General Linear Models and Analysis of Variance

Prof. Dr. Akın Pala

Replication

- Treatment is applied to more than one experimental unit.
- Example:
  - 3 pots get 8 hours of sunlight
  - 3 pots get 12 hours of sunlight
  - 3 pots get 16 hours of sunlight.
- Must have replication, you have to apply the same thing to more than one experimental unit/subject.

Replication

- Reasons for using replication:
- Estimate experimental error
- Improve precision of estimate
  - Larger samples tend to have smaller variance in life sciences (variance=differences/n)
- Increase scope of inference
  - Use different strains of mint
  - Different kinds of soil
  - Use different kinds of animals, as opposed to drawing from a restricted area of the population.

Analysis of Variance (ANOVA)

ANOVA was developed in the 1920s by Ronald A. Fisher (1890-1962) who worked for the British Government Agricultural Department.

Analysis of Variance (ANOVA)

- The Analysis of Variance (ANOVA) is a technique commonly used to test whether there are statistically significant differences between two or more independent groups.
- These groupings are made on the basis of levels of independent variables.
- The simplest form of ANOVA is a one way analysis, where the intention is to determine whether there is an overall main effect of different levels of an independent variable on a dependent variable.
- The outcome of performing an ANOVA is an F ratio, which can be used to determine whether statistically significant differences exist between the groups.

ANOVA example

- A simple experiment to determine the effect of caffeine on learning,
- You might have two levels of the independent variables
  - experimental: caffeine 100mg;
  - control: placebo.
- The F ratio can then be tested for significance at different levels of α, typically $P < 0.01$ or $P < 0.05$, for a dependent variable, such as number of correct responses in a math quiz.

Extensions of ANOVA

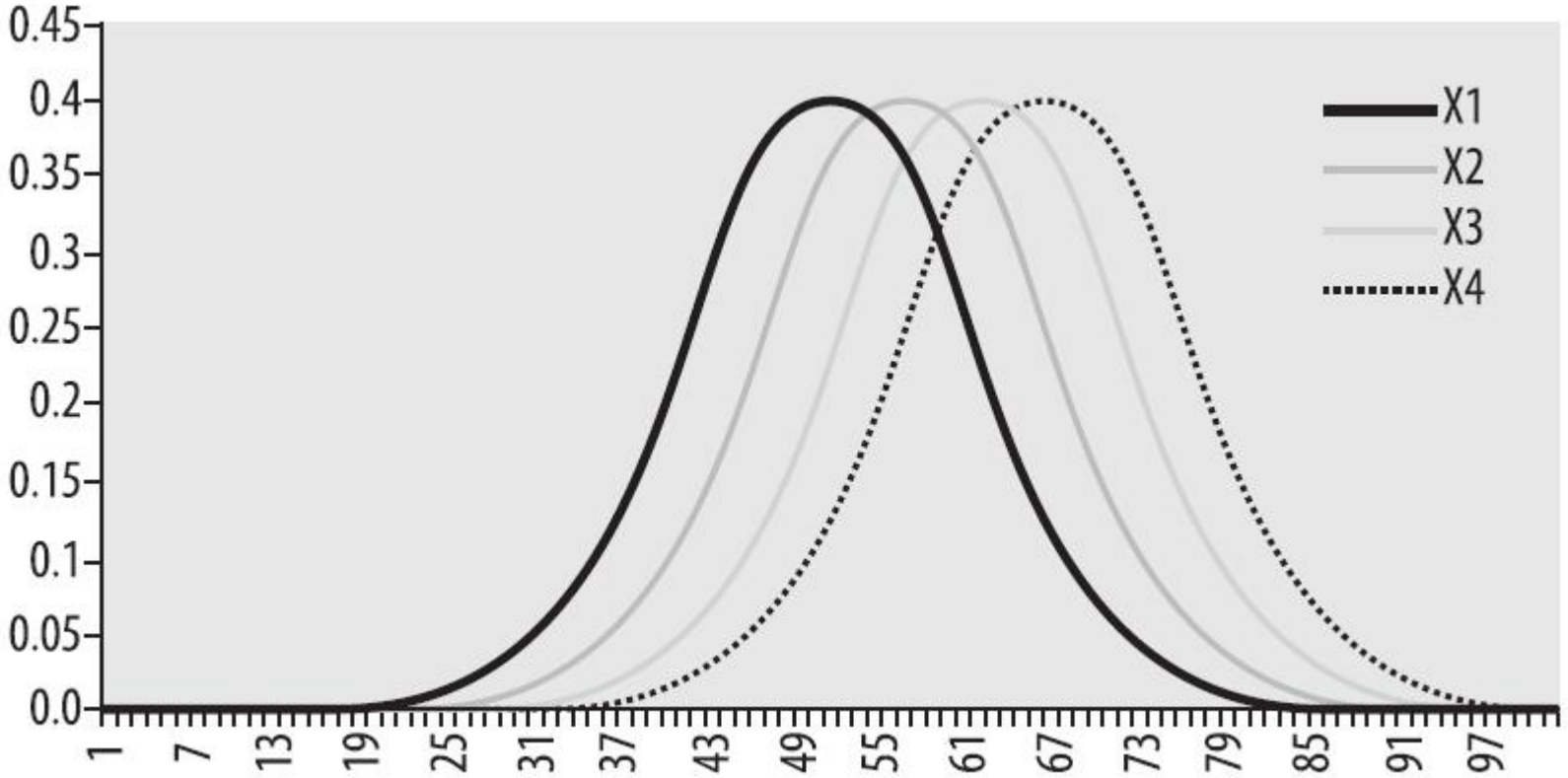
- Like linear regression, ANOVA can be extended to account for differences across multiple independent variables, as well as interactions between the variables.
- For example, you may wish to know
  - Do caffeine or nicotine have an effect on learning?
    - Main effects,
  - Is there any interaction between caffeine and nicotine?

One-Way ANOVA

- The simplest form of ANOVA is a one-way ANOVA, where the simple question of differences between two or more treatment levels can be tested in terms of a main effect.
- This is just like performing a t-test.
- The hypothesis is that one or more populations differ significantly in their means.

4 groups

- Slightly different means, significant overlap in their distributions; a large amount of variation within each sample. Probably non significant.

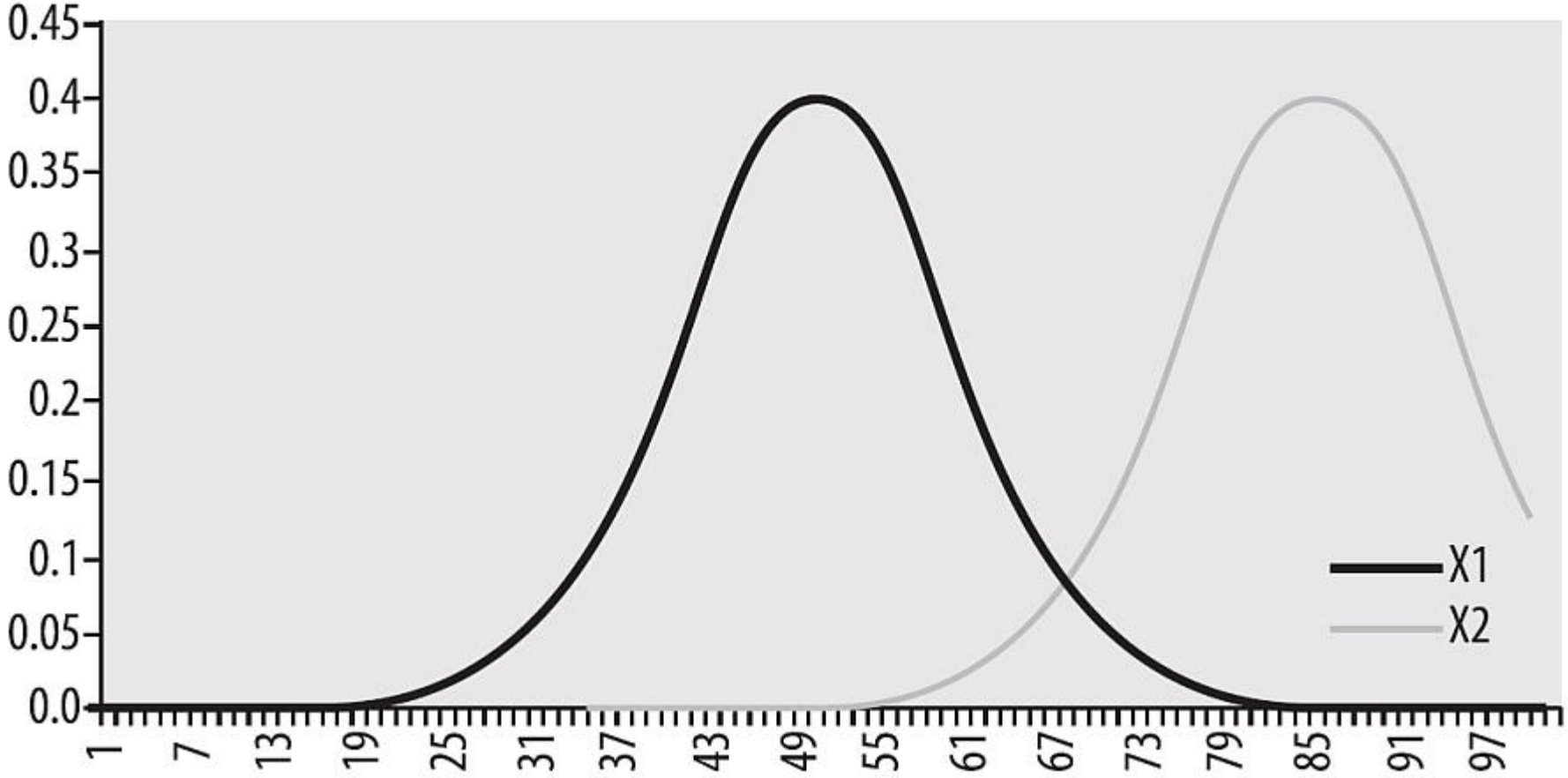


4 groups, overlaps

- $H_0 = X1 = X2 = X3 = X4$
- Fail to reject the null hypothesis.
- No difference:
  - the samples were drawn from the same population
  - the division into groups is not appropriate

2 groups, little overlapping

- Differences are large enough to be significant



2 groups, little overlapping

- The means are statistically significant
- The two samples have been drawn from different populations.

ANOVA

- The hypothesized variation between means is the basis for the analysis of variance.
- The total variation across all samples is known as the Total Sum of Squares (SS_{Total})
  - The total sum of squared deviations
- The variation between samples is known as the Sum of Squares Between ($SS_{Between}$)
- The variation within samples is known as the Sum of Squares Within (SS_{Within}).

ANOVA

- In simple terms, and taking into account degrees of freedom for each variation term:
- The F test for ANOVA is simply the ratio between the Sum of Squares Between and the Sum of Squares Within.
- If the variation between samples becomes relatively large compared to the variation within samples
  - You would expect the groups to be drawn from different populations, and a significant mean difference to exist.

Three Classes of Models

1. Fixed effects models
2. Random effects models
3. Mixed effects models

1. Fixed effects model

- The fixed effects model applies to situations where the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change.
- This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

2. Random effects model

- Random effects models are used when the treatments are not fixed.
- This occurs when the various factor levels are sampled from a larger population.
- Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.

3. Mixed effects model

- A mixed effects model contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

Three Classes of Models, Example

- Factors affecting gender of progeny:
  - Breed
  - Dam age
  - Sire id

Calculations in ANOVA

- The calculations of ANOVA can be characterized as computing a number of means and variances, dividing two variances and comparing the ratio to a handbook value to determine statistical significance.
- Calculating a treatment effect is then trivial, "the effect of any treatment is estimated by taking the difference between the mean of the observations which receive the treatment and the general mean.

Calculations in ANOVA

- Total sum of squares = explained/group sum of squares (SS for treatment) + residual sum of squares (error).
- SSE=TotalSS-GroupSS
- The total sum of squares = between samples sum of squares + within samples sum of squares
- In MANOVA:
- Total sum of squares and products (SSP) matrix= Between samples SSP matrix + Within samples SSP matrix.

t = treatment r = # of replicates within a treatment
i = treatments j = replicates within a treatment

Source of variation	df	Sums of squares: Definition	Sums of squares: Working	Mean squares†	F
Treatments	$t-1$	$r\sum_{i}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i}\frac{Y_{i\cdot}^2}{r}-\frac{Y_{\cdot\cdot}^2}{rt}$	√	√
Error	$t(r-1)$	$\sum_{i,j}(Y_{ij}-\bar{Y}_{i\cdot})^2$	by subtraction	√	
Total	$rt-1$	$\sum_{i,j}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i,j}Y_{ij}^2-\frac{Y_{\cdot\cdot}^2}{rt}$		

Calculations in ANOVA

$$Y = \beta X$$

$$X'Y = \beta X'X$$

ANOVA, example

- Polychlorinated biphenyls were widely used as dielectric and coolant fluids in electrical apparatus, cutting fluids for machining operations, carbonless copy paper and in heat transfer fluids.
- Due to PCBs' environmental toxicity and classification as a persistent organic pollutant, PCB production was banned by the United States Congress in 1979.
- PCBs cause cancer in animals and are probable human carcinogens.

Polychlorinated biphenyls, PCB

- Study effects of different levels of PCB on body weights of mice (in grams).
- $H_0 = \mu_1 = \mu_2 = \ldots\ldots = \mu_t$
- H_1 = not H_0 (some are different)

PCB

- Study effects of different levels of PCB on body weights of mice (in grams).

0ppm	62.5ppm	250ppm	1000ppm
55	47	49	36
47	51	44	41
46	40	46	
53	44	51	
		48	
$n_1=4$	$n_2=4$	$n_3=5$	$n_4=2$
$Y_{1.}=201$	$Y_{2.}=182$	$Y_{3.}=238$	$Y_{4.}=77$
$\overline{Y_1}=50.25$	$\overline{Y_2}=45.50$	$\overline{Y_3}=47.60$	$\overline{Y_4}=38.50$

$Y_{..}=201+182+238+77=698$ $n=4+4+5+2=15$

Source of variation	df	Sums of squares		Mean squares†	F
		Definition	Working		
Treatments	$t-1$	$r\sum_{i}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i}\frac{Y_{i\cdot}^2}{r}-\frac{Y_{\cdot\cdot}^2}{rt}$	√	√
Error	$t(r-1)$	$\sum_{i,j}(Y_{ij}-\bar{Y}_{i\cdot})^2$	by subtraction	√	
Total	$rt-1$	$\sum_{i,j}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i,j}Y_{ij}^2-\frac{Y_{\cdot\cdot}^2}{rt}$		

Calculations, Total Sum of Squares

$$TotalSS = \sum_{i,j} (Y_{ij} - \overline{Y..})^2$$

$$TotalSS = \sum Y_{i,j}{}^2 - \frac{(\sum Y)^2}{r * t}$$

$$\text{TotalSS} = 55^2 + 47^2 + \ldots.. + 41^2 - \frac{698^2}{15} = 359.73$$

Calculations, Treatment Sum of Squares

$$TreatmentSS = \sum \frac{Y_{i.}^2}{r} - \frac{(\sum Y)^2}{r * t}$$

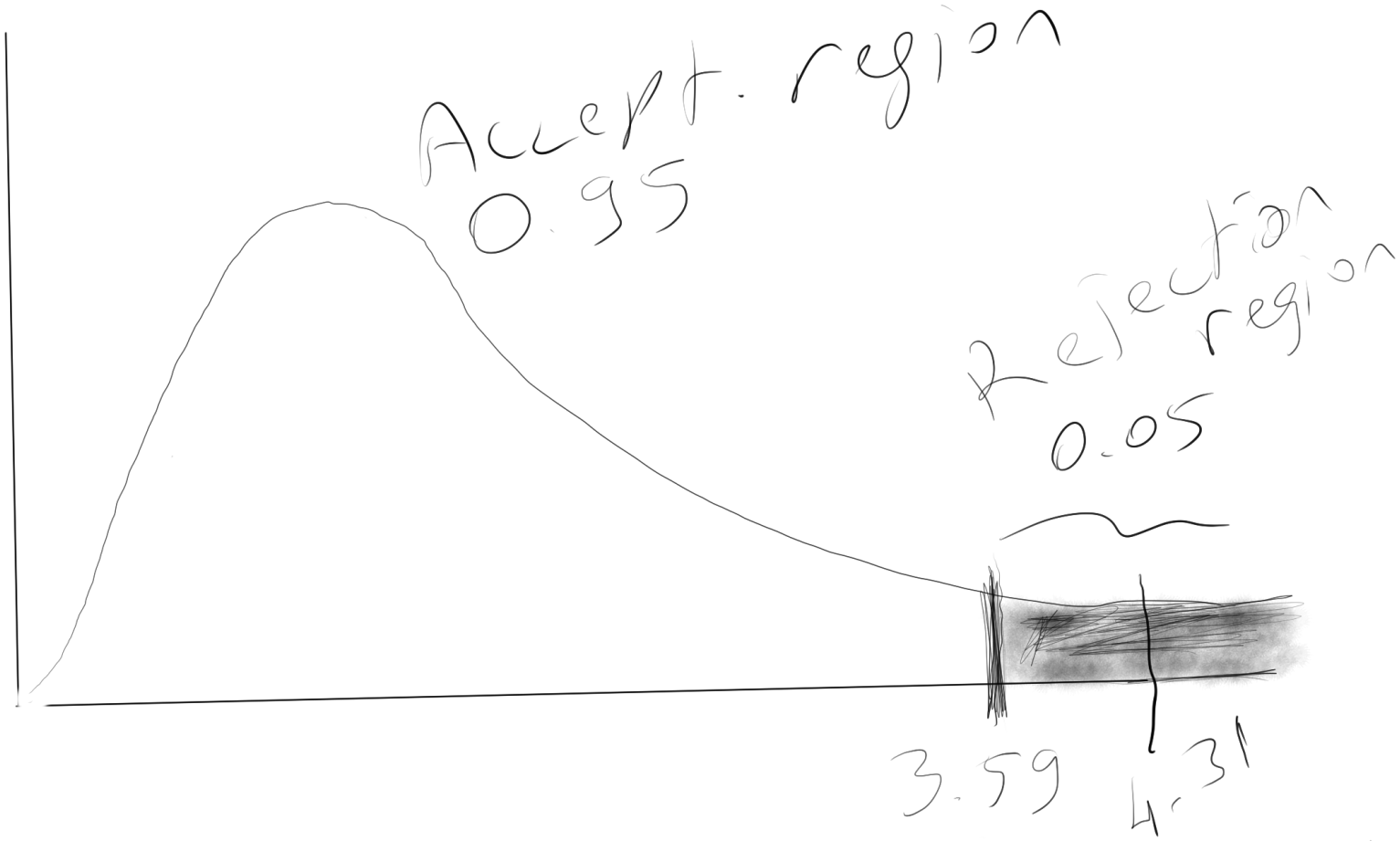
$$\text{TreatmentSS} = \frac{201^2}{4} + \frac{182^2}{4} + \frac{238^2}{5} + \frac{77^2}{2} - \frac{698^2}{15} = 194.28$$

Calculations, Error Sum of Squares

- TotalSS=SSTreatment+SSE
- SSE=TotalSS-SSTreatment
- SSE=359.73-194.28=165.45

Source	df	SS	MS	F	P
Total SS	14	359.74			
SSTrt	3	194.28	64.76	4.31	0.03
SSE	11	165.45	15.04		

$F_{0.05,\ 3,\ 11}$ from the table=3.59


Accept. region
0.95
Rejection region
0.05
3.59
4.31

Discussion

- Reject H_0 at level α if $F > F_{\alpha, t-e, n-t}$
- If the treatment mean square is less than the error, the result is declared nonsignificant not matter how small the ratio.
- A significant F implies that the evidence is sufficiently strong to indicate that all the treatments do not belong to populations with a common μ

Discussion

- $F(4.31) > F_{0.05,\ 3,\ 11}$ (3.59)
- Reject H_0 and Conclude:
- Level of PCB affects body weight.
- Not all the means of the groups are equal.

ANOVA example with SAS

```
data ANOVA;
input dose weight;
cards;
0 55
0 47
0 46
0 53
62.5 47
62.5 51
62.5 40
62.5 44
250 49
250 44
250 46
250 51
250 48
1000 36
1000 41
;
run;
proc glm; class dose;
model weight=dose;
run;
quit;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
dose	4	0 62.5 250 1000

Number of Observations Read	15
Number of Observations Used	15

The GLM Procedure
Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	194.2833333	64.7611111	4.31	0.0307
Error	11	165.4500000	15.0409091		
Corrected Total	14	359.7333333			

R-Square	Coeff Var	Root MSE	weight Mean
0.540076	8.334372	3.878261	46.53333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
dose	3	194.2833333	64.7611111	4.31	0.0307

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dose	3	194.2833333	64.7611111	4.31	0.0307

CV

- Coefficient of Variation=ratio of the standard deviation to the mean.
- The CV is widely used in analytical chemistry to express the precision and repeatability of an assay.
- CV is useful because the standard deviation of data must always be understood in the context of the mean of the data.
  - In contrast, the actual value of the CV is a standard number.
  - For comparison between data sets with different units or widely different means, one should use CV instead of the standard deviation.

CV

- Coefficient of Variation=ratio of the standard deviation to the mean.
- The CV is widely used in analytical chemistry to express the precision and repeatability of an assay.
- A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100:
- $100\% \times 0 / 100 = 0\%$
- A data set of [90, 100, 110] has more variability. Its standard deviation is 8.16 and its average is 100:
- $100\% \times 8.16 / 100 = 8.16\%$
- A data set of [1, 5, 6, 8, 10, 40, 65, 88] has more variability again. Its standard deviation is 30.78 and its average is 27.875:
- $100\% \times 30.78 / 27.875 = 110.4\%$

The coefficient of determination

- The **coefficient of determination**, denoted R^2 and pronounced **R squared**, is a number that indicates how well data fit a statistical model.
- An R^2 of 1 indicates that the regression line perfectly fits the data, while an R^2 of 0 indicates that the line does not fit the data at all.

The coefficient of determination

- It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model
- In our example, R^2 was 0.54, so it means that we are missing more factors affecting the body weight in mice.

$$R^2 = 1 - \frac{SSE}{SSTotal} = \frac{165.45}{359.73} = 1 - 0.46 = 0.54$$

Post Hoc test

- In the one-way ANOVA example, the F-test was statistically significant, and the null hypothesis was rejected.
- However, you still need a way to determine which of the means were significantly different from each other
- The result of the F-test does not imply that all samples were drawn from different populations, merely that at least one was drawn from a different population from the others.
- At the alpha value= 0.05, if 20 comparisons are made, you would expect at least one erroneous result.
- Probabilities of mistakes add up.

Post Hoc tests

- To determine which groups were actually different, you need to undertake post hoc tests.
- The Scheffe test is the most conservative test.
- Fisher's least significant difference
- Bonferroni correction
- Duncan's new multiple-range test
- Student-Newman-Keuls' test
- Tukey's honestly significant difference test
  - When the data is unbalanced, Tukey/Kramer test

In papers

- Difference tests were adjusted by the Tukey method (TUKEY, 1953). Since the data was unbalanced, adjustments were approximated using a method defined by KRAMER (1956).
  - TUKEY, J. W. The problem of multiple comparisons. (1953). Mimeographed Monograph, appears in full in the "Collected work of J. W. Tukey" Vol. VII, (edited by H. Braun) 1994, Chapman & Hall Inc., N.Y. Princeton University, Princeton, NJ.
  - KRAMER, C. Y. Extension of multiple range tests to group means with unequal numbers of replications. Biometry, 12 (1956), 309-310

```
data ANOVA;
input dose weight;
cards;
0 55
0 47
0 46
0 53
62.5 47
62.5 51
62.5 40
62.5 44
250 49
250 44
250 46
250 51
250 48
1000 36
1000 41
;
run;
proc glm; class dose;
model weight=dose;
lsmeans dose/pdiff stderr;
lsmeans dose/pdiff stderr adjust=tukey;
run;
quit;
```

LSMeans Comparisons

The GLM Procedure
Least Squares Means

dose	weight LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
0	50.2500000	1.9391305	<.0001	1
62.5	45.5000000	1.9391305	<.0001	2
250	47.6000000	1.7344111	<.0001	3
1000	38.5000000	2.7423447	<.0001	4

Least Squares Means for effect dose
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: weight

i/j	1	2	3	4
1		0.1112	0.3303	0.0050
2	0.1112		0.4367	0.0613
3	0.3303	0.4367		0.0171
4	0.0050	0.0613	0.0171	

Tukey/Kramer Adjustment

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

dose	weight LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
0	50.2500000	1.9391305	<.0001	1
62.5	45.5000000	1.9391305	<.0001	2
250	47.6000000	1.7344111	<.0001	3
1000	38.5000000	2.7423447	<.0001	4

Least Squares Means for effect dose
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: weight

i/j	1	2	3	4
1		0.3537	0.7424	0.0221
2	0.3537		0.8498	0.2173
3	0.7424	0.8498		0.0702
4	0.0221	0.2173	0.0702	

Homework

- Form a data set with at least 15 observations and 3 levels of a factor.
- Calculate F manually first, show your calculations including SST etc.
- Use SAS for the same and include Tukey.
- Make comments on your results.