Lecture 6: A/B/n testing with ANOVA

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

January 15, 2024

Plan

Plan

1) ANOVA

2) ANCOVA and other variations

3) Multiple hypothesis comparison

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances
- ANOVA uses variance-based F test to check the group mean equality. Sometimes, ANOVA F test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances
- ANOVA uses variance-based F test to check the group mean equality. Sometimes, ANOVA F test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal

What is ANOVA (ANalysis Of VAriance)?

- ANOVA test used to compare the means of more than 2 groups (t-test and it's variations can be used to compare 2 groups)
- Groups mean differences inferred by analyzing variances
- ANOVA uses variance-based F test to check the group mean equality. Sometimes, ANOVA F test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal

Null hypothesis:

Groups means are equal (no variation in means of groups)

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

k – is the number of groups

Alternative hypothesis:

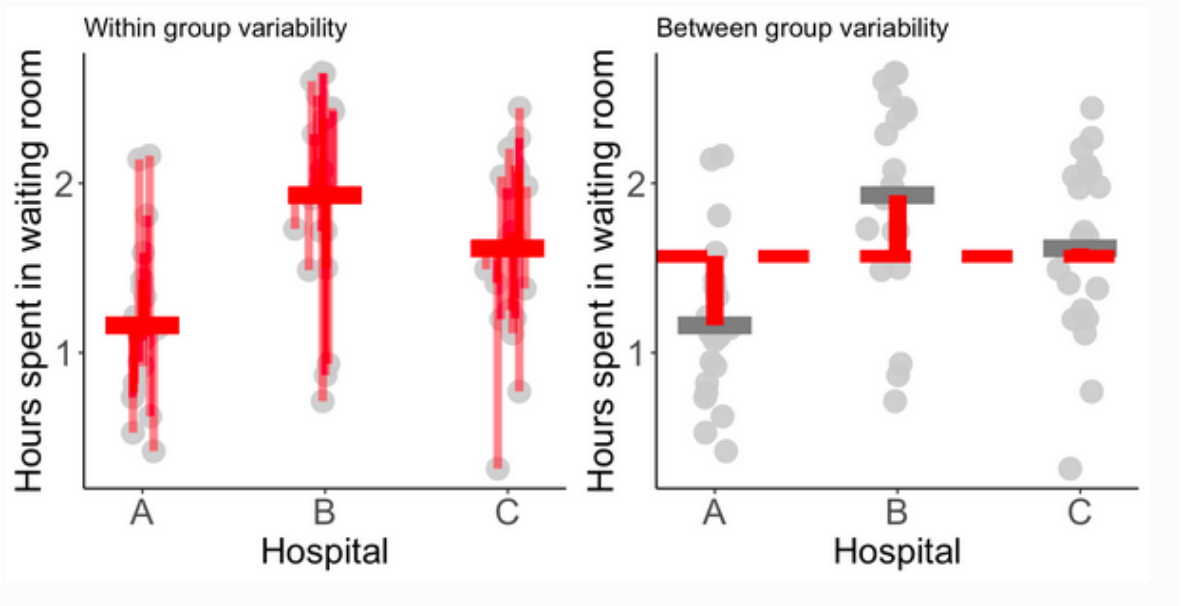
At least, one group mean is different from other groups H_1 : All μ are not equal

Idea

Compare

- within-group variability: the variance of the individual observations within a group, and

- between-group variability: the variance between the averages of the groups.

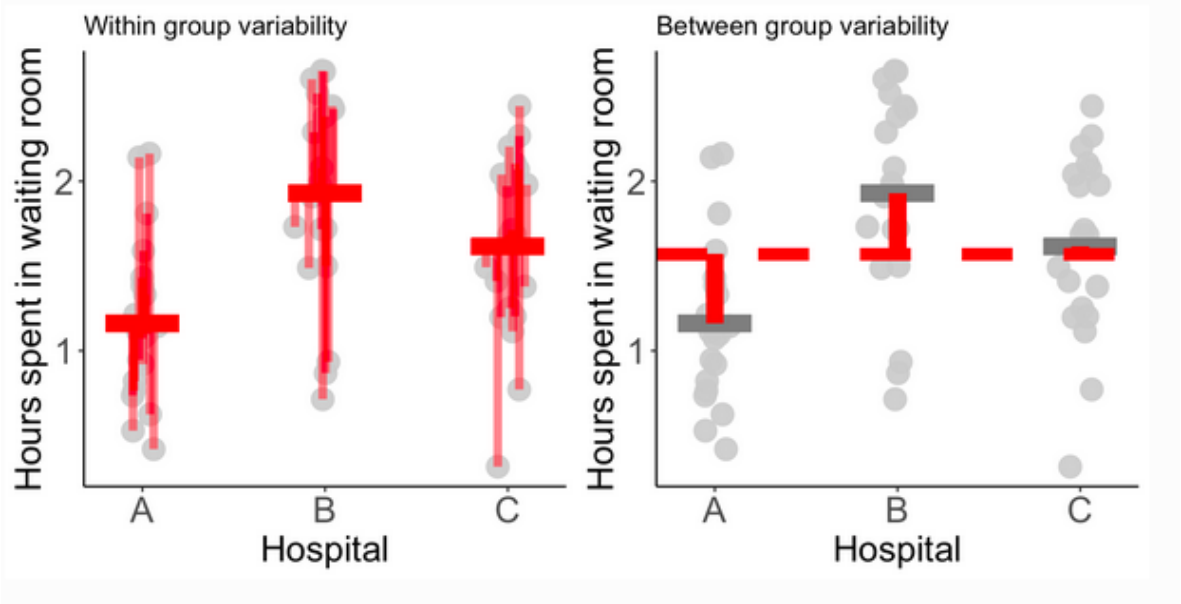


Idea

Compare

- within-group variability: the variance of the individual observations within a group, and

- between-group variability: the variance between the averages of the groups.

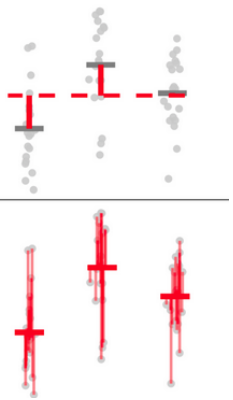


The basic idea is that if the variability between the groups is greater than the variability within the groups, then we have evidence that the differences between the groups is not simply reflecting random noise.

Idea

Related F-statistics:

$$\frac{\text{BSS / (K - 1)}}{\text{WSS / (N - K)}} =$$



$$WSS = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 \quad \text{and} \quad BSS = \sum_{i=1}^{K} (\bar{y}_{..} - \bar{y}_{i.})^2$$

where y_{ij}, defines the waiting room time (outcome) for patient j from hospital i, $\bar{y}_{..}$ defines the global average waiting time and $\bar{y}_{i.}$ defines the average waiting time for hospital i.K is the number of hospitals, and n_i is the number of patients sampled from hospital i.

Hence:

$$F = \frac{Var_{between}}{Var_{within}} = \frac{BSS/(K-1)}{WSS/(N-K)} \sim F_{K-1,N-k}$$

Sum of Squares (SS)

Inside the One-Way ANOVA Table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

Sum of Squares (SS)

Inside the One-Way ANOVA Table:

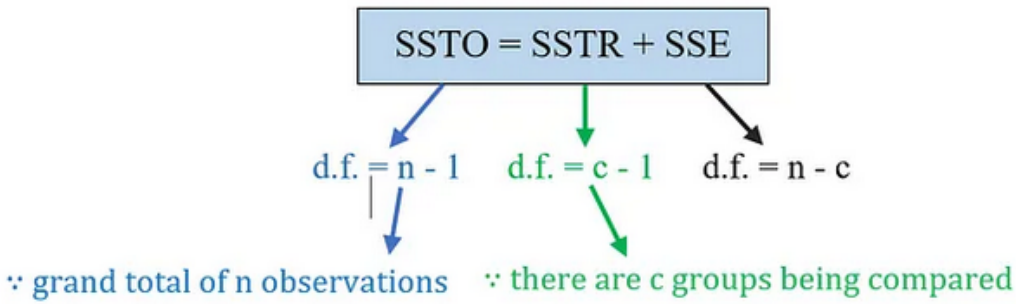
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

The total amount of variability comes from two possible sources, namely:

1. Difference among the groups, called treatment (TR)

2. Difference within the groups, called error (E)

The sum of the squares due to treatment (SSTR) and the sum of squares due to error (SSE) are listed in the one-way ANOVA table. The sum of SSTR and SSE is equal to the total sum of squares (SSTO).



Regression form

$$y_{ik} = \mu + \alpha_k + \epsilon_{ik}$$

$$SS_T = SS_B + SS_E$$

$$TSS = ESS + RSS$$

Where, y_{ik} – i^{th} observation of k^{th} level of groups, μ = overall population mean (unknown), α_k = Main effect for groups (deviation from the μ), ϵ_{ik} = Error, k = levels for groups $k = 1, 2 \ldots, \mathrm{p})$, i = Observations or replicates for each group $(i = 1, 2 \ldots, r)$,

Where,
$SS_B = \sum_i p_i \left(\overline{y_{i.}} - \bar{y}..\right)^2$, $SS_E = \sum_{ik} \left(y_{ik} - \overline{y_{i.}}\right)^2$, $SS_T = SS_B + SS_E = \sum_{ik} \left(y_{ik} - \overline{y_{..}}\right)^2$

F-test for regression significance:

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \sim F_{k-1,n-k}$$

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.
- **Homoscedasticity or Homogeneity of variance** – all the k group variances are equal, that is ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = \sigma_k^2$.
  Rule of thumb: the ratio of the largest to the smallest sample standard deviation is less than 2
  Check: Levene's, Bartlett's, or Brown-Forsythe test

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.
- **Homoscedasticity or Homogeneity of variance** – all the k group variances are equal, that is ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = \sigma_k^2$.
  Rule of thumb: the ratio of the largest to the smallest sample standard deviation is less than 2
  Check: Levene's, Bartlett's, or Brown-Forsythe test
- **Dependent variable is continuous**

  If the dependent variable is ordinal or rank (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances.

  If these assumptions are violated, you should consider the non-parametric tests (e.g. Mann-Whitney U test, Kruskal-Wallis test).

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.
- **Homoscedasticity or Homogeneity of variance** – all the k group variances are equal, that is ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = \sigma_k^2$.
  Rule of thumb: the ratio of the largest to the smallest sample standard deviation is less than 2
  Check: Levene's, Bartlett's, or Brown-Forsythe test
- **Dependent variable is continuous**

  If the dependent variable is ordinal or rank (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances.

  If these assumptions are violated, you should consider the non-parametric tests (e.g. Mann-Whitney U test, Kruskal-Wallis test).

Assumptions for the one-way ANOVA hypothesis test

- **Observations are i.i.d** – sample data are randomly selected from populations and randomly assigned to each of the treatment groups
- **Normality**
  Check: normal probability plot, Q-Q plot, Shapiro-Wilks test, etc.
- **Homoscedasticity or Homogeneity of variance** – all the k group variances are equal, that is ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = \sigma_k^2$.
  Rule of thumb: the ratio of the largest to the smallest sample standard deviation is less than 2
  Check: Levene's, Bartlett's, or Brown-Forsythe test
- **Dependent variable is continuous**

  If the dependent variable is ordinal or rank (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances.

  If these assumptions are violated, you should consider the non-parametric tests (e.g. Mann-Whitney U test, Kruskal-Wallis test).

ANOVA is a powerful method when the assumptions of normality and homogeneity of variances are valid.

ANOVA is less powerful, if the assumption of normality is violated while variances are equal.

ANOVA types

Main types:

(i)

- One-way ANOVA (one factor)

- Two-way ANOVA (two factors)

ANOVA types

Main types:

(i)

- One-way ANOVA (one factor)

- Two-way ANOVA (two factors)

(ii)

- Univariate ANOVA – only one dependent variable in the model

- MANOVA – multiple dependent variables in the dataset

- ANCOVA – an additional continuous independent variable in the model is used

ANOVA types

Main types:

(i)

- One-way ANOVA (one factor)

- Two-way ANOVA (two factors)

(ii)

- Univariate ANOVA – only one dependent variable in the model

- MANOVA – multiple dependent variables in the dataset

- ANCOVA – an additional continuous independent variable in the model is used

(iii)

- Repeated Measure ANOVA – if you have repeated measurements for treatments or time on same subjects

ANOVA variations

1) **Repeated measures ANOVA**

Earlier: Two investigations of the same sample

$\longrightarrow$ paired t-test, not two-sample t-test

ANOVA: Generalisation from 2 to more measuring times

Subject	Time (min) 0	30	60	120
1	96	92	86	92
2	110	106	108	114
3	89	86	85	83
4	95	78	78	83
5	128	124	118	118
6	100	98	100	94
7	72	68	67	71
8	79	75	74	74
9	100	106	104	102

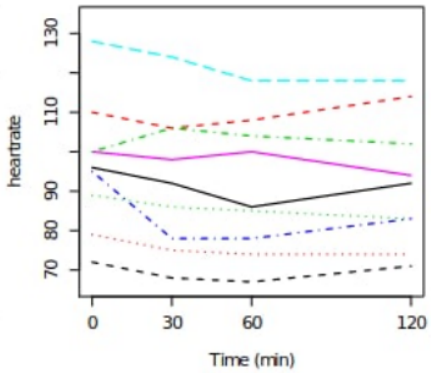


Figure: Example: Short-term effect of a drug on the heart-rate of 9 patients with heart disease

ANOVA variations

Model: repeated measures ANOVA

$$y_{ij} = \mu + \alpha_i + b_j(t_i) + \varepsilon_{ij}$$

$t_i -$ time points, measuring times, $i = 1, \ldots, m$

$\mu + \alpha_i -$ mean trend

$j = 1, \ldots, J -$ individuals

$b_j(t_i) -$ individual (random) effect of person j at time t_i

$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

(i) Univariate ANOVA for repeated measures

$$\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_s^2 \text{ (compound symmetry)}$$

(ii) Multivariate one-way model (MANOVA)

$$\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_{i_1 i_2} \text{ (un-structured)}$$

ANOVA variations

Repeated measures ANOVA

$$y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$
$$b_j \sim \mathcal{N}\left(0, \sigma_s^2\right) - \text{ person (subject) effect}$$

where, $\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_s^2$ (compound symmetry)

Assumptions:

- Normality (Do not assume, but verify)
  If normal distribution or equal variances cannot be confirmed, transform data or use Kruskal-Wallis test.

ANOVA variations

Repeated measures ANOVA

$$y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$
$$b_j \sim \mathcal{N}\left(0, \sigma_s^2\right) - \text{ person (subject) effect}$$

where, $\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_s^2$ (compound symmetry)

Assumptions:

- Normality (Do not assume, but verify)
  If normal distribution or equal variances cannot be confirmed, transform data or use Kruskal-Wallis test.
- Samples stem from population with equal variances (Do not assume, but verify)
  Why: as variances are equal within all groups, all observations are used to estimate the variance (pooling). $\longrightarrow$ more degrees of freedom, better power

ANOVA variations

Repeated measures ANOVA

$$y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$
$$b_j \sim \mathcal{N}\left(0, \sigma_s^2\right) - \text{ person (subject) effect}$$

where, $\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_s^2$ (compound symmetry)

Assumptions:

- Normality (Do not assume, but verify)
  If normal distribution or equal variances cannot be confirmed, transform data or use Kruskal-Wallis test.
- Samples stem from population with equal variances (Do not assume, but verify)
  Why: as variances are equal within all groups, all observations are used to estimate the variance (pooling). $\longrightarrow$ more degrees of freedom, better power
- Assumption of "compound symmetry" rarely valid for more than 2 measuring times
  Solution: Greenhouse-Geisser correction for deviations from "compound symmetry"

ANOVA variations

Repeated measures ANOVA

$$y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$
$$b_j \sim \mathcal{N}\left(0, \sigma_s^2\right) - \text{ person (subject) effect}$$

where, $\text{Cov}\left(b_j\left(t_{i_1}\right), b_j\left(t_{i_2}\right)\right) = \sigma_s^2$ (compound symmetry)

Assumptions:

- Normality (Do not assume, but verify)
  If normal distribution or equal variances cannot be confirmed, transform data or use Kruskal-Wallis test.
- Samples stem from population with equal variances (Do not assume, but verify)
  Why: as variances are equal within all groups, all observations are used to estimate the variance (pooling). $\longrightarrow$ more degrees of freedom, better power
- Assumption of "compound symmetry" rarely valid for more than 2 measuring times
  Solution: Greenhouse-Geisser correction for deviations from "compound symmetry"
- Sphericity – the variances of the differences between all possible pairs of within-subject conditions (i.e., levels of the independent variable) are equal.
  Check: Mauchly's sphericty test or Mauchly's W test

  If sphericity is violated, then the variance calculations may be distorted, which would result in an F-ratio that is inflated

ANOVA variations

2) **Two-Way Mixed ANOVA**

'Two-Way' – how many Independent Variables you have in your experimental design, in this case: two.

'Mixed' the nature of these variables.

While

a **'repeated-measures ANOVA'** contains only within participants variables (where participants take part in all conditions) and

an **'independent ANOVA'** uses only between participants variables (where participants only take part in one condition),

'Mixed ANOVA' contains BOTH variable types. In this case, one of each.

ANOVA variations

3) **ANCOVA**

(i) One-way ANOVA structural model

$$\boldsymbol{X}_{ij} = \mu + \tau_j + \mathbf{e}_{ij}$$

(ii) One-way ANCOVA structural model

$$X_{ij} = \mu + \alpha_j + \beta Z_{ij} + e_{ijk}$$

Covariate is just another source of variance

- Use the term βZ_{ij} because of continuous nature;

- Implicitly, we have specified no interaction between covariate and the independent variable (α)

Uses of ANCOVA

1. To control unwanted variation that would otherwise inflate the error with which we test our models (classical usage)

2. To control for group differences, esp. in the analysis of clinical trials or other pre/post designs

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

Assumptions of ANCOVA

1) all the regular ANOVA assumptions:

- homogeneous variance

- normal distribution

- independence of errors

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

Assumptions of ANCOVA

1) all the regular ANOVA assumptions:

- homogeneous variance

- normal distribution

- independence of errors

2) plus:

- relationship between covariate and DV is **linear**

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

Assumptions of ANCOVA

1) all the regular ANOVA assumptions:

- homogeneous variance

- normal distribution

- independence of errors

2) plus:

- relationship between covariate and DV is **linear**

- relationship between covariate and DV is linear **within each group**

ANOVA variations

How ANCOVA reduces error variance

- covariate = another predictor in the model but continuous
- if the covariate is associated with the DV and this relationship accounts for some systematic variance unexplained by the focal IV
- hence, a smaller error term because we've partioned out the variance due to the covariate means an increase in statistical power

Assumptions of ANCOVA

1) all the regular ANOVA assumptions:

- homogeneous variance

- normal distribution

- independence of errors

2) plus:

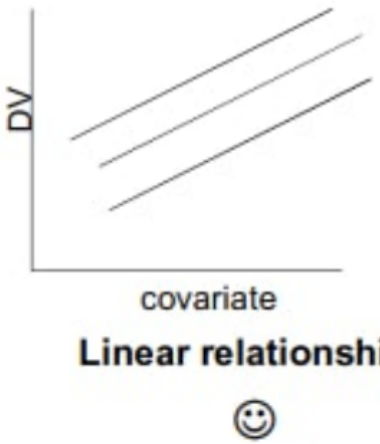
- relationship between covariate and DV is **linear**

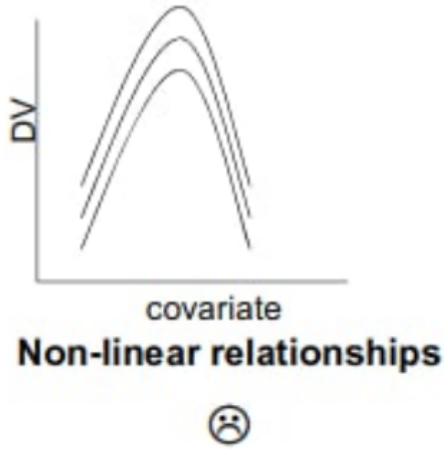
- relationship between covariate and DV is linear **within each group**

- relationship between DV and covariate is equal across treatment groups - **homogeneity of regression slopes**

ANOVA variations

Assumption: relationship between covariate and DV is **linear**

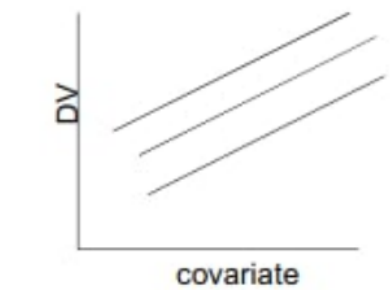




Non-linear relationships generally cannot be detected with ANCOVA - degrades power.

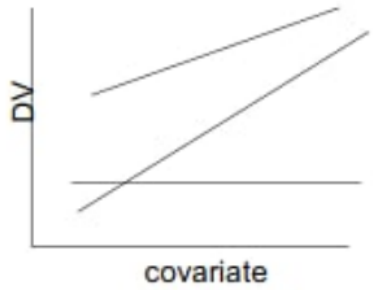
ANOVA variations

Assumption: relationship between DV and covariate is equal across treatment groups - **homogeneity of regression slopes**



homogeneity of regression slopes

☺



heterogeneity of regression slopes

☹

Homogeneity of regression slopes is important because adjustments to treatment means are based upon an average within-cell regression coefficient

Multiple hypothesis testing

The p value obtained from ANOVA analysis is significant ($p < 0.05$), we conclude that treatment differences are statistically significant

Problem:

ANOVA does not tell which treatments are significantly different from each other.

Multiple hypothesis testing

The p value obtained from ANOVA analysis is significant ($p < 0.05$), we conclude that treatment differences are statistically significant

Problem:

ANOVA does not tell which treatments are significantly different from each other.

Solution: To know the pairs of significant different treatments, we will can perform **multiple pairwise comparison** (post hoc comparison) analysis

Note: When the ANOVA is significant, post hoc tests are used to see differences between specific groups.

Post hoc tests should control the family-wise error rate (inflated type I error rate) due to multiple comparisons

Post hoc tests adjust the p-values (e.g. Bonferroni correction) or critical value (e.g. Tukey's HSD test).

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Bad idea:

To test each feature using hypothesis testing separately with some level of significance $\alpha = 0.05$.

Why? Let's calculate the probability of one significant result just due to chance?

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Bad idea:

To test each feature using hypothesis testing separately with some level of significance $\alpha = 0.05$.

Why? Let's calculate the probability of one significant result just due to chance?

$P($ at least one significant result $) = 1 - P($ no significant results $)$

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Bad idea:

To test each feature using hypothesis testing separately with some level of significance $\alpha = 0.05$.

Why? Let's calculate the probability of one significant result just due to chance?

$P($ at least one significant result $) = 1 - P($ no significant results $)$

$P($ at least one significant result $) = 1 - (1 - 0.05)^{20} \approx 0.64$

Multiple hypothesis testing

Example:

There are 20 features you are interested in as independent (predictor) features to create your machine learning model.

We want to select which features are useful for our prediction model

Null Hypothesis (H0): There is no relationship between the variables

Alternative Hypothesis (H1): There is a relationship between variables

Bad idea:

To test each feature using hypothesis testing separately with some level of significance $\alpha = 0.05$.

Why? Let's calculate the probability of one significant result just due to chance?

$P($ at least one significant result $) = 1 - P($ no significant results $)$

$P($ at least one significant result $) = 1 - (1 - 0.05)^{20} \approx 0.64$

With 20 hypotheses were made, there is around a 64% chance that at least one hypothesis testing result is significant, even if all the tests are actually not significant.

With a higher number of features to consider, the chance would even higher.

That is why there are methods developed for dealing with *multiple testing error.*

Multiple hypothesis testing

This method is called the **multiple testing correction**.

Number of errors committed when testing m null hypotheses

	Declared non-significant	Declared signiflicant	Total
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m -$ **R**	**R**	m

What was actually corrected?

Multiple hypothesis testing

This method is called the **multiple testing correction**.

Number of errors committed when testing m null hypotheses

	Declared non-significant	Declared significant	Total
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m -$ **R**	**R**	m

What was actually corrected?

(1) Controlling the Family-wise error rate (FWER)

FWER is the probability of making at least one false discoveries (type I errors)

$$\text{FWER} = \Pr(V \geq 1) = 1 - \Pr(V = 0)$$

Thus, by assuring FWER $\leq \alpha$, the probability of making one or more type I errors in the family is controlled at level α.

(2) Controlling the False Discovery Rate (FDR) = Type I error/False Positive Error.

Multiple hypothesis testing

1) Bonferroni correction – simplest yet the strictest method, controls FWER

α is divide it with the number of the testing/number of the hypothesis for each hypothesis.

$$\alpha_{Bon} = \alpha/m$$

Multiple hypothesis testing

1) Bonferroni correction – simplest yet the strictest method, controls FWER

α is divide it with the number of the testing/number of the hypothesis for each hypothesis.

$$\alpha_{Bon} = \alpha/m$$

Example:

Let's assume we have 10 features.

Normally, when we get the P-value < 0.05, we might see a significant result due to a chance.

In our case if we have 20 hypothesis testing.

$$\alpha_{Bon} = \alpha/m = 0.05/20 = 0.0025$$

Hence

$P($ at least one significant result $) = 1 - (1 - 0.0025)^{20} \approx 0.049$

Multiple hypothesis testing

1) Bonferroni correction – simplest yet the strictest method, controls FWER

α is divide it with the number of the testing/number of the hypothesis for each hypothesis.

$$\alpha_{Bon} = \alpha/m$$

Example:

Let's assume we have 10 features.

Normally, when we get the P-value < 0.05, we might see a significant result due to a chance.

In our case if we have 20 hypothesis testing.

$$\alpha_{Bon} = \alpha/m = 0.05/20 = 0.0025$$

Hence

$P($ at least one significant result $) = 1 - (1 - 0.0025)^{20} \approx 0.049$

Note:

Bonferroni Correction is proven too strict at correcting the α level where Type II error/False Negative rate is higher than what it should be.

Multiple hypothesis testing

2) Holm-Bonferroni correction method – less strict, controls FWER

The α level correction is not uniform for each hypothesis testing; instead, it was varied depending on the P-value ranking.

By ranking, it means a P-value of the hypothesis testing we had from lowest to highest.

Feature	P-Value
Feature #4 – Rank 1	0.001
Feature #3	0.003
Feature #1	0.01
Feature #8	0.0134
Feature #7	0.02
Feature #10	0.025
Feature #9	0.044
Feature #2	0.067
Feature #6	0.33
Feature #5 – Rank 10	0.5

(1)

Let's try to rank our previous hypothesis from the P-value we have before. The rank should look like this.

Multiple hypothesis testing

After we rank the P-value, we would the correct α level and test the individual hypothesis using this equation below.

$$P_k < \frac{\alpha}{m + 1 - k}$$

Where k is the ranking and m is the number of hypotheses tested.

Example, we test rank 1:

$$P_1 < \frac{0.05}{10+1-1}$$
$$P_1 < 0.005$$

Example, we test rank 10:

$$P_1 < \frac{0.05}{10+1-10}$$
$$P_1 < 0.05$$

3) Shidak correction – controls FWER

$P(V \leq 1) = 1 - P(V = 0) \leq 1 - (1 - \alpha_1)^m = \alpha$, where α is the significance level we set for the family hypotheses and α_1- the desired significance level for testing each single hypothesis.

Let's express α_1 in terms of α and get $\alpha_1 = 1 - (1 - \alpha)^{1/m}$ |

Multiple hypothesis testing

3) Shidak correction – controls FWER

$P(V \leq 1) = 1 - P(V = 0) \leq 1 - (1 - \alpha_1)^m = \alpha$, where α is the significance level we set for the family hypotheses and α_1- the desired significance level for testing each single hypothesis.

Let's express α_1 in terms of α and get $\alpha_1 = 1 - (1 - \alpha)^{1/m}$ |

4) Shidak-Holm method – controls FWER

Iterative adjustment. Similarly, we sort our p-values in ascending order and correct them according to the Shidak correction:

$\alpha_1 = 1 - (1 - \alpha)^{\frac{\pi}{m}}$

$$\alpha_i = 1 - (1 - \alpha)^{\frac{\alpha}{w - l + 1}}$$
$$\dots$$
$$\alpha_m = \alpha$$

'=

Multiple hypothesis testing

Has several properties:

1. Controls FWER at the α significance level if the statistics are collectively independent.

2. If the statistics are collectively independent, it is impossible to construct a procedure that controls FWER at the α level and is more powerful than the Shidak-Holm method.

3. For large m it differs little from the Holm method

Multiple hypothesis testing

Has several properties:

1. Controls FWER at the α significance level if the statistics are collectively independent.

2. If the statistics are collectively independent, it is impossible to construct a procedure that controls FWER at the α level and is more powerful than the Shidak-Holm method.

3. For large m it differs little from the Holm method

Note:

- Without additional assumptions it is impossible to construct a more powerful procedure than Holm's method

- Given the independence of the experiments, it is impossible to construct a more powerful procedure than the Shidak-Holm method

But you can create a powerful procedure for FDR - and, as practice shows.

Multiple hypothesis testing

5) Benjamini-Hochberg method – controls FDR

Recall, FDR (False Discovery Rate) is the average proportion of falsely rejected $\boldsymbol{H_0}$ among all rejected

$$H_0, FDR = E\left(\frac{V}{R}\middle| R > 0\right)$$

- when considering FWER, we were concerned about the probability that at least one null hypothesis would be falsely rejected

- when considering FDR, we lower the bar and assume that there will be several such hypotheses - but no more than α.

Note that $FDR \leq FWER$

Benjamini-Hochberg (BH) method or often called the BH Step-up procedure:

First, ranks the P-value from the lowest to the highest.

The hypothesis is then compared to the α level by the following equation.

$$P_k < \frac{k}{m}\alpha$$

- where k is the rank and m is the number of the hypotheses.

Summary

Summary:

1) ANOVA compares

- within-group variability: the variance of the individual observations within a group, and

- between-group variability: the variance between the averages of the groups.

2) ANOVA variations:

- One-way ANOVA vs Two-way ANOVA

- ANOVA vs MANOVA

- Repeated measures ANOVA

- ANCOVA

3) ANOVA does not tell which treatments are significantly different from each other $=>$ **multiple pairwise comparison**

4) Post hoc tests should control the family-wise error rate (inflated type I error rate) or False discovery rate· due to multiple comparisons