STA302 - Lecture 2

Cedric Beaulac

May 9, 2019

Quick Review

Let's go slowly today

Statistical significance

- We compared two groups :
  - Group 1 : 48, 56, 58
  - Group 2 : 44, 46, 51
- To assess if the groups are truly different, we asked ourselves how exceptional the observed difference between means ($\bar{x}_1 - \bar{x}_2 = 7$) is.

Statistical significance

- Assuming the groups are the same we could have observed any permutations with equal probability.
- Assuming the groups are the same what we observed (difference of at least 7) has low probability.

P-value

- The p-value is the probability of a result as improbable as we've observed given the groups are the same.
- If the p-value is small, we either observed something that is unlikely OR the groups ARE NOT the same.

Hypothesis testing

The basics

- H_0 : The null hypothesis; We compute the probability of the observed event under the null.
- H_1 : The alternative; the event
- A statistical hypothesis testing is the evaluation of the compatibility of H_0 with the observed data.

Last lecture problem

- H_0 : The null hypothesis; Groups have no effect
- H_1 : The alternative: Group 1 has larger value than group 2
- We observed a p-value of 0.1, as this is unlikely we claim the assumption that H_0 is true is incorrect.

Last lecture problem

- H_0 : The null hypothesis;$\mu_A = \mu_B$
- H_1 : The alternative: $\mu_A \geq \mu_B$.
- We observed a p-value of 0.1, as this is unlikely we claim the assumption that H_0 is true is incorrect.

What is unlikely ?

- Defining what is unlikely is a difficult task
- This threshold is called *significance level*
- A good scientist defines a significance level ahead of time.
- Statistical significance makes NO SENSE without a significance level.
- Something is statistically significant with respect to a pre-defined significance level.

Significance level and type I error

- By definition, the p-value is the probability of the data given the H_0 is true.
- If this probability is small, we reject the null.
- It is fundamental to understand that if we fix signficance level to α we will be wrong to reject the null $\alpha\%$ of the time.
- If H_0 is true, the p-value $\sim U(0,1)$.

Significance level and type I error

- Rejecting H_0 when it is true is the type 1 error
- With signficance level α we expect to have to reject the null even though it is true with probability α.

Table of error types

Table of error types	H_0 is true	H_0 is False
Fails to reject	Correct inference	Type II error
Reject	Type I error	Correct inference

Table of error types

- Tradeoff between type I and type II error.
- What if we always reject of accept ?
- It's common to fix one and optimize for the other.

The two-sample t-test

Introduction

- Given a large data set, we have n_A observations coming from group A and n_B observations form group B, we can not use exact permutation test.
- Let us use some probability theory to establish some distributions.
- To use probability theory we will need some assumptions.
- Usually we MUST verify our assumptions, we will do that later!

The basics : Distribution Theory

- Assume the two samples are independant random samples from a normal distribution with means μ_A and μ_B with the same variance σ.
- To compare the two groups we will compare their empirical means $\bar{y}_A$ and $\bar{y}_B$.

$$\bar{y}_A - \bar{y}_B \sim N(\mu_A - \mu_B, \sigma^2(1/n_A + 1/n_B))$$

- Which implies:

$$\frac{(\bar{y}_A - \bar{y}_B) - (\mu_A - \mu_B)}{\sigma\sqrt{(1/n_A + 1/n_B)}} \sim N(0, 1)$$

The basics : Distribution Theory

- Since σ is unkown, we estimate it by s where

$$s^2 = \frac{\sum_{i=1}^{n_A}(y_{iA} - \bar{y}_A)^2 + \sum_{i=1}^{n_B}(y_{iB} - \bar{y}_B)^2}{n_A + n_B - 2}$$

- We no longer have a Normal distribution but a student distribution instead (Normal divided by independant chi-square):

$$\frac{(\bar{y}_A - \bar{y}_B) - (\mu_A - \mu_B)}{s\sqrt{(1/n_A + 1/n_B)}} \sim t_{n_A + n_B - 2}$$

The two-sample t-test : hypothesis testing

- H_0 The null : $\mu_a = \mu_b$.
- Under the null :

$$\frac{(\bar{y}_A - \bar{y}_B)}{s\sqrt{(1/n_A + 1/n_B)}} \sim t_{n_A+n_B-2}$$

The procedure

- The test goes as follow :
  - Decide on the significance level α
  - Check assumptions
  - Compute the test statistic $\frac{(\bar{y}_A - \bar{y}_B)}{s\sqrt{(1/n_A + 1/n_B)}}$
  - Establish the probability of such observation given that the null is true.
  - If the probability is below the significance level, this is evidence against the null.
- We will be wrong to reject the null with probability α.

The two-sample t-test

- Illustrate a simple way to use probability theory to perform statistical analysis.
- That is mostly what we will do for the next 5 weeks.
- We will do everything on R later today.

One-way Analysis of variance (ANOVA)

The basics

- Extension of the basic test when there is more than 2 groups.
- We already discussed the natural variability in the data, an ANOVA is an analysis of variance (ahahah)
- If groups are different we expect there is a bigger difference between groups (reflecting the group effect) than within groups (natural variability of the data).
- In order to stay consistent with the litterature, let's considers the groups represent different **treatments**.

The basics

- SST : Total Sum of Square $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- This is the unormalized sample variance, it represents the variable in the data.
- If we have T treatments, and n_t observations for treatment t the SST can be written as $\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t} - \bar{y})^2$
- This can be decompose this term to observe the variable between groups and within groups.

Sum of Squares decomposition

$$\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t}-\bar{y})^2 = \sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t}-\bar{y}_t)^2 + \sum_{t=1}^{T}\sum_{i=1}^{n_t}(\bar{y}_t-\bar{y})^2$$

- Proving the decompositon is left as an exercise (I've been dreaming to say this my whole life)
- Seriously do it! (hint : add $\bar{y}_t - \bar{y}_t$ inside the squarred error term)

Sum of Squares decomposition

- SSG $= \sum_{t=1}^{T} \sum_{i=1}^{n_t} (\bar{y_t} - \bar{y})^2 = \sum_{t=1}^{T} n_t (\bar{y_t} - \bar{y})^2$ is the sum of squares of groups/treatments (between groups sum of squares). (we assume n_t are equals $\forall t$)
- It is the sum of squarred distance between groups mean and the grand mean: also known as the explained variance.
- It is an unormarlized estimation of the between-groups variance.
- SSE $= \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y_t})^2$ is the sum of squares of error (within-groups sum of squares).
- It is the squarred prediction error for every observation if we use their group mean as predicted values: the unexplained variance.
- An unormalized estimation of the within-groups variance.

ANOVA

- We want to assess how large is SSG relatively to SSE. (IMPORTANT)
- We could look at SSG/SSE but it would be hard to establish a distribution for those things.
- We know a sum of squares devided by its degrees of freedom has a chi-square distribution.
- Thus MSG = SSG/(T-1) $\sim \chi^2_{T-1}$ and MSE = SSE/(n-T) $\sim \chi^2_{n-T}$ where $n \sum_t n_t$.
- MSG/MSE $\sim F_{T-1,n-T}$. (Really ? Why ?).
- There you go we're done!

ANOVA

- That was a bit too quick.
- Let us introduce our first *model* and prove things using distribution theory.

ANOVA

- The (effect) model :

$$y_{i,t} = \mu + \tau_t + \varepsilon_{i,t}$$

  where $\varepsilon \sim N(0, \sigma^2)$.
- μ is the global mean.
- τ_t is the effect of the tth treatment with $\sum_{t=1}^{T} \tau_t = 0$
- ε are errors reprenting the natural variability in real-life data.

Distribution Theory

- $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}$ where $Z_i \sim N(0,1)$
- If $X_i \sim N(\mu, \sigma)$ then $\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$
- $\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} + \frac{n(\bar{x}-\mu)^2}{\sigma^2}$
- It implies (not directly) that $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}$
- Ouff. . . ..

Distribution Theory

- Finally $z \sim \chi^2_{d_1}$ and $w \sim \chi^2_{d_2}$ then $\frac{z/d_1}{w/d_2} \sim F(d_1, d_2)$ it is known!
- Thus $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma_x^2(n-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/\sigma_y^2(n-1)} \sim F_{n-1,n-1}$
- And $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)} \sim F_{n-1,n-1}$ IF AND ONLY IF $\sigma_x^2 = \sigma_y^2$ (the null).
- YES!

Distribution Theory

- Here is the well know estimator for σ_x^2

$$\hat{\sigma_x}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- A ratior of estimator such $\hat{\sigma}_x^2/\hat{\sigma}_y^2$ has a F distribution if and only if $\sigma_x^2 = \sigma^2$:

$$\frac{\hat{\sigma_x}^2}{\hat{\sigma_y}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/n-1}{\sum_{i=1}^{n}(y_i - \bar{y})^2/n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma_x^2 n-1}{\sum_{i=1}^{n}(y_i - \bar{y})^2/\sigma_y^2 n-1} \sim F_{n-1,n-1}$$

ANOVA

- $n_t \sum_{t=1}^{T} (\bar{y}_t - \bar{y})^2 / (T - 1)$ is an estimation for the variation between groups (σ_T)
- SSE $= \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y_{i,t} - \bar{y}_t)^2 / (n - T)$ is an estimation for the variation within groups (σ_ϵ)
- Thus $\frac{SSG/(T-1)}{SSE/(n-T)} \sim F(T-1, n-T)$ if and only if the between-groups and within-groups variance are equal.
- Thus a small p-value indicates these variances are different, which is evidence for the existence of some group effect.

I’m exhausted

- Let’s take a break
- Let’s work with R when we are back!

Practice problem

- SST Decomposition
- Experimental Design: Procedures for Behavioral Sciences Chapter 3 : (#4 & #5)
- A modern introduction to probability and statistics : understanding why and how : (QE 28.1, QE 28.2, QE 28.3, QE 25.1, QE 25.3, QE 25.4)
- Download the R lab data set, run the ANOVA function and make sure you get the same p-value.

You want more ?!?!

- A modern introduction to probability and statistics : understanding why and how : selected exerices for ch. 25,26.

External Ressources

- Experimental Design: Procedures for Behavioral Sciences Ch. 3.
- A modern introduction to probability and statistics : understanding why and how Ch. 25, 26, 27 & 28
- A Modern Approach to Regression with R Ch.2
- Wikipedia

Bonus slides

ANOVA

- For ANOVA $\frac{n_t \sum_{t=1}^{T}(\bar{y}_t - \bar{y})^2}{T-1} = \hat{\sigma_T}$ and $\frac{\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{i,t} - \bar{y}_t)^2}{n-T} = \hat{\sigma}_r$.

Distribution Theory

- Well-know unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- So $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$
- We are getting there slowly but surely.