Chapter 2: Linear Regression
Part 3: t-test and ANOVA

MATH 60604: Statistical Modelling

HEC Montréal
Department of decision sciences

Overview of course material

Type of variable Y	Independent Observations	Method
Continuous	Yes	**Simple linear regression (chap 2 part 1)** **Multiple linear regression (chap 2 part 2)** **Special cases: t-test and ANOVA (chap 2 part 3)** **Models for survival data (chap 6)**
Continuous	No (ex : longitudinal study)	**Regression with random effects (chap 5)**
Binary	Yes	**Logistic Regression (chap 4)**
Count	Yes	**Poisson Regression (chap 4)**

Table of contents

The t-test

- Generally speaking, there are two types of t-test, which allow to test different types of hypotheses:
  - One population t-test which tests the mean μ of a population, ex:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

  - Two population t-test which tests equality of two means μ_1 and μ_2:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

- These two t-tests are <u>not</u> identical and answer different questions.
- The one-population t-test was reviewed in chapter 1 (where we focused on the case of paired-samples).
- The focus in here is the t-test for two populations

Motivating example

Motivating example:

- **Study question**: Does paying by credit card impact consumers' spending?
- **Context**: In potentially expensive purchases, the amount that consumers are willing to spend when paying by credit card could be different (higher?) than that of consumers who pay by cash.
- **Objectives**: Present new evidence in favour of the proposition that individuals have different spending habits when paying by credit card.

Reference

Prelec, D. et Simester, D. (2001). Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay. Marketing Letters 12, 5-12.

Motivating example

- The product being sold: two tickets for the last regular season game of the Boston Celtics (NBA basketball team). This game was very important, since it would decide who would end up in first place in the division.
- 64 subjects were randomly placed into the two groups:
  - Group 1 (33 subjects): must pay by cash
  - Group 2 (31 subjects): must pay by credit card
- All the subjects had to fill out a questionnaire asking how much they were willing to pay for the two tickets.

Randomization

- **Randomization**: the *random* assignment of subjects into different experimental groups.
- **Idea**: we want to have groups that are as similar as possible with regards to characteristics that could potentially influence the response variable, thereby diminishing or distorting their impact.
  - Ex: it's possible that age and sex, among other variables, could influence the amount of money someone is willing to spend.
  - Randomization tries to minimize the effects of other (confounding) variables by creating balanced groups.
    - The distribution of all other variables should be comparable in the two groups.
  - Any differences between the two groups can then be attributed to the treatment (here, paying by credit card vs. cash).
- Note that another way of accounting for other variables: directly control for the effects of these variables using a regression model

Motivating example

Example: cash vs. credit

Comparison of two groups: context and notation

- We want to compare the means of two populations
  - We want to see whether the two means μ_1 and μ_2 are significantly different.
  - We also want to quantify the difference between these two means.

Notation

- Group 1:
  - population mean μ_1
  - observations $(Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{n_1}^{(1)})$ (sample size n_1)
- Group 2:
  - population mean μ_2
  - observations $(Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_{n_2}^{(2)})$ (sample size n_2)

Comparison of two groups: hypothesis testing

Hypothesis to test

We want to test the following hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

or equivalently:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

- We're thus interested in estimating the difference $\mu_1 - \mu_2$, along with a confidence interval.
- The two sample t-test allows us to test this.

Comparing 2 groups: t-test

Notations:

- Let

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^{(1)} \quad \text{and} \quad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_i^{(1)} - \bar{Y}_1)^2$$

denote the sample mean and sample variance for group 1

- In the same way, let

$$\bar{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)} \quad \text{and} \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i^{(2)} - \bar{Y}_2)^2$$

denote the sample mean and sample variance for group 2

Comparing 2 groups: t-test

- There exists many different versions of the t-test (as we already saw), depending on the context of the problem and the underlying assumptions.
- If $\sigma_1^2 = \sigma_2^2$, the test statistic is

$$\boxed{T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

where S_p^2 is an estimator of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

this *pooled variance* S_p is simply a weighted average of the two sample variances S_1^2 and S_2^2

Comparing 2 groups: t-test

$$\boxed{T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

- Underlying assumptions necessary for the validity of this two sample t-test
  - equality of variance, i.e. $\sigma_1^2 = \sigma_2^2$
  - normality, i.e. $Y_i^{(j)} \sim \mathcal{N}$ for $j = 1, 2$
  - independence, i.e. the observations from the two groups are independent
- If these assumptions hold, it can be shown that under the null hypothesis $H_0 : \mu_1 = \mu_2$, the test statistic T will follow a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.
- The test will reject H_0 when the value of the observed test statistic is either too small or too large (i.e. whenever $|T|$ is large)
  - this will lead to a small p-value

Comparing 2 groups: confidence interval

- When we're interested in comparing two groups, it's often useful to also provide a confidence interval for the difference in the means $\mu_1 - \mu_2$
- This provides more information than simply a p-value as it allows to quantify and assess the magnitude of the difference
- A $100(1-\alpha)\%$ confidence interval for the difference $\mu_1 - \mu_2$ is given by

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

  where $t_{\nu,\alpha/2}$ is the $1-\alpha/2$ quantile from the Student t distribution with ν degrees of freedom

- Recall that we can also use a C.I. to make a conclusion:
  - if 0 is not contained in the C.I., then we can reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 \neq \mu_2$ (at the α level)

The t-test is a special case of regression

- Let's re-examine the credit card vs. cash example:
  - Goal: test whether the amount of money consumers are willing to spend is the same in the "cash" and the "credit card" groups.
- Define
  - Y: the offer amount in dollars (`offre`)
  - X: indicator variable for type of payment, where $X = 0$ for cash (i.e. `groupe=0`) and $X = 1$ for credit (i.e. `groupe=1`)
- Testing whether the two means are equal is equivalent to testing whether the effect of the binary variable X (`groupe`) on the response variable Y (`offre`) is zero
  - We can formulate the problem using a linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

testing $H_0 : \mu_{cash} = \mu_{card}$ is the same as testing $H_0 : \beta_1 = 0$

The t-test is a special case of regression

- The (simple) linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

implies that that:

$$\begin{aligned} E(Y|X=0) &= E(Y|\text{ cash}) \\ &= \beta_0 \\ &= \mu_1 \end{aligned}$$

$$\begin{aligned} E(Y|X=1) &= E(Y|\text{ credit}) \\ &= \beta_0 + \beta_1 \\ &= \mu_2 \end{aligned}$$

$\rightarrow$ it is clear that testing $H_0 : \mu_1 = \mu_2$ (i.e. the means are equal) is equivalent to testing $H_0 : \beta_1 = 0$ since $\beta_1 = \mu_2 - \mu_1$

Classic t-test

Example: cash vs. credit

Overview of t-test: two ways of thinking

- The t-test can be written in the classic form, or completely equivalently - in the form of a simple linear regression model
- These two equivalent formulations are presented in the table below:

	Regression model	**Classic t-test**
Context	Y=response variable X=group indicator Model $Y = \beta_0 + \beta_1 X + \epsilon$	$(Y_1^{(1)}, Y_{2,}^{(1)}, \ldots, Y_{n_1}^{(1)})$: obs group 1, of size n_1 and mean μ_1 $(Y_1^{(2)}, Y_{2,}^{(2)}, \ldots, Y_{n_2}^{(2)})$: obs group 2, of size n_2 and mean μ_2
Null hypothesis	$H_0 : \beta_1 = 0$	$H_0 : \mu_1 = \mu_2$
Test statistic	$t = \hat{\beta}_1 / \hat{se}(\hat{\beta}_1)$	$t = (\bar{Y}_1 - \bar{Y}_2) / \left(S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$ with $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$, pooled variance
Distribution of the test stat. under H_0	Student$(n-2)$	Student$(n-2)$ (where $n = n_1 + n_2$)

Overview of t-test: assumptions

- The results of the t-test are valid only if the underlying assumptions hold
- Below are the assumptions for the t-test in the regression context, as well as (equivalently) in the classic t-test form:

	Regression model	**Classic t-test**
Context	Y=Offer and X=Group Model $Y = \beta_0 + \beta_1 X + \epsilon$	$(Y_1^{(1)}, Y_{2,}^{(1)}, \ldots, Y_{n_1}^{(1)})$: obs group 1, of size n_1 and mean μ_1 $(Y_1^{(2)}, Y_{2,}^{(2)}, \ldots, Y_{n_2}^{(2)})$: obs group 2, of size n_2 and mean μ_2
Assumption #1	The ϵ_i's are independent ($\Leftrightarrow$ the observations are independent)	The observations between the two groups are independent, and both consist of random samples
Assumption #2	$E[\epsilon_i] = 0$ ($\Leftrightarrow E[Y_i \vert X = x_i] = \beta_0 + \beta_1 x_i$)	$E(Y_i^{(1)}) = \mu_1$ $E(Y_i^{(2)}) = \mu_2$
Assumption #3	$Var[\epsilon_i] = \sigma^2$ is constant for all i ($\Leftrightarrow$ the variance of the Y_i's is constant)	The variance of Y is the same between the two groups ($\sigma_1^2 = \sigma_2^2$)
Assumption #4	The ϵ_i's follow a normal distribution ($\Leftrightarrow Y_i \vert X_i$ follows a normal distribution)	The Y_i's follow a normal distribution in **each** of the two groups

Assumption: normality

- The normality assumption ensures that the test statistic T follows a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom:
  - this allows to calculate the p-value
  - also allows to calculate C.I.s
- If the Y_i's are not normally distributed, the results may not be valid...
- In large samples, this is not problematic:
  - Even if Y is not normally distributed in each population, the test and CIs will actually still be valid for large sample sizes (we can rely on asymptotic statistical results)
  - *A rule of thumb is often: $n_1 \geq 20$ and $n_2 \geq 20$. This actually depends on the symmetry of the distributions of Y in each group (the more symmetric, the smaller the sample size needs to be for the CIs to be valid). For very large sample sizes, inference will still be valid even for asymmetric distributions.*
- In the regression framework, the normality assumption can easily be verified through the model residuals.

Assumption: equality of/constant variance

- Another important assumption for the t-test is equality of variance between the two groups (constant variance in the regression model)
  - We have seen how to verify this assumption in the context of regression using residual diagnostics
  - In the context of a t-test (comparing the means from two groups), we can also formally test the assumption that $\sigma_1^2 = \sigma_2^2$

Assumption: equality of/constant variance

- To test equality of the variances of the two groups

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

  we can carry out an F-test

- The test statistic for the F-test is

$$T = \frac{S_1^2}{S_2^2}$$

- It can be shown that under H_0, the test statistic T follows an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of liberty

Classic t-test assumptions

Example: credit vs. cash

What to do when the variances are not equal?

- We will see that it is possible to slightly modify the classic t-test so that it is valid in the case where the variances are not equal:
  - This modified version of the t-test is called Welch's test.
- In the general context of linear regression, there exist other alternatives to classic linear regression that address the fact that the variance is not constant:
  - ex: variance-stabilizing transformations
  - ex: weighted least squares
  - we won't directly cover these alternative approaches here...

Classic t-test
Welch test
Remarks for t-test
ANOVA
One-way ANOVA
Two-way ANOVA
main effects
interactions

Table of contents

T-test: unequal variances in the two groups

- The classical t-test can be modified to allow for different variances in the two groups
  - The modified version of the t-test is called Welch's test.
  - The difference between the classic t-test and Welch's test is the way in which the variance is calculated in the test statistic.
- For $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$, the test statistic for Welch's test is

$$\boxed{T = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}}$$

- Under H_0, T follows a Student t distribution with m degrees of freedom, where

$$m \approx \frac{(w_1 + w_2)^2}{\left\{\frac{w_1^2}{(n_1-1)} + \frac{w_2^2}{(n_2-1)}\right\}}, \quad \text{for } w_i = S_i^2/n_i,\ i = 1, 2$$

  - this approximation for the degrees of freedom is called the Satterthwaite approximation

Comparing 2 groups: confidence interval

- When the variances of the two groups are not equal, the formula for the C.I. is a little different:
  - A $100(1-\alpha)\%$ confidence interval for the difference $\mu_1 - \mu_2$ is given by

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{m,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

  where $t_{m,\alpha/2}$ is the $1-\alpha/2$ quantile from the Student t distribution with m degrees of freedom (where m is obtained from the Satterthwaite approximation)

Welch's test

Example: credit vs. cash

Table of contents

Discussion: other tests

- The t-test (both the classical and Welch's) relies on the assumption of normality, as is the case in linear regression.
- If the sample size is large, this will generally not be problematic...
- Note that there is another test that can be used to compare independent population means: the Mann-Whitney-Wilcoxon (MWW) test.
  - The MWW test is non-parametric and is based on *ranks*
  - It can thus be used for non-normal data; and it is also robust to extreme values.
  - We will not cover this test here.

Discussion: one-sided tests

- In the credit card example, we tested the hypothesis:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

$\rightarrow$ this is a two-sided test

- Sometimes the research question requires hypothesis specifying the direction of the difference in means:
  - i.e. $H_1 : \mu_1 < \mu_2$,
  - or $H_1 : \mu_1 > \mu_2$
  - $\rightarrow$ these are both one-sided tests

Discussion: one-sided tests

- Suppose we're interested in testing

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 > \mu_2$$

- The t-test (classic or Welch) is based on the same test statistic as before:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{or} \quad \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- If the value of T is large, we'll be inclined to believe that H_1 is true, but if T is small, H_0 will seem more likely.

- As we've already seen, we use a p-value to be able to formally conclude.

Discussion: one-sided tests

- The calculation of the p-value depends on the form of H_1
- Suppose that based on the sample data, we observe a value of the test statistic $T = T_{obs}$
  - If $H_1 : \mu_1 < \mu_2$ then the p-value is the probability $P(T < T_{obs} | H_0)$
  - If $H_1 : \mu_1 > \mu_2$ then the p-value is the probability $P(T > T_{obs} | H_0)$
  - Recall: under H_0, T follows a Student t distribution, so we'll calculate the p-values using this distribution (with degrees of freedom as defined by either the classic t-test or Welch's test).
- **Careful:** the hypotheses should be specified before the actual research process begins, even before collecting the data.
- You should never look at the data and then decide what H_1 sould be!
  - In doing this, the true level of the test and the p-value will actually be distorted!

Discussion: significance vs. practical importance

- In general, the power of a test increases with the sample size.
  - *Recall: the power of a test is defined as*

$$1 - \beta = 1 - P(\text{fail to reject } H_0 | H_0 \text{ false})$$

- This means that even a tiny deviation from H_0 will be detected if the sample sizes are large enough.

Very important:

It is the responsibility of the researcher to distinguish between statistical significance (i.e. rejecting H_0) and the practical importance of what has been detected.

Discussion: significance vs. practical importance

- In our example, we found that those who pay by credit card are willing to spend \$15 more than those who pay by cash.
- This amount could possibly be judged as an appreciable difference given the mean amount offered in each group (\$56.6 and \$71.6).
  - This is a difference of 27% !
- If, for example, we had observed a difference of \$1, corresponding to group means of \$200 and \$201, the difference would only be 0.5%. This would likely not be considered as an important difference in practice, even if the difference was statistically significant.
- In a research paper, you should not only report the results of the tests with p-values, but you should also allow the readers to see the size of the detected effects
- In the case of a difference in means, it is important to report the values $\bar{Y}_1$, $\bar{Y}_2$, S_1^2, S_2^2 as well as the sample sizes
- It is also important to provide the CI for the difference $\mu_1 - \mu_2$

Review of the t-test

- The t-test allows testing equality between two means in two populations
  - can be performed using a classical t-test framework
  - can be thought of as a special case of simple linear regression
- In both cases, you need to pay attention to the assumptions of the test/model:
  - Normality of observations in each group
  - Equality of variances between the groups (i.e. constant in the two groups)
- When the variances of the two groups are NOT equal we must use Welch's test

Table of contents

T-test and ANOVA in regression framework

- Simple linear regression (one predictor variable in the model): tests the effect of a variable X on a continuous variable Y
  - **Special case**: when X is binary, simple linear regression is equivalent to testing the means in two populations.
  - This is a t-test
- Multiple linear regression (several predictor variables in the model): tests the joint effect of X on the continuous variable Y
  - **Special case**: when X is categorical (with k categories), we have the equivalent of testing k means in k populations.
  - This is an ANOVA.

Motivating example

- **Study objective**: compare the *perceived* quality of service vs. the *expectations* regarding quality of service.
- The quality of service is measured through a tool called SERVQUAL, which is more often called RATER
  - Studies different factors offered by businesses that are linked to quality of service.
  - The original tool measured 10 factors but the more recent version, RATER, only measures 5.

Reference

Bose, S. et Gupta, N. (2013). Customer Perception of Services Based on the SERVQUAL Dimensions : A Study of Indian Commercial Banks. Services Marketing Quartely 34, 49-66.

Motivating example

Factors in the RATER tool:

- Reliability: the ability to perform the promised service dependably and accurately
- Assurance: the knowledge and courtesy of employees and their ability to convey trust and confidence
- Tangibles: the appearance of physical facilities, equipment, personnel and communication materials
- Empathy: the provision of caring, individualized attention to customers
- Responsiveness: the willingness to help customers and to provide prompt service.

Source : Wikipedia (https://en.wikipedia.org/wiki/SERVQUAL) accessed June 22, 2015.

Motivating example

- Customers for three public sector banks were included in the study:
  - United Bank of India
  - State Bank of India
  - Allahabad Bank
- For this example, we will focus on the “reliability” score
  - The higher the score, the more the customer perceives the bank as reliable.
- We want to test whether the mean reliability score for the three banks are significantly different from one another.

Motivating example

Example: reliability score of banks

Table of contents

Introduction and terminology

- The goal of a one-way ANOVA is to compare the means for k populations
- Bank example: we want to compare the quality of service between 3 banks
  - Here, the factor is the bank.
  - There are three different banks, so we say that the factor "bank" has 3 levels
- We will perform a one-way ANOVA on a factor with 3 levels.

Notation

- The general one-way ANOVA framework allows to compare the means of k populations

Notation

- Group 1:
  - population mean μ_1 and variance σ_1^2
  - observations $(Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{n_1}^{(1)})$ (sample size n_1)
- Group 2:
  - population μ_2 and variance σ_2^2
  - observations $(Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_{n_2}^{(2)})$ (sample size n_2)
- ...
- Group k:
  - population mean μ_k and variance σ_k^2
  - observations $(Y_1^{(k)}, Y_2^{(k)}, \ldots, Y_{n_k}^{(k)})$ (sample size n_k)

Comparing k groups: hypothesis testing

- We want to test whether the k means $\mu_1, \mu_2, \ldots, \mu_k$ are equal and we want to measure the size of the differences between these means.
- This is a generalization of the problem of comparing 2 groups, as seen in the context of the t-test.
  - the t-test tests whether the means of two groups are equal
  - the ANOVA tests whether or not the means of k groups are equal

ANOVA

Hypothesis to test:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

H_1 : at least two means are different from one another

ANOVA is a special case of linear regression

- In the bank example, we're interested in investigating whether quality of service (as measured through the reliability score) is the same among the 3 banks
  - $\rightarrow$ we're interested in comparing the means of 3 groups, i.e. we're interested in testing

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 \\ H_1 &: \text{at least 2 means are different from one another} \end{aligned}$$

    where μ_j is the mean reliability score for bank j, $j = 1, 2, 3$ (1=United Bank of India, 2=State Bank of India, 3= Allahabad Bank)

- There are different ways to test this hypothesis
  - ANOVA
  - multiple linear regression

ANOVA is a special case of linear regression

- Testing equality of means across k groups is equivalent to carrying out a global test in a multiple linear regression model that has a single categorical explanatory variable with k levels
- Suppose X is a categorical variable with k levels labelled as $1, 2, \dots, k$
  - let X_1 denote the indicator variable for level 1, i.e. $X_1 = 1$ if $X = 1$ and $X_1 = 0$ otherwise
  - let X_2 denote the indicator variable for level 2
  - ...
  - let X_{k-1} denote the indicator variable for level $k-1$
- The linear regression model which includes only the categorical variable X can then be specified as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \epsilon$$

ANOVA is a special case of linear regression

- From the model we have that

$$\begin{aligned}
\mu_1 &= \mathrm{E}(Y|X=1) = \mathrm{E}(Y|X_1=1, X_2=0, \ldots, X_{k-1}=0) \\
&= \beta_0 + \beta_1 \\
\mu_2 &= \mathrm{E}(Y|X=2) = \mathrm{E}(Y|X_1=0, X_2=1, \ldots, X_{k-1}=0) \\
&= \beta_0 + \beta_2 \\
&\vdots \\
\mu_{k-1} &= \mathrm{E}(Y|X=k-1) = \mathrm{E}(Y|X_1=0, X_2=0, \ldots, X_{k-1}=1) \\
&= \beta_0 + \beta_{k-1} \\
\mu_k &= \mathrm{E}(Y|X=k) = \mathrm{E}(Y|X_1=0, X_2=0, \ldots, X_{k-1}=0) \\
&= \beta_0
\end{aligned}$$

- Thus

$$\mu_1 = \mu_2 = \cdots = \mu_k \quad \Leftrightarrow \quad \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

ANOVA is a special case of linear regression

- Thus testing equality of the means across all groups

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

is equivalent to testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

which is a global test in the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \epsilon$$

ANOVA is a special case of linear regression

- Back to our example, testing equality between the mean score for the 3 banks is equivalent to testing the effect of the bank on the score in a multiple regression model that includes only the categorical explanatory variable bank.

- We could formulate this problem through the following linear regression model:

$$Y = \beta_0 + \beta_1 X_{bank1} + \beta_2 X_{bank2} + \epsilon$$

  where Y denotes the reliability score and the variables X_{bank1} and X_{bank2} are the binary indicator variables for banks 1 and 2, respectively.

- Testing

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

  is equivalent to testing

$$H_0 : \beta_1 = \beta_2 = 0$$

ANOVA: special case of regression

- If we assume the model:

$$Y = \beta_0 + \beta_1 X_{bank1} + \beta_2 X_{bank2} + \epsilon$$

- Then we have:

$$\begin{aligned}
E[Y|X_{bank1} = 1, X_{bank2} = 0] &= E[Y|\text{ Banque=1 }] \\
&= \mu_1 \\
&= \beta_0 + \beta_1 \\
E[Y|X_{bank1} = 0, X_{bank2} = 1] &= E[Y|\text{ Banque=2 }] \\
&= \mu_2 \\
&= \beta_0 + \beta_2 \\
E[Y|X_{bank1} = 0, X_{bank2} = 0] &= E[Y|\text{ Banque=3 }] \\
&= \mu_3 \\
&= \beta_0
\end{aligned}$$

- We see that testing $H_0 : \mu_1 = \mu_2 = \mu_3$ (the 3 means are equal) is the same as testing $H_0 : \beta_1 = \beta_2 = 0$
- This is just an F-test (global) of the effect of the bank variable

One-way ANOVA

Example: reliability score of banks

ANOVA: assumptions

- The results of the ANOVA are only valid if we can the underlying model assumptions are met.
- Below we see the assumptions for the ANOVA in the context of regression, as well as in the classical ANOVA form:

	Regression	**Classic ANOVA**
Context	Y, $\{X_1, \ldots, X_{k-1}\}$ = binary variables (for a categorical variable with k levels) Regression Model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \epsilon$	$(Y_1^{(1)}, \ldots, Y_{n_1}^{(1)})$: obs group 1, size n_1, mean μ_1 ... $(Y_1^{(k)}, \ldots, Y_{n_k}^{(k)})$: obs group k, size n_k, mean μ_k
Assumption #1	The ϵ_i's are independent ($\Leftrightarrow$ the obs. are indep.)	The observations in the k groups are independent
Assumption #2	$E[\epsilon_i] = 0$ ($\Leftrightarrow E[Y\|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}$)	$E(Y_i^{(j)}) = \mu_j, j = 1, \ldots, k$
Assumption #3	$Var[\epsilon_i]$ is constant for all i ($\Leftrightarrow$ the variance of Y_i is constant)	The variance for Y is the same in the k groups $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
Assumption #4	The ϵ_i's follow a normal dist. ($\Leftrightarrow [Y_i]$ given X_i follows a normal dist.)	The Y_i follow a normal dist. in **each** of the k groups

Normality assumption

- In the context of regression, the normality assumption can be easily verified through examination of the residuals (see parts 1 and 2)
- In the context of ANOVA, you can simply look at the histogram and qq-plot for each group to assess normality
- Note: if the variable Y is not normally distributed within each population (group), the tests and CIs are still valid if the sample sizes are large
  - (minimum 20 observations per group)

Effects of non-normality in small sample sizes

- As for the t-test (corresponding to $K = 2$ groups), if the distribution of Y is not normal in one or more of the populations, the F-test and Welch's test will not be affected if the sample sizes are sufficiently large
  - The reference distribution for the test (Fisher distribution) is still approximately valid when the sample size is large.
- However, when the distributions are non-normal, the F-test and Welch's are not necessarily the best choice.
- An option in this case is the Kruskal-Wallis (KW) test, which is a non-parametric test based on the ranks (and thus does not rely on the assumption of normality). We will not cover this.

Constant variance assumption

- In the context of regression, the assumption of constant variance in the 3 groups can be verified through the model's residual plots.
- A formal test for equality of variances can also be carried out
  - the underlying hypotheses for testing the equality of the variances across the different groups are

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$
$$H_1 : \text{at least 2 of these variances differ}$$

  - there exists different tests to do this, e.g., Bartlett's test, Levene's test
  - Note: Levene's test is less sensitive to the assumption of normality in comparison to Bartlett's test. There are also different versions of Levene's test depending on what is used to center the observations: the mean, the median, the trimmed mean.
- When the variances are not constant, a generalization of Welch's test can be used for comparisons of multiple groups.

Example: reliability score of banks

Pairwise comparisons

- In a one-way ANOVA, rejecting $H_0 : \mu_1 = \cdots = \mu_K$ with the F-test (or Welch's test) only tells us that the population means are not all equal; i.e. that at least two means are different.
  - We don't get any additional information. We might wonder which particular groups have differing means.
  - The F-test is only the first step in an analysis.
- Sometimes, we're interested in specific comparisons, or we may want to make all possible pairwise comparisons of μ_i and μ_j.

Pairwise comparisons

- For example, we could be interested in testing:

$$H_0 : \mu_i = \mu_j \text{ vs. } H_1 : \mu_i \neq \mu_j$$

  for all possible combinations (i, j) and obtain CIs for each of the differences $\mu_i - \mu_j$
- With K groups, there are $K(K-1)/2$ possible comparisons.
  - This means we're really performing $K(K-1)/2$ tests.
- Each comparison between 2 groups is essentially a t-test to compare two groups
- This can be done for each individual test using a t-test, or by refitting the linear regression model changing the reference level.
- There are also ways to carry out all pairwise comparisons automatically directly in R.

One-way ANOVA

Example: reliability score of banks

Review of one-way ANOVA

- ANOVA is equivalent to a linear regression model with a categorical predictor variable
- The F-test allows us to test whether the categorical variable X has a global effect on Y
- We saw a test allowing us to verify the constant variance assumption for the k groups.
- We also saw a variation of the F-test to use when the variances are not equal: a generalization of Welch's test

Review of one-way ANOVA

- In statistics, you sometimes see the term “ANCOVA”, which is in fact an ANOVA test where we can adjust for other variables in the model.
  - This is actually just a regression with a categorical variable (for the ANOVA) where other predictors X are added to the model.
- When the factor variable has only 2 groups, the one-way ANOVA is equivalent to the t-test.

Review of one-way ANOVA

- Test the global effect of a factor:
  - **Parametric test**:
    - Classic F test (equal variances)
    - Welch's F test (unequal variances)
  - We can formally test the equality of variances between groups

- Perform all pairwise comparisons:
  - This can be done through individual t-tests
  - We can also directly perform these t-tests in the ANOVA in R:
    - The classical set-up: parametric t-tests, equal variances
    - The generalized Welch's set-up: parametric Welch tests, unequal variances

Table of contents

Motivating example

We'll look at the two-way ANOVA through an example...

- In a university, 120 participants tried a new course registration system. They were actually using a fake registration system where the two factors of interest could be manipulated: the stage of advancement in the registration process and the type of delay.
- The aim of the study was to assess the effect of the type of delay and where in the process the delay occurred (stage of advancement) on the evaluation of the service and on the perceived waiting time.

Reference: *Hui, M. K., Thakor, M. V. et Gill, R. (1998). The Effects of Delay Type and Service Stage on Consumers' Reaction to Waiting. Journal of Consumer Research 24, 469-479.*

Motivating example

- The two levels for stage of advancement were: “close to end” and “far from end”.
  - For the stage “far from end”, a message indicating a delay appears immediately at the beginning of the registration process.
  - For the stage “close to end”, the delay message appears after the participant has entered all important information (personal information and course choices).
- The three types of delays were : “procedural”, “correctional”, and “unknown”.
  - For a procedural type delay, the delay message indicates that the server is currently busy.
  - For the correctional type delay, the delay message indicates that there are problems with the terminal and that the system needs to re-establish the connection.
  - For the unknown type delay, the delay message simply states “please wait”.

Motivating example

- At the end of the registration process, the participants were asked to estimate the time delay (in minutes) lost during the registration process. They were also asked to provide an evaluation of the service using two measurement scales.
- If we consider all possible combinations of the two factors, there are 6 groups.
  - The 120 participants were randomly assigned to one of these 6 conditions (20 per cell).

    However, 9 of the 120 participants were removed, as they were not able to specify the type of delay that occurred. So, there were not actually 20 observations in each cell in the final dataset.

Study objective

- Evaluate the effect of the factors type of delay and stage of advancement on the evaluation of service

Two-way ANOVA

Example: user experience

Two-way ANOVA

- In a two-way ANOVA, in general, we consider two factors A and B.
  - the factor A has a levels,
  - the factor B has b levels
  - $\rightarrow$ there are ab groups in total
- We're interested in studying how the mean of Y is influenced by the two factors A and B.
- We can think of this schematically, where the (population) mean of the variable Y can be represented as follows:

	Factor B			
Factor A	level 1	level 2	...	level b
level 1	μ_{11}	μ_{12}	...	μ_{1b}
level 2	μ_{21}	μ_{22}	...	μ_{2b}
...	...	...	...	...
level a	μ_{a1}	μ_{a2}	...	μ_{ab}

where μ_{ij} is the (population) mean of Y for the group corresponding to level i of factor A and level j of factor B

Two-way ANOVA

- The objective of the ANOVA is to study
  - how and to what extent the levels of the factors A and B influence the mean of Y
  - how these two factors interact on the mean of Y.
- This is exactly equivalent to fitting a linear regression model with two categorical variables, along with an interaction term between them.

Two-way ANOVA - motivating example

- In the example, there are two factors:
  - stage (2 levels)
  - delay (3 levels)
- Let μ_{ij} denote the mean of Y (`eval`) for level i of `stade` and level j of `delay`:

	delay		
stade	level 1	level 2	level 3
level 1	μ_{11}	μ_{12}	μ_{13}
level 2	μ_{21}	μ_{22}	μ_{23}

- An ANOVA allows to study the effect of the stage of advancement and the type of delay on the mean μ_{ij} of the response variable (evaluation)

Main effects model

- We'll start off with an ANOVA model with additive effects (that is, a main effects only model, without an interaction term).
- Continuing with our motivating example, we're interested in assessing the effects of stage of advancement and type of delay on the evaluation (response variable).
- Define the following binary (indicator) variables:
  - X_{stage2}: indicator for level 2 of the factor `stage`
  - X_{delay2}, X_{delay3}: indicator for levels 2 and 3, respectively, of the factor `delay`
- The ANOVA with additive effects (or main effects) corresponds to the following linear regression model:

$$Y = \beta_0 + \beta_1 X_{stage2} + \beta_2 X_{delay2} + \beta_3 X_{delay3} + \epsilon$$

Main effects model

- The main effects (i.e. no interaction) model is

$$E(Y|X_{stage2}, X_{delay2}, X_{delay3}) = \beta_0 + \beta_1 X_{stage2} + \beta_2 X_{delay2} + \beta_3 X_{delay3}$$

thus the means μ_{ij} are:

$$\begin{aligned}
\mu_{11} &= E(Y|\text{stage=1, delay=1}) = \beta_0 \\
\mu_{12} &= E(Y|\text{stage=1, delay=2}) = \beta_0 + \beta_2 \\
\mu_{13} &= E(Y|\text{stage=1, delay=3}) = \beta_0 + \beta_3 \\
\mu_{21} &= E(Y|\text{stage=2, delay=1}) = \beta_0 + \beta_1 \\
\mu_{22} &= E(Y|\text{stage=2, delay=2}) = \beta_0 + \beta_1 + \beta_2 \\
\mu_{23} &= E(Y|\text{stage=2, delay=3}) = \beta_0 + \beta_1 + \beta_3
\end{aligned}$$

Main effects model

- In this model, the effect of `stage` is the same for all levels of `delay`:

$$\beta_1 = E(Y|\text{stage}{=}2,\ \text{delay}{=}1) - E(Y|\text{stage}{=}1,\ \text{delay}{=}1)$$
$$\beta_1 = E(Y|\text{stage}{=}2,\ \text{delay}{=}2) - E(Y|\text{stage}{=}1,\ \text{delay}{=}2)$$
$$\beta_1 = E(Y|\text{stage}{=}2,\ \text{delay}{=}3) - E(Y|\text{stage}{=}1,\ \text{delay}{=}3)$$

$\Rightarrow$ the effect of `stage` does not depend on the variable `delay`.

- And vice versa - in the same way, the effect of `delay` does not depend on the level of `stage`:

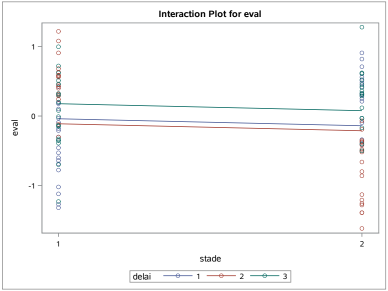
$$\beta_2 = E(Y|\text{stage}{=}1,\ \text{delay}{=}2) - E(Y|\text{stage}{=}1,\ \text{delay}{=}1)$$
$$\beta_2 = E(Y|\text{stage}{=}2,\ \text{delay}{=}2) - E(Y|\text{stage}{=}2,\ \text{delay}{=}1)$$

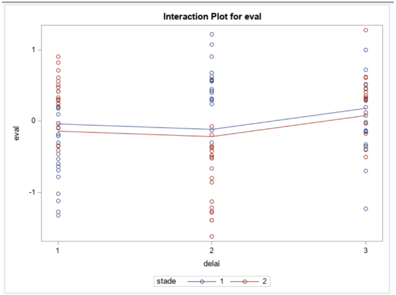
$$\beta_3 = E(Y|\text{stage}{=}1,\ \text{delay}{=}3) - E(Y|\text{stage}{=}1,\ \text{delay}{=}1)$$
$$\beta_3 = E(Y|\text{stage}{=}2,\ \text{delay}{=}3) - E(Y|\text{stage}{=}2,\ \text{delay}{=}1)$$

$$\beta_3 - \beta_2 = E(Y|\text{stage}{=}1,\ \text{delay}{=}3) - E(Y|\text{stage}{=}1,\ \text{delay}{=}2)$$
$$\beta_3 - \beta_2 = E(Y|\text{stage}{=}2,\ \text{delay}{=}3) - E(Y|\text{stage}{=}2,\ \text{delay}{=}2)$$

Main effects model

Visualisation:





Note: these plots were made in SAS

Main effects model

Example: user experience

Two-way ANOVA: interaction

- In the two-way ANOVA, the concept of interaction refers to the following question

  to what extent is the effect of a factor different with respect to the level of another factor (and vice-versa)?

- This is the same definition of interaction that we used in the context of multiple linear regression, and thus the procedure is the same....

Two-way ANOVA: interaction

- In the motivating example
  - the categorical variable `stage` is represented by a single indicator variable X_{stage2}
  - the categorical variable `delay` is represented by two indicator variables X_{delay2} and X_{delay3}
- Thus, if we want to include an interaction between the categorical variables `stage` and `delay`, we'll need to include two interaction terms:
  - $X_{stage2}X_{delay2}$
  - $X_{stage2}X_{delay3}$
- The linear regression model with an interaction is thus

$$
\begin{aligned}
Y = \beta_0 + \beta_1 X_{stage2} + \beta_2 X_{delay2} + \beta_3 X_{delay3} \\
+ \beta_4 X_{stage2} X_{delay2} + \beta_5 X_{stage2} X_{delay3} + \epsilon
\end{aligned}
$$

Two-way ANOVA: interaction

- The interaction model is

$$E(Y|X_{stage2}, X_{delay2}, X_{delay3} = \beta_0 + \beta_1 X_{stage2} + \beta_2 X_{delay2} + \beta_3 X_{delay3} + \beta_4 X_{stage2} X_{delay2} + \beta_5 X_{stage2} X_{delay3}$$

- The model implies that the (population) means μ_{ij} are:

$$\begin{aligned}
\mu_{11} &= E(Y|\text{stage=1, delay=1}) = \beta_0 \\
\mu_{12} &= E(Y|\text{stage=1, delay=2}) = \beta_0 + \beta_2 \\
\mu_{13} &= E(Y|\text{stage=1, delay=3}) = \beta_0 + \beta_3 \\
\mu_{21} &= E(Y|\text{stage=2, delay=1}) = \beta_0 + \beta_1 \\
\mu_{22} &= E(Y|\text{stage=2, delay=2}) = \beta_0 + \beta_1 + \beta_2 + \beta_4 \\
\mu_{23} &= E(Y|\text{stage=2, delay=3}) = \beta_0 + \beta_1 + \beta_3 + \beta_5
\end{aligned}$$

Two-way ANOVA: interaction

- The interaction between `stage` and `delay` causes the effect of `stage` on Y to depend on the level of `delay`:

$$\begin{aligned}
\beta_1 &= E(Y|\text{stage=2, delay=1}) - E(Y|\text{stage=1, delay=1}) \\
\beta_1 + \beta_4 &= E(Y|\text{stage=2, delay=2}) - E(Y|\text{stage=1, delay=2}) \\
\beta_1 + \beta_5 &= E(Y|\text{stage=2, delay=3}) - E(Y|\text{stage=1, delay=3})
\end{aligned}$$

- And vice-versa - the effect of `delay` on Y depends on the level of `stage`:

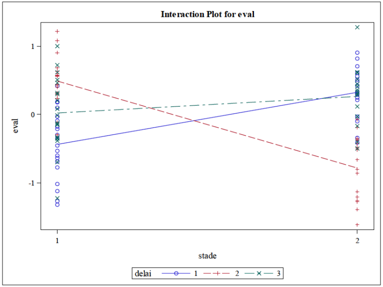
$$\begin{aligned}
\beta_2 &= E(Y|\text{stage=1, delay=2}) - E(Y|\text{stage=1, delay=1}) \\
\beta_2 + \beta_4 &= E(Y|\text{stage=2, delay=2}) - E(Y|\text{stage=2, delay=1})
\end{aligned}$$

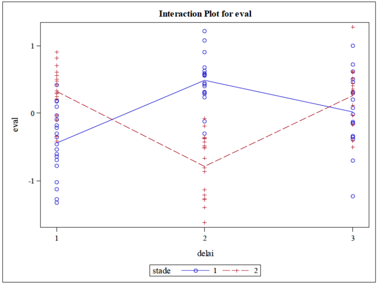
$$\begin{aligned}
\beta_3 &= E(Y|\text{stage=1, delay=3}) - E(Y|\text{stage=1, delay=1}) \\
\beta_3 + \beta_5 &= E(Y|\text{stage=2, delay=3}) - E(Y|\text{stage=2, delay=1})
\end{aligned}$$

$$\begin{aligned}
\beta_3 - \beta_2 &= E(Y|\text{stage=1, delay=3}) - E(Y|\text{stage=1, delay=2}) \\
\beta_3 + \beta_5 - \beta_2 - \beta_4 &= E(Y|\text{stage=2, delay=3}) - E(Y|\text{stage=2, delay=2})
\end{aligned}$$

Interaction model

Visualisation





Note: these plots were done in SAS

Interaction model

Example: user experience

2-way ANOVA: general steps to take

- If there's an interaction between the factors A and B, then testing the main effect of A (or B) doesn't say much, since the effect depends on the level of the other factor (by definition of interaction)
- When an interaction is significant and important (and thus the effect of one factor depends on the level of the other factor, and vice-versa), we'll be more interested in exploring the effect of one factor within each level of the other factor:
  - Ex: we can compare the means across all levels of factor A within a fixed level of factor B, say j. That is, compared $\mu_{1j}, \mu_{2j}, \ldots, \mu_{aj}$. And we would repeat this for each level of B, $j = 1, \ldots, b$ (and vice-versa).
  - The mean differences of a factor for a given level of another factor are called simple effects

	delay		
stage	level 1	level 2	level 3
level 1	μ_{11}	μ_{12}	μ_{13}
level 2	μ_{21}	μ_{22}	μ_{23}

2-way ANOVA: general steps to take

2-way ANOVA Steps

1. Test if there's an interaction between the factors A and B
2. If no interaction was detected in step 1, then we can test the global effects of A and B.
3. If an interaction was detected in step 1, then we can test the effect of A within each level of B - or vice versa. That is, we can compare the simple effects of the factors.

Interaction model: comparisons

Example: user experience

What to do when the variances differ between cells?

- Until now, the 2-way ANOVA analyses we've done have assumed equal variances among the cells in the table resulting from crossing the two factors, that is, assuming a equal variances across groups.
  - In the context of a regression model, this is the constant variance assumption.
- If we have reason to doubt this assumption, we can perform the analysis without making this assumption, in the same way we did for the t-test and the one-way ANOVA.
  - It's also possible to test equality of variances as before (Bartlett's test, Levene's test).

Interaction model: non-constant variance

Example: user experience

COMMON ERRORS ON EXAMS OR ASSIGNMENTS

You will lose points if you do the following...

Writing hypotheses

- When you carry out a test:
  - ALWAYS formally write out the underlying hypotheses, i.e. H_0 AND H_1
  - If you only write:

    $$H_0 : \mu_1 = \mu_2$$

    you will lose points if you do not specify what μ_1 and μ_2 represent!
  - You should write:

    $$H_0 : \mu_1 = \mu_2 \text{ (or another notation of your choice)}$$

    where μ_1 represents the mean of ... in group ... and μ_2 represents the mean of ... in group ...

Conclusion of a hypothesis test

- In the context of a test (whatever it may be)...
  - You should always make a conclusion and interpret the results
  - You will lose marks if you simply write: *we reject* H_0 (or the opposite)
  - You must write: we reject H_0, and so .... (interpretation in the form of a sentence)
  - Remember to say WHY you can make that conclusion (e.g. p-value $< \alpha$).

- Careful when applying the t-test and ANOVA - these rely on the assumption of equal variances (and normality). It's important to test the assumption of equality of variances.