,2}(df_i)$$

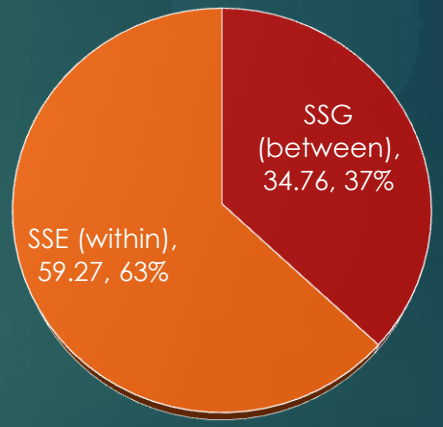
$$SSG = \sum_{obs} (\bar{x}_i - \bar{\bar{x}})^2 = \sum_{groups} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

$$SSE + SSG = SST; \quad MS = \frac{SS}{DF}; \quad F = \frac{MSG}{MSE}$$

R^2 Statistic

R^2 gives the percent of variance due to between group variation

$$R^2 = \frac{SS[Between]}{SS[Total]} = \frac{SSG}{SST}$$



We will see R^2 again when we study regression.

Where's the Difference?

Dr Kyaw Oo

Once ANOVA indicates that the groups do not all appear to have the same means, what do we do?

```
Analysis of Variance for days
Source      DF         SS         MS         F         P
treatmen     2      34.74      17.37      6.45     0.006
Error       22      59.26       2.69
Total       24      94.00

                                          Individual 95% CIs For Mean
                                          Based on Pooled StDev
Level        N       Mean      StDev  ----------+---------+---------+------
A            8      7.250      1.669  (-------*-------)
B            8      8.875      1.458            (-------*-------)
P            9     10.111      1.764                     (------*-------)
                                          ----------+---------+---------+------
Pooled StDev =     1.641                        7.5       9.0      10.5
```

Clearest difference: P is worse than A (CI's don't overlap)

Multiple Comparisons

Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two using the 2-sample t test

- We need to adjust our p-value threshold because we are doing multiple tests with the same data.

- There are several methods for doing this.

- If we really just want to test the difference between one pair of treatments, we should set the study up that way.

Pairwise Comparisons

Multiple Comparisons

Dependent Variable:

LSD

(I) Employment Category		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Clerical	Custodial	-9.283	7.932	0.242	-24.87	6.30
	Manager	-20.614*	4.814	0.000	-30.07	-11.15
Custodial	Clerical	9.283	7.932	0.242	-6.30	24.87
	Manager	-11.331	8.797	0.198	-28.62	5.95
Manager	Clerical	20.614*	4.814	0.000	11.15	30.07
	Custodial	11.331	8.797	0.198	-5.95	28.62

*. The mean difference is significant at the 0.05 level.

- Alpha $_{(FE)}$=Alpha*K
- K=number of comparison
- Alpha = Usually 0.05

Assessment1

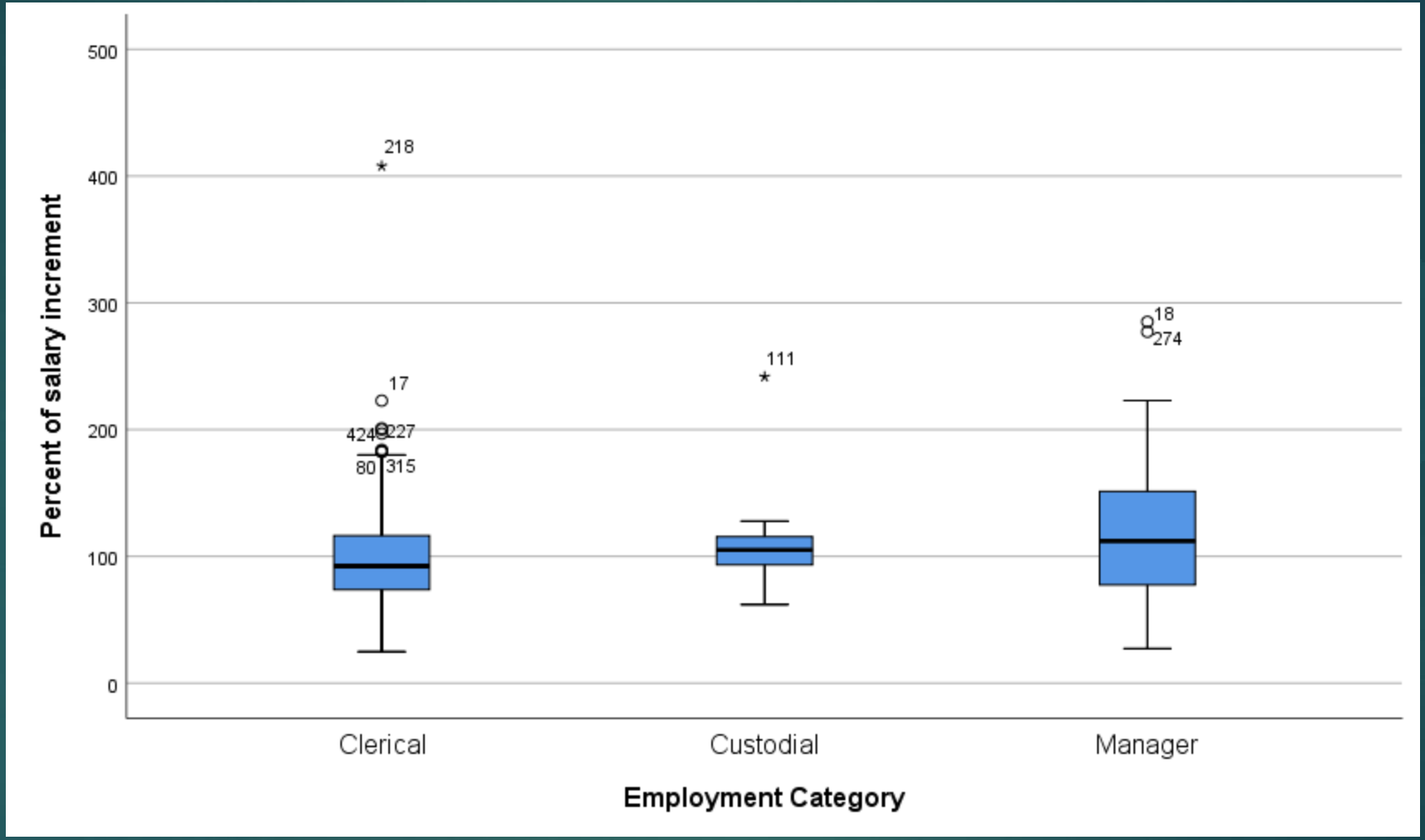
In the one-way analysis of variance table for the comparison of three groups:

- (a) the group mean square + the error mean square = the total mean square;
- (b) there are two degrees of freedom for groups;
- (c) the group sum of squares + the error sum of squares = the total sum of squares;
- (d) the numbers in each group must be equal;
- (e) the group degrees of freedom + the error degrees of freedom = the total degrees of freedom.

FTTFT

Assessment2
How will you interpret?

- Does the ANOVA analysis meet assumptions?
- Are the group summaries different? If yes, how much?
- Is the difference statistically significant?
  - at which alpha value?
  - F value and R^2 and P value?
- If yes, which groups are different? How much?


Percent of salary increment
500
400
300
200
100
0
218
17
424
227
80
315
111
18
274
Clerical
Custodial
Manager
Employment Category


Frequency
80
60
40
20
0
Clerical
Custodial
Manager
Employment Category
0
100
200
300
400
500
Percent of salary increment

Descriptives

Percent of salary increment

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Clerical	363	98.21	37.184	1.952	94.37	102.05	25	408
Custodial	27	107.49	30.680	5.904	95.36	119.63	62	242
Manager	84	118.82	51.450	5.614	107.66	129.99	27	285
Total	474	102.39	40.463	1.859	98.74	106.04	25	408

ANOVA

Percent of salary increment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	29731.769	2	14865.885	9.402	.000
Within Groups	744704.592	471	1581.114		
Total	774436.362	473			

Multiple Comparisons

Dependent Variable: Percent of salary increment

Bonferroni

(I) Employment Category	(J) Employment Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Clerical	Custodial	-9.283	7.932	.727	-28.34	9.77
	Manager	-20.614*	4.814	.000	-32.18	-9.05
Custodial	Clerical	9.283	7.932	.727	-9.77	28.34
	Manager	-11.331	8.797	.595	-32.47	9.80
Manager	Clerical	20.614*	4.814	.000	9.05	32.18
	Custodial	11.331	8.797	.595	-9.80	32.47

*. The mean difference is significant at the 0.05 level.

Multiple Comparisons

Dependent Variable: Percent of salary increment

	(I) Employment Category	(J) Employment Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	Clerical	Custodial	-9.283	7.932	.242	-24.87	6.30
		Manager	-20.614*	4.814	.000	-30.07	-11.15
	Custodial	Clerical	9.283	7.932	.242	-6.30	24.87
		Manager	-11.331	8.797	.198	-28.62	5.95
	Manager	Clerical	20.614*	4.814	.000	11.15	30.07
		Custodial	11.331	8.797	.198	-5.95	28.62
Bonferroni	Clerical	Custodial	-9.283	7.932	.727	-28.34	9.77
		Manager	-20.614*	4.814	.000	-32.18	-9.05
	Custodial	Clerical	9.283	7.932	.727	-9.77	28.34
		Manager	-11.331	8.797	.595	-32.47	9.80
	Manager	Clerical	20.614*	4.814	.000	9.05	32.18
		Custodial	11.331	8.797	.595	-9.80	32.47

*. The mean difference is significant at the 0.05 level.

Thank you.