Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

DATA-DRIVEN TRANSFORMATION

INTRODUCTION TO VARIANCE

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

WHAT IS VARIANCE?

- Variance refers to the **variations** of measurements drawn from **samples** (i.e., sets of data points)

- Variance refers to the **differences** between the measurements for **each sample** element and the **sample mean**

- We use **squared** deviations from the mean because the mean deviation is zero.

$$V(X) = \frac{1}{n-1} \sum_i (x_i - \mu)^2$$

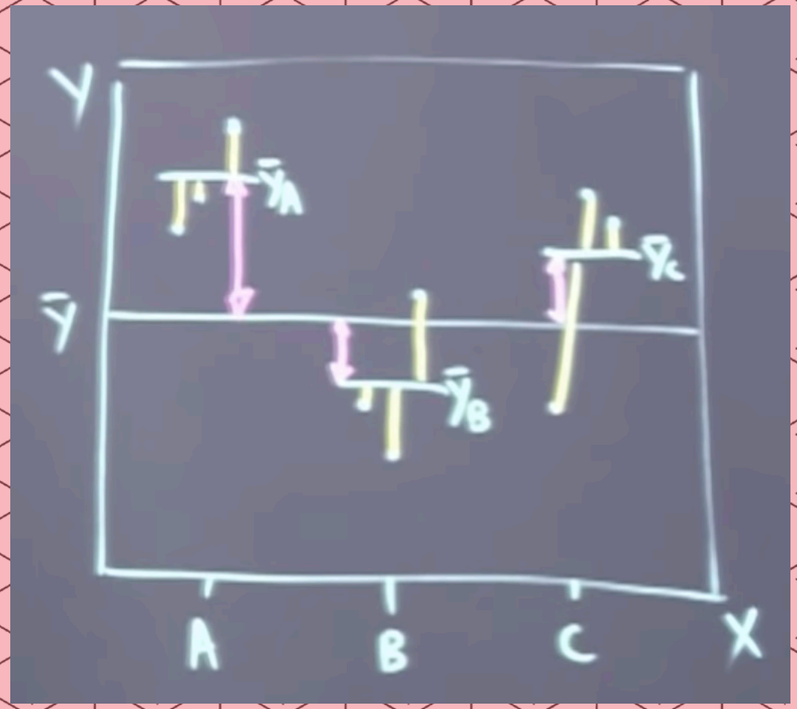
WHAT IS COVARIANCE?

- Covariance refers to the **variability of two variables**.
- For example, if data points have high values for both variables, and low values for both variables: covariance is positive.
- Normalized covariance is **correlation**.

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i} (x_i - \mu_x)(y_i - \mu_y)$$

WHAT IS ANOVA?

- ANOVA stands for **analysis of variance**.
- It compares the variance **within samples** (i.e., within groups or experimental conditions), and **between samples**.
- Components of the variance are either **explained** (due to the experimental conditions) or **unexplained** (due to random sample variance).

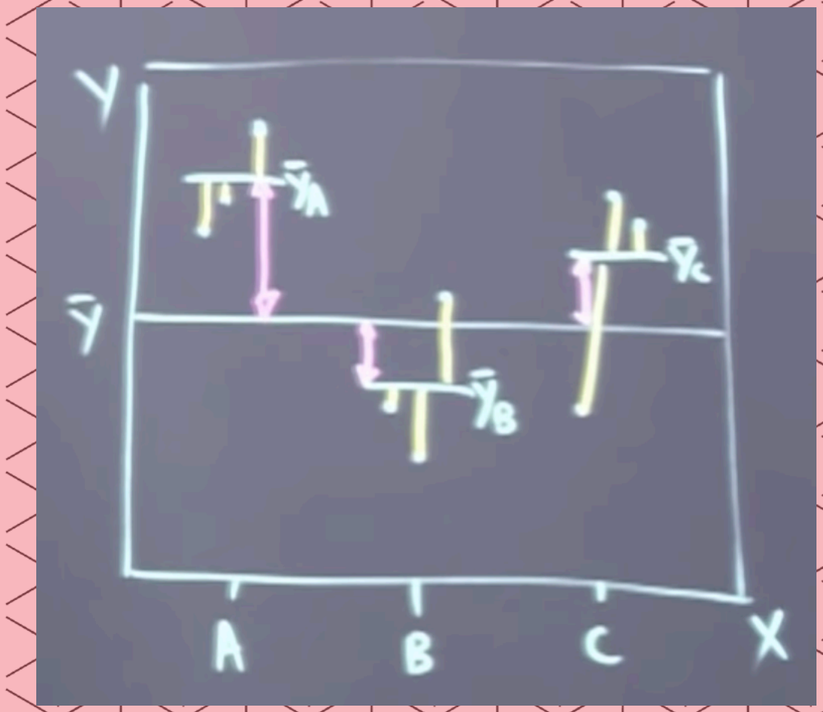


WHAT IS ANOVA?

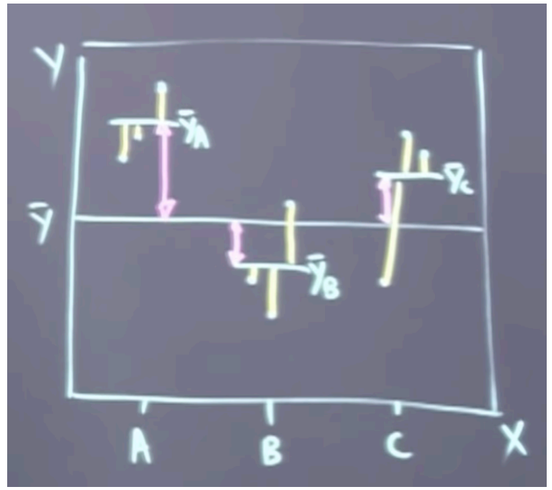
- **If variances within samples & between samples are equal**, the experimental conditions have no significant effect on the dependent variable. Their effect is similar to random deviations.

- **If variances between samples is greater than within sample**, the experimental conditions have an effect on the dependent variable. But it may be due to chance.

- Only **statistical tests** can estimate the effect of chance.



ANOVA IN DETAILS



$$Explained\ Variance = \frac{Sum\ of\ Squared\ Deviations}{Degree\ of\ Freedom} = \frac{\sum Group\ Size \times (Group\ Mean - Overall\ Mean)^2}{Number\ of\ Groups - 1} = \frac{\sum_i n_i (y_i - \bar{y})^2}{k - 1}$$

$$Unexplained\ Variance = \frac{Sum\ of\ Squared\ Deviations}{Degree\ of\ Freedom} = \frac{\sum (Data\ Point - Group\ Mean)^2}{Number\ of\ Data\ Points - Number\ of\ Groups} = \frac{\sum_j (y_{ij} - \bar{y}_i)^2}{n - k}$$

ANOVA IN DETAILS

$$F = \frac{Explained\ Variance}{Unexplained\ Variance}$$

- This F value is what we need to run a statistical test

MANOVA

- MANOVA stands for **multivariate analysis of variance**.

- It handles **multiple dependent variables**.

https://www.youtube.com/watch?v=e0EBqqwEZr4

ANCOVA

- ANCOVA stands for **analysis of covariance**.
- Accounts for confounding factors (**covariates**) that impact the dependent variable (thus covary with it).
- **Group means are adjusted** to account for covariance between dependent variable and covariates.

Covariates are variables measured for each data point.

https://www.youtube.com/watch?v=0e8BI2u6DU0

MANCOVA

- It stands for **multivariate analysis of covariance**.

- It handles **covariates** and **multiple dependent variables**.

ASSUMPTIONS FOR ANOVA

- **Observations are independent** (e.g., no relationship between observations of each group).
- Independent variables are **categorical**.
- Dependent variables are **numerical**, **normally distributed**, and their **variance within groups is equivalent** (homoscedacity).

ASSUMPTIONS FOR ANCOVA

- Covariates are **numerical**, **normally distributed**, with **equivalent variance** (homoscedacity).
- The relationships between covariates and dependent variables are **linear**.
- **Covariance** between covariates and dependent variable is **equivalent among groups** (covariates must affect all groups in the same manner).

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

DATA-DRIVEN TRANSFORMATION

INTRODUCTION TO STATISTICAL TESTS

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

WHAT IS A STATISTICAL TEST?

- Formally called **statistical hypothesis testing**
- **Hypotheses** are interpretations of the facts represented in the data
- The **null hypothesis** assumes that samples (data subsets) are **not significantly different**, i.e., they come from the same population
- **Statistical tests** compares the null hypothesis against other hypotheses

FIND EXAMPLES OF HYPOTHESES

- Did this campaign improve sales?
- Is this kind of people more responsive to emails?
- Do we make more sales in December?

USE CASES & TYPES OF TESTS

- How many **dependent variables**?
  Are they continuous, discrete or categorical?

- How many **independent variables**?
  Are they continuous, discrete or categorical?

Variables that you observe,
that are not fixed before the experiment starts

Variables that you control,
to create experimental conditions

http://www.dataanalytics.org.uk/Data%20Analysis/Statistics/choosing-your-stats-test.htm

USE CASES & TYPES OF TESTS

- **How many data points & samples**?
- Do we **describe or compare** them?
- What was the **sampling method**?
  Random, stratified, paired, panel sampling?
  With or without replacement?

https://en.wikipedia.org/wiki/Sampling_(statistics)#Sampling_methods

T-TEST

- When comparing **2 samples** or more,
  are the **dependent variables** different?
  Are the differences **statistically significant?**

- <u>Assumption</u>: the dependent variables are **numerical** and **normally distributed**.

T-tests are ANOVA F-tests with 2 groups (i.e., independent variables)

https://www.youtube.com/watch?v=vwWEa8wU_6U

T-TEST

$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$

A big t-value = different groups

A small t-value = similar groups

https://www.youtube.com/watch?v=0Pd3dc1GcHc

T-VALUE TO P-VALUE

- **What is a p-value?**

- The **probability** that the **differences** observed between the samples **have occurred by chance**.

- In other words, it is the **probability that samples are drawn from the same population**, and differences occurred because we randomly sampled subsets of the population.

P-value = 0.05

It means there is 5% probability that the differences occurred by chance.

Of all samples drawn from the same population, 5% of them would display such differences (or more differences).

5% occurrence is quite frequent. It's a 1/20 chance.

P-value of 0.01 are more conservative.

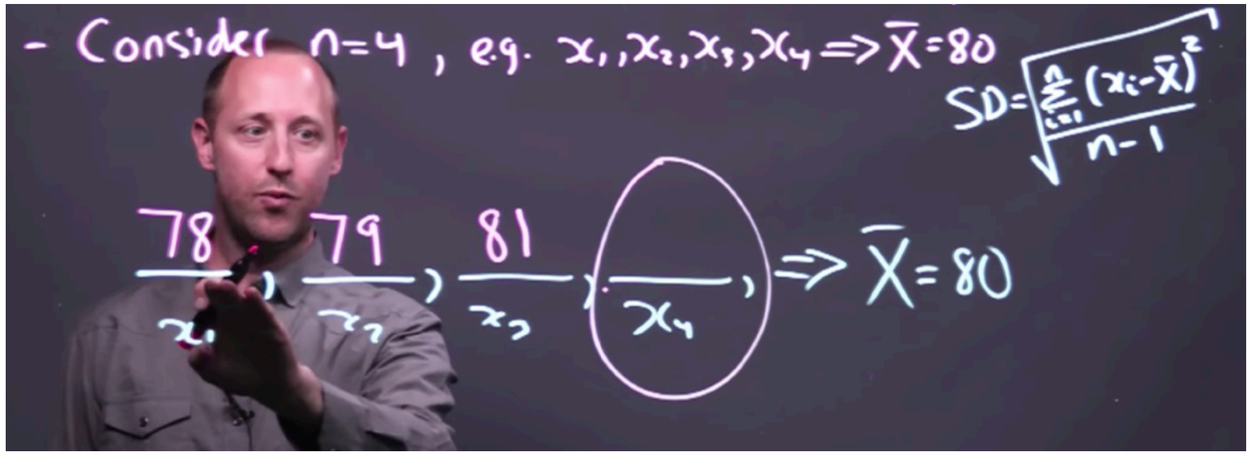
T-VALUE TO P-VALUE

- There is **no exact equation** to calculate p-values from t-values.
- **Approximation** methods exist. They are used in R libraries, and others.
- Approximations are also available in **tables**, e.g., giving p-values in function of t-values.

https://en.wikipedia.org/wiki/Student%27s_t-distribution#Table_of_selected_values

T-VALUE TO P-VALUE

- P-values depend on the **degrees of freedom**.
- Basically **each data point adds a degree** of freedom. Except one point.



https://www.youtube.com/watch?v=nlm9gfso4mw

With fewer degrees of freedom, i.e., fewer data points, the p-value is larger.

T-TEST VARIANTS DEPEND ON USE CASES

Equal or unequal sample sizes, unequal variances [edit]

Main article: Welch's t-test

This test, also known as Welch's *t*-test, is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal) and hence must be estimated separately. The t statistic to test whether the population means are different is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Here s_i^2 is the unbiased estimator of the variance of each of the two samples with n_i = number of participants in group i (1 or 2).

https://en.wikipedia.org/wiki/Student%27s_t-test#Independent_two-sample_t-test

CHI-SQUARED TEST

- Compares **counts of data points per category**.
- Tests if **observed values** are different from **expected values** if classes are not associated (e.g., probability of autism is independent of vaccination).
- Assumption: the independent variables are **categorical** and **mutually exclusive**.

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

https://www.youtube.com/watch?v=vwWEa8wU_6U

CHI-SQUARED TEST

	Autism	No Autism	(total)
Vaccin	?	?	40 000
No Vaccin	?	?	100 000
(total)	500	130 200	140 000

Null Hypothesis

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

Observed

https://www.youtube.com/watch?v=vwWEa8wU_6U

CHI-SQUARED TEST

	Autism	No Autism	(total)
Vaccin	$\frac{40\,000 \times 500}{140\,000}$	?	**40 000**
No Vaccin	?	?	100 000
(total)	**500**	130 200	**140 000**

Null Hypothesis

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

Observed

https://www.youtube.com/watch?v=vwWEa8wU_6U

CHI-SQUARED TEST

	Autism	No Autism	(total)
Vaccin	142	?	**40 000**
No Vaccin	?	?	100 000
(total)	**500**	130 200	**140 000**

Null Hypothesis

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

Observed

- What is the **probability** that the **differences** observed between the tables **have occurred by chance**?

RISK RATIO *a.k.a. Relative Risk*

- Chi-Squared tests may identify an effect between variables, but not **what kind of effect**?
  (e.g., more autism if vaccine?)

- Risk ratios compare the **probability of belonging to a category** given membership to another category.
  (e.g., there 3.75 times more chance of "Autism" if one is "Vaccine" rather than "No Vaccine"?)

$$P(Autism|Vaccine) = \frac{300}{40\,000} = 0.75\,\%$$

$$P(Autism|NoVaccine) = \frac{200}{100\,000} = 0.2\,\%$$

$$\frac{P(Autism|Vaccine)}{P(Autism|NoVaccine)} = \frac{0.75\,\%}{0.2\,\%} = 3.75$$

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

https://www.youtube.com/watch?v=NqsSun9HZfI

ODD RATIO

- Odd ratios compare the **odds of belonging to a category**, given membership to another category.

- **Odds** are the probability of an event happening, divided by the probability of it not happening.

$$Odd(A) = \frac{P(A)}{P(\neg A)} \qquad Odd(A \mid B) = \frac{P(A \mid B)}{P(\neg A \mid B)}$$

- Risk & odd ratios are similar if a category is rare.

$$Odd(Autism \mid Vaccine) = \frac{300}{39\,700} = 0.76\,\%$$

$$Odd(Autism \mid NoVaccine) = \frac{200}{99\,800} = 0.2\,\%$$

$$\frac{Odd(Autism \mid Vaccine)}{Odd(Autism \mid NoVaccine)} = \frac{0.76\,\%}{0.2\,\%} = 3.8$$

	Autism	No Autism	(total)
Vaccin	300	39 700	40 000
No Vaccin	200	99 800	100 000
(total)	500	130 200	140 000

https://www.youtube.com/watch?v=NqsSun9HZfI

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

DATA-DRIVEN TRANSFORMATION

QUESTION / DISCUSSION