Introduction to data analysis

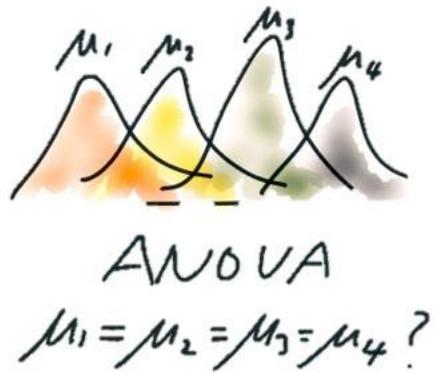
Part 6. Introduction to statistics (part 2/2)

ANOVA

Hypothesis test for more than one population mean
ANOVA

ANOVA = Analysis Of Variance

Hypothesis test for more than one population mean



Null hypothesis

H_0

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_g$

Alternative hypothesis

H_a

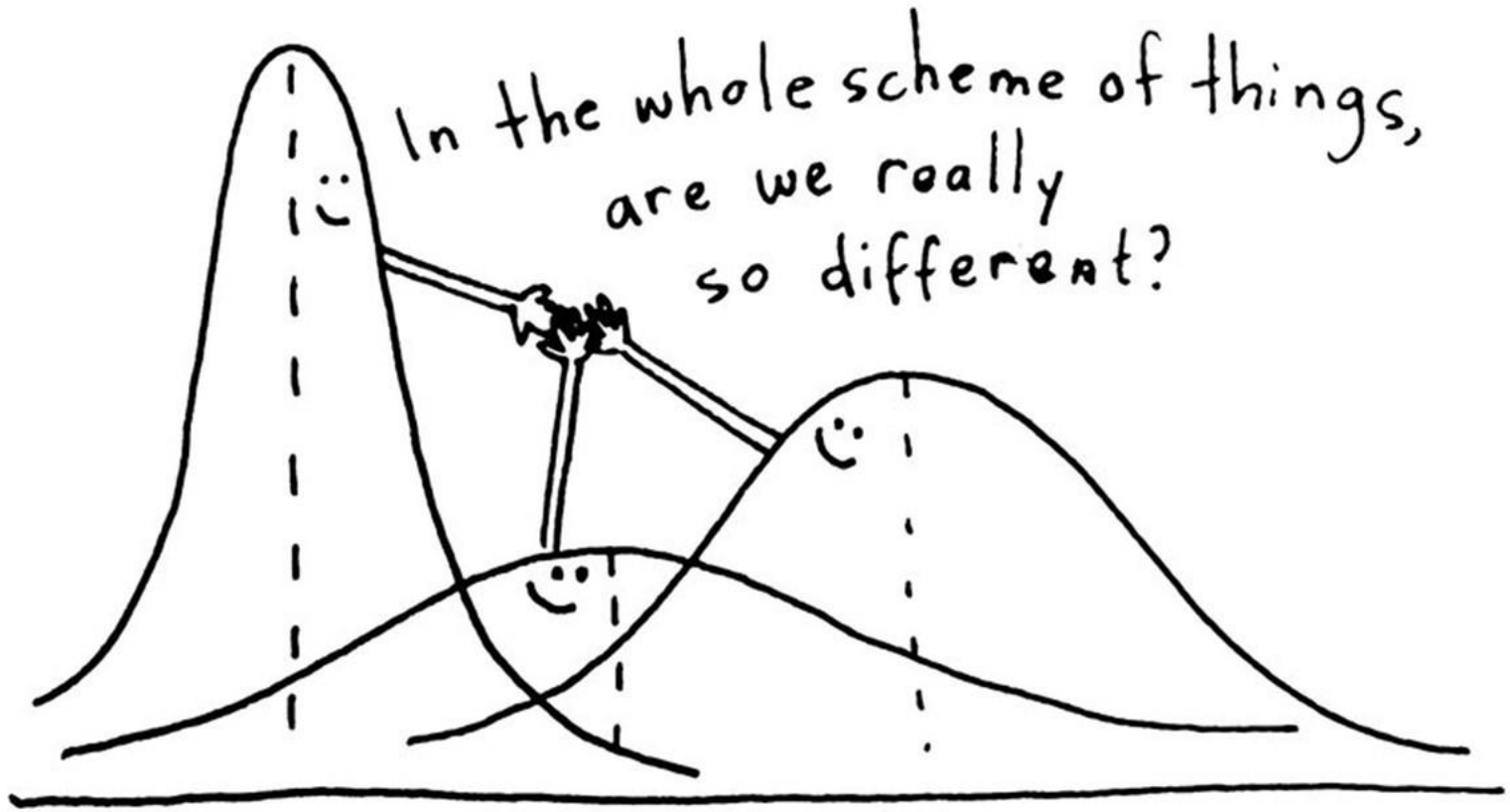
H_a: At least 1 mean differs from the others

Assumptions:

All the populations (g) have a normal distribution

All the populations have the same variance: $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \ldots = \sigma^2_g$

Hypothesis test for more than one population mean
ANOVA



Hypothesis test for more than one population mean
ANOVA vs T-test

BASIS FOR COMPARISON	T-TEST	ANOVA
Meaning	T-test is a hypothesis test that is used to compare the means of two populations.	ANOVA is a statistical technique that is used to compare the means of more than two populations.
Test statistic	$(\bar{x}-\mu)/(s/\sqrt{n})$	Between Sample Variance/Within Sample Variance

Hypothesis test for more than one population mean
ANOVA

Number of samples = g
With g > 1

Sample 1	Sample 2	Sample 3	Sample g
x_{11}	x_{21}	x_{31}	x_{g1}
x_{12}	x_{22}	x_{32}	x_{g2}
...	...	...	...
$x_{1;n1-1}$	$x_{2;n2}$	$x_{3;n3-2}$	$x_{g;ng}$
$x_{1;\,n1}$		$x_{3;n3-1}$	
		$x_{3;n3}$	

Each sample can have a different number of observations

Hypothesis test for more than one population mean
ANOVA

Sample 1	Sample 2	Sample 3	Sample g
x_{11}	x_{21}	x_{31}	x_{g1}
x_{12}	x_{22}	x_{32}	x_{g2}
...	...	...	...
$x_{1;n1-1}$	$x_{2;n2}$	$x_{3;n3-2}$	$x_{g;ng}$
$x_{1;n1}$		$x_{3;n3-1}$	
		$x_{3;n3}$	

For each sample, one can calculate the mean

$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_g$

Hypothesis test for more than one population mean
ANOVA

Sample 1	Sample 2	Sample 3	Sample g
x_{11}	x_{21}	x_{31}	x_{g1}
x_{12}	x_{22}	x_{32}	x_{g2}
...	...	...	...
$x_{1;n1-1}$	$x_{2;n2}$	$x_{3;n3-2}$	$x_{g;ng}$
$x_{1;\,n1}$		$x_{3;n3-1}$	
		$x_{3;n3}$	

For all the samples, one can calculate the mean as

$$\bar{x} = \frac{1}{n_1+n_2+..+ng}\,(x_{11}+x_{12}+x_{13}+\ ...+x_{1;n1}+x_{21}+...+x_{2,\,n2}+x_{31}+...\,x_{3;n3}+...+x_{g1}+...+x_{g,\,ng})$$

NOT

$$\bar{x} = \frac{1}{g}\,(\bar{x}_1+\bar{x}_2+\bar{x}_3+\ ...+\bar{x}_g)$$

Hypothesis test for more than one population mean
ANOVA

Sample 1	Sample 2	Sample 3	Sample g
x_{11}	x_{21}	x_{31}	x_{g1}
x_{12}	x_{22}	x_{32}	x_{g2}
...	...	...	...
$x_{1;n1-1}$	$x_{2;n2}$	$x_{3;n3-2}$	$x_{g;ng}$
$x_{1;n1}$		$x_{3;n3-1}$	
		$x_{3;n3}$	

$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_g$	$\overline{x}$

Goal: Are the differences between the population means significant?

Hypothesis test for more than one population mean
ANOVA

Sample 1	Sample 2	Sample 3	Sample g
x_{11}	x_{21}	x_{31}	x_{g1}
x_{12}	x_{22}	x_{32}	x_{g2}
...	...	...	...
$x_{1;n1-1}$	$x_{2;n2}$	$x_{3;n3-2}$	$x_{g;ng}$
$x_{1;\,n1}$		$x_{3;n3-1}$	
		$x_{3;n3}$	

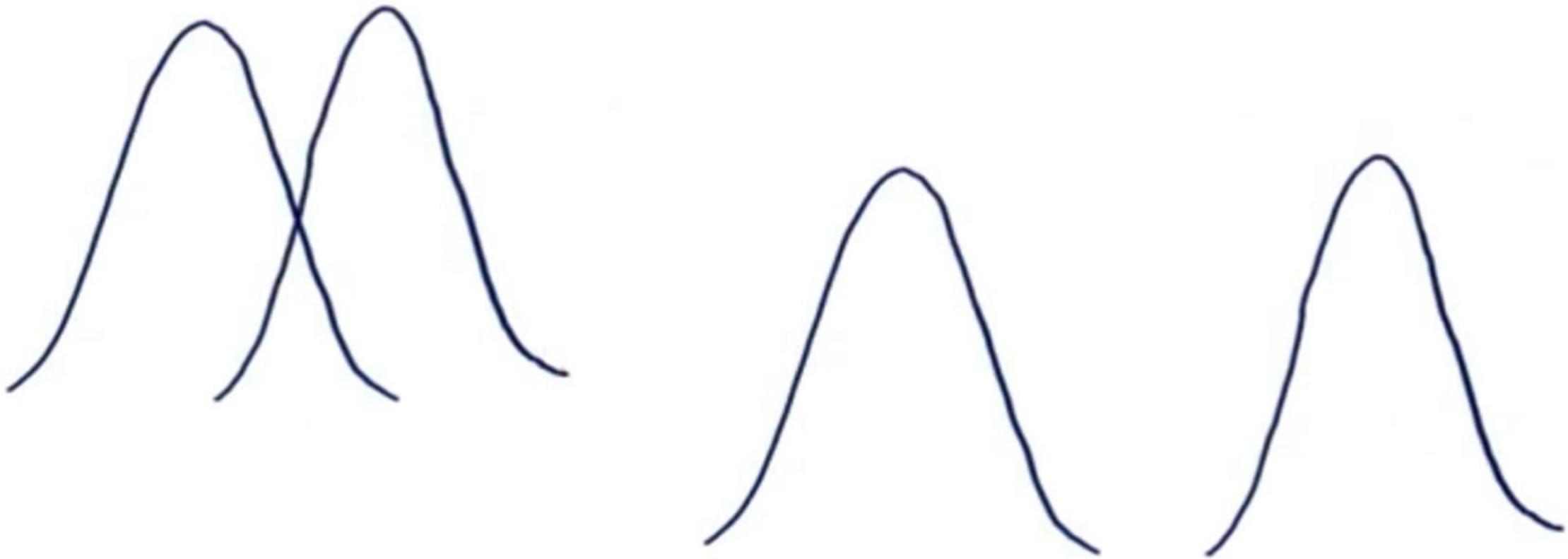
$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_g$	$\overline{x}$

Goal: Are the differences between the population means significant?

1. The differences between the population means

2. The variation of the data within one sample

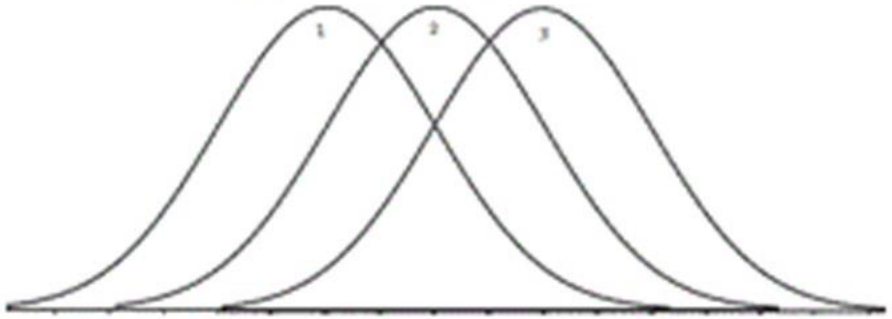
Hypothesis test for more than one population mean
Between group variability



Hypothesis test for more than one population mean
Between group variability

Little discrimination

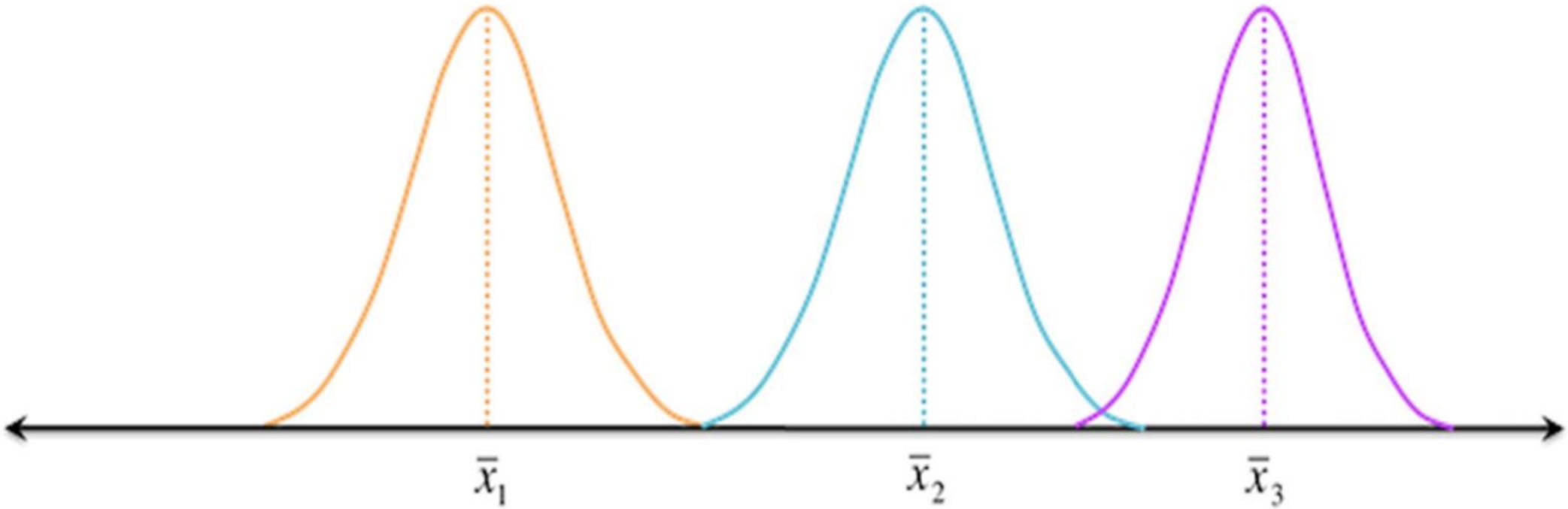
Some Discrimination



Discrimination between Two Groups, but not the third

Large Discrimination

Hypothesis test for more than one population mean
Within group variability



Hypothesis test for more than one population mean
ANOVA

For each element x_{ij}, one can define a total deviation from the mean $\bar{x}$

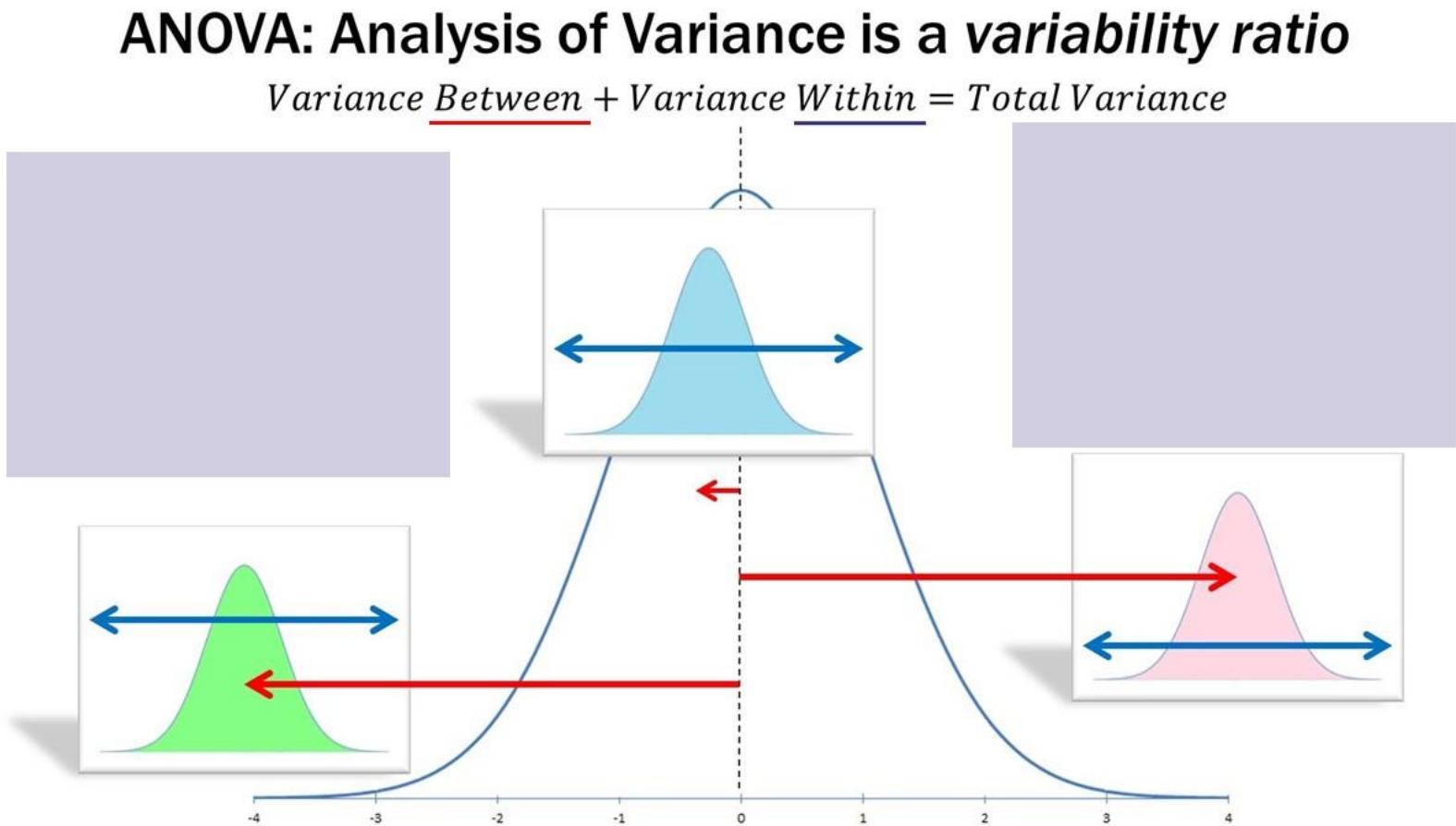
$$x_{ij} - \bar{x} = x_{ij} - \bar{x} + \bar{x}_i - \bar{x}_i$$

$$= (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})$$

1. The differences between the population means

2. The variation of the data within one sample

Hypothesis test for more than one population mean
ANOVA



Hypothesis test for more than one population mean
ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$

Sum of Squares Within (SSW) =

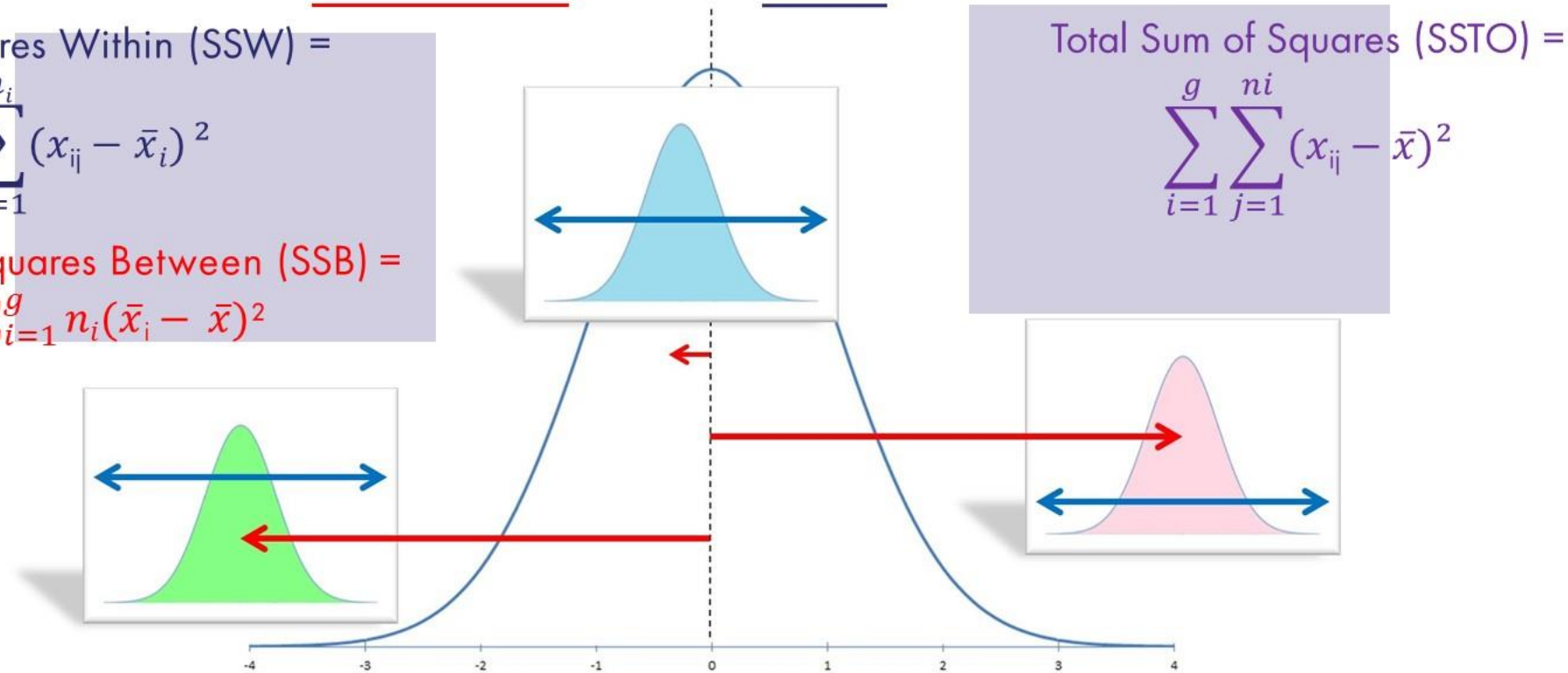
$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Sum of Squares Between (SSB) =

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

Total Sum of Squares (SSTO) =

$$\sum_{i=1}^{g}\sum_{j=1}^{ni}(x_{ij} - \bar{x})^2$$



Sum Of Squares Within = binnenkwadraatsom
Sum of Squares Between = tussenkwadraatsom
Total Sum of Squares = totale kwadraatsom

Hypothesis test for more than one population mean
ANOVA – output interpretation

Source of **variation**	Sums **of Squares**	Degrees **of Freedom**	Mean **Squares**	F-**statistic**	P-**value**
Between variation	SSB	g - 1	MSB	f	p
Within variation	SSW	n - g	MSW		
Total variation	SSTO	n - 1			

Degrees of freedom: The number of values in the final calculation of a statistic that are free to vary.

$$\frac{MSB}{MSW} = \frac{\frac{SSB}{g-1}}{\frac{SSW}{n-g}}$$

$$F_{g-1,n-g} = \frac{\frac{SSB}{g-1}}{\frac{SSW}{n-g}} = \frac{MSB}{MSW}$$

F = Between group variability / Within group variability

Hypothesis test for more than one population mean
ANOVA – output interpretation

Source of **variation**	Sums **of Squares**	Degrees **of Freedom**	Mean **Squares**	F-**statistic**	P-**value**
Between variation	SSB	g - 1	MSB	f	p
Within variation	SSW	n - g	MSW		
Total variation	SSTO	n - 1			

$p = P(F_{g-1,n-g} > f)$

e.g. $F_{\alpha,g-1,n-g} = F_{0,05,10,20}$

$= 2,348$

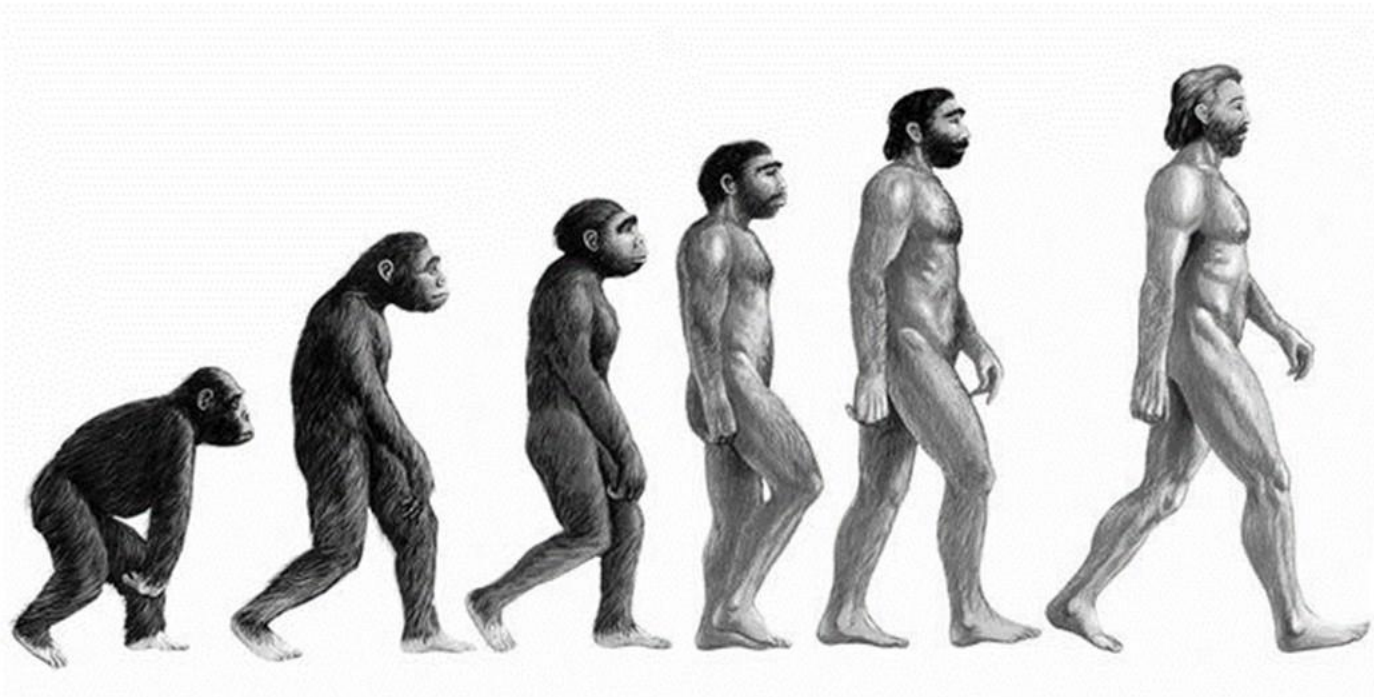
Suppose F-statistic: f =3

$P < \alpha \Rightarrow H_o$ verwerpen

Linear regression

Linear regression

If you could change one thing in this world, what would it be?



Linear regression

How much will our sales volume increase?

If we change the pricing of our product?

Keeping all other factors constant

How much will our sales volume increase?

If we change the positioning of our product?

Keeping all other factors constant

Disentangle the different predictors

The different predictors might be correlated with each other

Which predictors are significant?

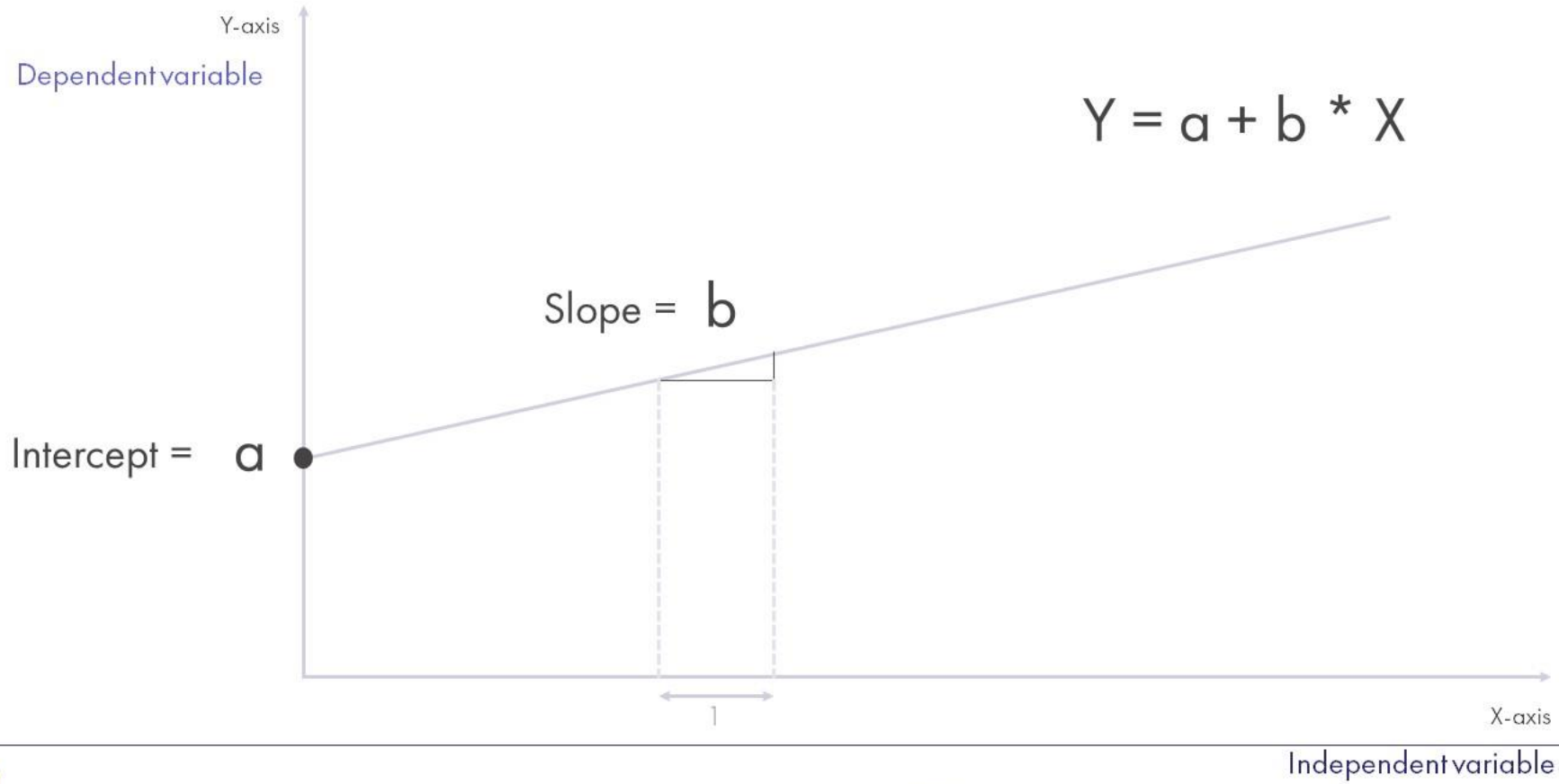
Linear regression

How does the change in one predictor relate to changes the dependent variable?

The other predictors are kept ‘Ceteris paribus’ / constant.

Linear Regression
The basics

Two-dimensional model



Linear Regression equation

Step one: specify the determinstic part of the model

$$E(y) = \beta_0 + \beta_1 * x_i$$

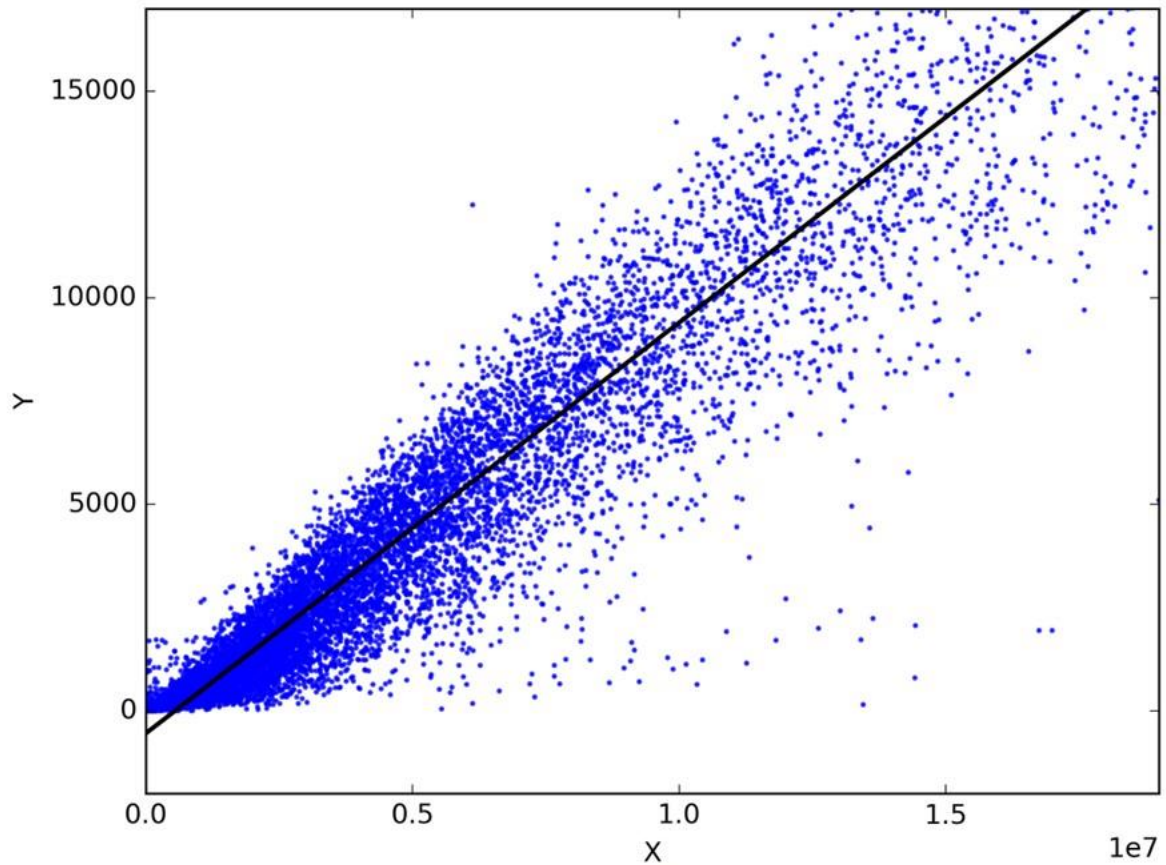
Step two: check plausability of an effect by drawing a scatterplot

Step three: look for the line that minimizes the deviations between all the points on the scatterplot

We call this the least squares line.

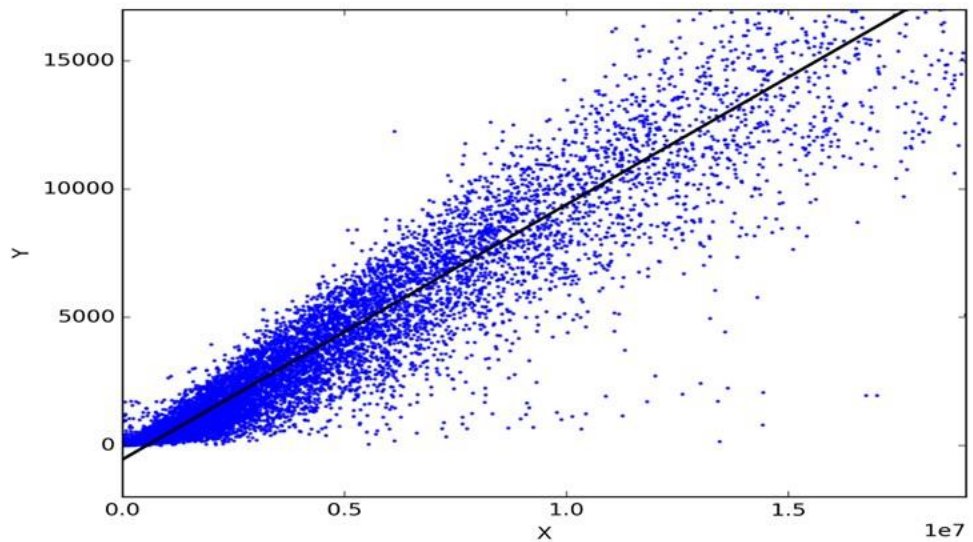
Linear regression

The regression line is to a scatter diagram as the average is to a list.



Linear regression

The regression line estimates the average value for the dependent variable, corresponding to each value of the independent variable.



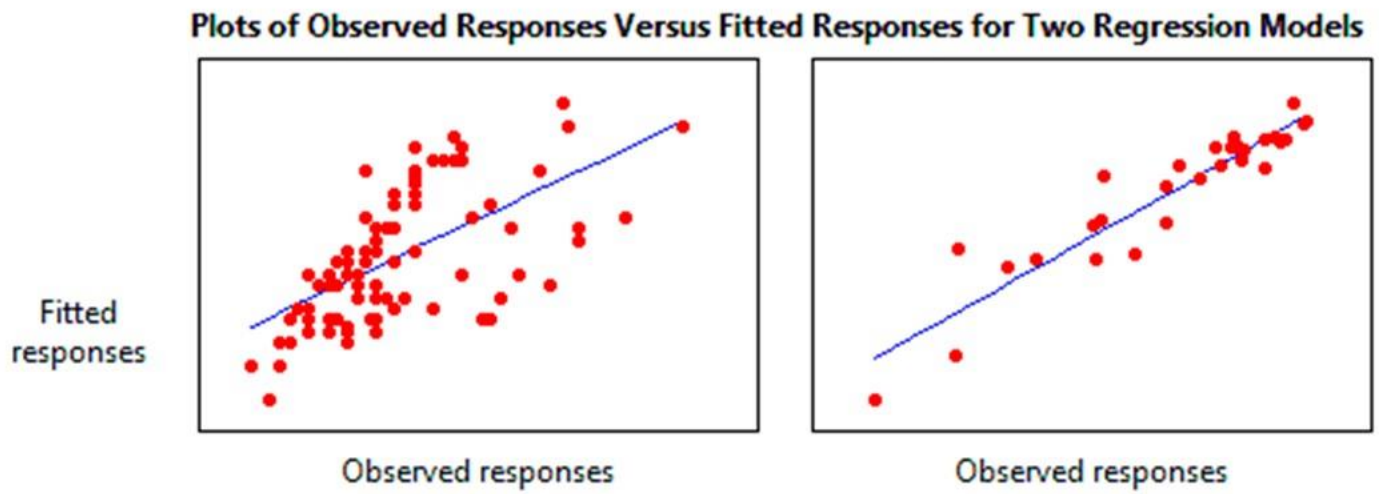
The regression line predicts the average y value assiociated with a given x value.

How good is our estimation?
Goodness-of-fit for a linear model

R^2 is a measure to indicate how close the data are to the fitted regression line.

$0\% \leq R^2 \leq 100\%$

General rule: The higher R^2, the better the model fits your data



$R^2 = 38\%$

$R^2 = 87{,}4\%$

Linear regression
Link to hypothesis testing

H_0: The predictor has no effect

$p > \alpha$

The coefficient of the predictor is equal to 0

$H_0: \beta_1 = 0$

H_a: The predictor has a significant effect

$p < \alpha$

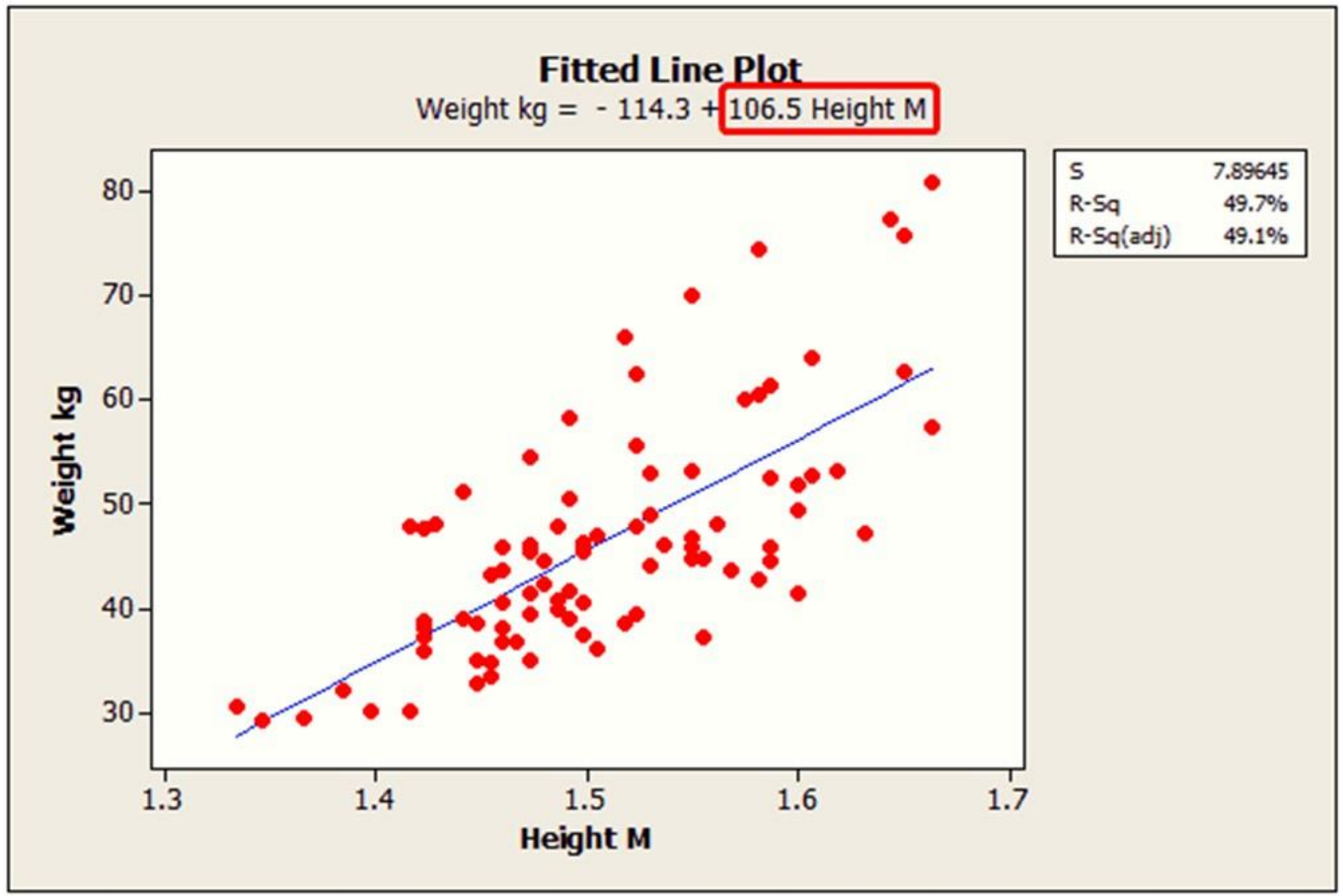
The coefficient of the predictor can be found in the output

$H_a: \beta_1 \neq 0$

Linear regression
Regression coefficients

Coefficients

Term	Coef	SE Coef	T	P
Constant	-114.326	17.4425	-6.55444	0.000
Height M	106.505	11.5500	9.22117	0.000



Linear regression
Tips

1. Which variables to include in your study?

2. Keep it simple
   In many cases only 3 predictors are sufficient

3. Use sound reasoning to include predictors in the model

4. Use large amounts of trustworthy data

Interpreting your results

Linear Regression
Excel– Interpret output

13	SUMMARY OUTPUT	
14		
15	*Regression Statistics*	
16	Multiple R	0,97102749
17	R Square	0,942894387
18	Adjusted R Square	0,91434158
19	Standard Error	379,349634
20	Observations	7

1. Adjusted R^2 should be close to 1

22	ANOVA					
23		*df*	*SS*	*MS*	*F*	*Significance F*
24	Regression	2	9504375,421	4752187,71	33,0228269	0,003261051
25	Residual	4	575624,5793	143906,1448		
26	Total	6	10080000			

2. Significance F should be < α (0,005)

Linear Regression
Excel– Interpret output

28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
29	Intercept	8324,406193	493,921259	16,85371107	7,26514E-05
30	Price	-825,56015	121,7130698	-6,782838946	0,002466352
31	Advertising	0,690451005	0,153379861	4,501575361	0,010809504

3. Look at the p-values. When $p < \alpha$, the corresponding predictor is significant

4. Quantity Sold = intercept + a * Price + b * Advertising
=> Quantity Sold = 8324 -825,56 * Price + 0,69 * Advertising

5. This can be used to make a forecast of how much will be sold
If the price equals 2 and advertising equals 2800.
⇒ Quantity Sold = 8324 -825,56 * 2 + 0,69 * 2800
⇒ 8604,88

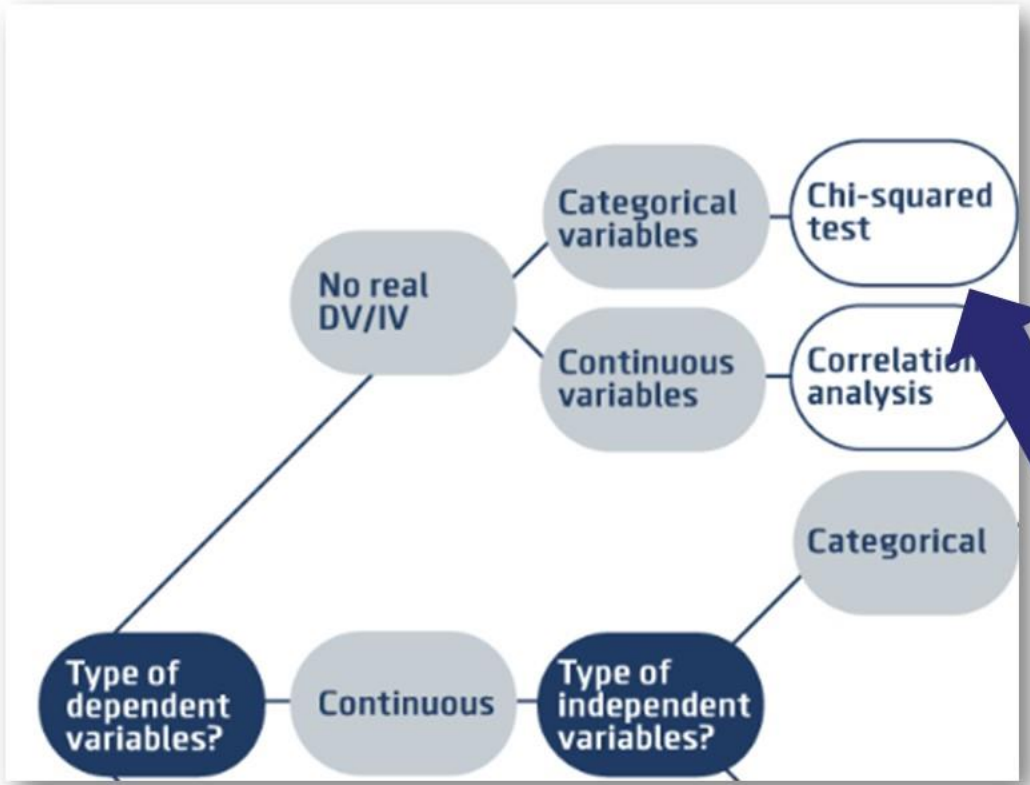
Linear Regression
Excel – Interpret output

35	RESIDUAL OUTPUT		
36			
37	*Observation*	*Predicted Quantity sold*	*Residuals*
38	1	8606,548707	-106,5487066
39	2	4610,876046	89,12395423
40	3	6400,086547	-600,0865468
41	4	7018,511395	381,4886047
42	5	6129,868257	70,13174344
43	6	7090,537552	209,4624483
44	7	5643,571497	-43,57149726

5. Residual: What is the difference between the forecast and the actual quantity sold

Other possible outputs
Interpretation exercise

Is there a difference in the amount of smokers and non-smokers comparing men to women?



Count

		Gender		
		Male	Female	Total
Do you smoke cigarettes?	Nonsmoker	149	148	297
	Past smoker	13	24	37
	Current smoker	31	37	68
Total		193	209	402

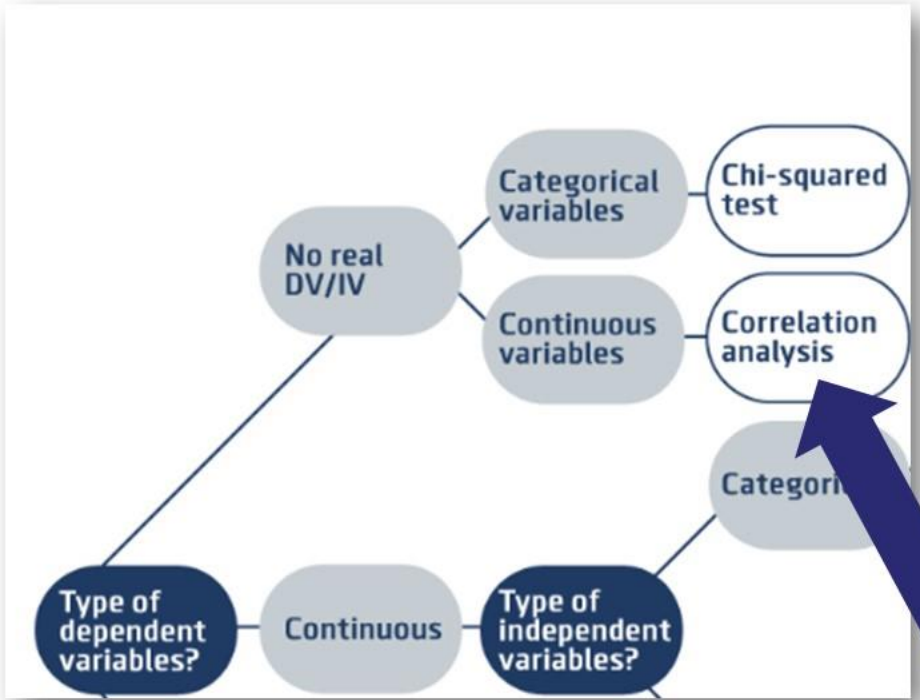
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.171[a]	2	.205
Likelihood Ratio	3.217	2	.200
Linear-by-Linear Association	1.106	1	.293
N of Valid Cases	402		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.76.

Other possible outputs
Interpretation exercise

Are student typically better at certain groups of courses?



Correlations

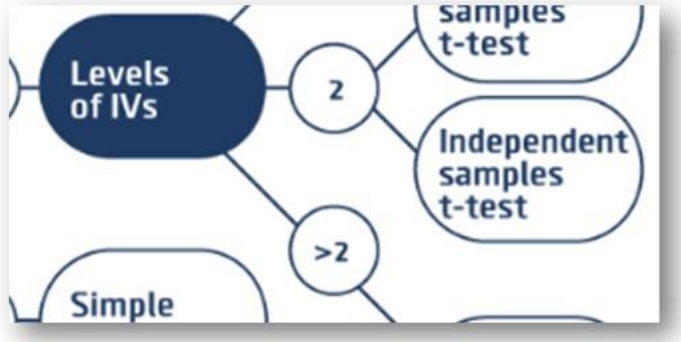
		reading score	writing score	math score	science score	female
reading score	Pearson Correlation[a]	1	.597**	.662**	.630**	-.053
	Sig. (2-tailed)[b]	.	.000	.000	.000	.455
	N[c]	200	200	200	200	200
writing score	Pearson Correlation	.597**	1	.617**	.570**	.256**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	200	200	200	200	200
math score	Pearson Correlation	.662**	.617**	1	.631**	-.029
	Sig. (2-tailed)	.000	.000	.	.000	.680
	N	200	200	200	200	200
science score	Pearson Correlation	.630**	.570**	.631**	1	-.128
	Sig. (2-tailed)	.000	.000	.000	.	.071
	N	200	200	200	200	200
female	Pearson Correlation	-.053	.256**	-.029	-.128	1
	Sig. (2-tailed)	.455	.000	.680	.071	.
	N	200	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

Other possible outputs
Interpretation exercise

Group Statistics

	female	N	Mean	Std. Deviation	Std. Error Mean
write writing score	.00 male	91	50.1209	10.30516	1.08027
	1.00 female	109	54.9908	8.13372	.77907



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
write writing score	Equal variances assumed	11.133	.001	-3.734	198	.000	-4.86995	1.30419	-7.44183	-2.29806
	Equal variances not assumed			-3.656	169.70	.000	-4.86995	1.33189	-7.49916	-2.24073

Other possible outputs
Interpretation exercise

One-Sample Kolmogorov-Smirnov Test

	N	Most Extreme Differences			Test Statistic	Asymp. Sig. (2-tailed)
		Absolute	Positive	Negative		
Reaction time trial 1	233	.073	.073	-.031	.073	.0047[c]
Reaction time trial 2	233	.090	.090	-.086	.090	.0001[c]
Reaction time trial 3	235	.385	.385	-.219	.385	.0000[c]
Reaction time trial 4	226	.045	.045	-.027	.045	.2000[c,d]
Reaction time trial 5	235	.120	.120	-.067	.120	.0000[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

P < 0.05? REJECT NULL HYPOTHESIS OF NORMAL POPULATION DISTRIBUTION

Classifiers
Interpretation exercise

Kappa statistic

Mean Absolute Error

Root mean squared error

Relative absolute error

Root relative squared error

```
Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

outlook:
        sunny    -> no
        overcast        -> yes
        rainy    -> yes
(10/14 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.2444
Mean absolute error                      0.5714
Root mean squared error                  0.7559
Relative absolute error                120       %
Root relative squared error            153.2194 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area
               0.556     0.800     0.556       0.556   0.556      -0.244   0.378
               0.200     0.444     0.200       0.200   0.200      -0.244   0.378
```

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Kappa statistic

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

A

$$totalAccuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B

$$randomAccuracy = \frac{(TN + FP)*(TN + FN) + (FN + TP)*(FP + TP)}{Total*Total}$$

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Kappa statistic

Our labels

Machine's labels	CATS	DOGS
CATS	10	7
DOGS	5	8

$$totalAccuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{10+8}{30} = 0.6$$

$$randomAccuracy = \frac{(8+5)\times(7+8)+(10+5)\times(10+7)}{30\times 30} = 0.5$$

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Kappa statistic

Our labels

Machine's labels	CATS	DOGS
CATS	10	7
DOGS	5	8

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy} = \frac{0.6 - 0.5}{1 - 0.5} = \frac{0.1}{0.5} = 0.2$$

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Kappa statistic

Our labels

Machine's labels	CATS	DOGS
CATS	10	7
DOGS	5	8

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy} = \frac{0.6 - 0.5}{1 - 0.5} = \frac{0.1}{0.5} = 0.2$$

A kappa value of 0 means that the result is the same as would be expected by chance.

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Mean absolute error

Quantity used to measure how close forecasts or predictions are to the eventual outcomes.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Root mean squared error

Similar to MAE, more emphasis on outside boundaries.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}$$

Classifiers
Interpretation exercise

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

Relative errors

The error is made relative to what it would have been if a simple predictor had been used. The simple predictor in question is just the average of the actual values from the training data. Thus relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the default predictor

Thank you

See you in the next part!