Statistical analysis in R

MiCM Fall Workshop Series (2022)

Alex Diaz-Papkovich

McGill

McGill initiative in Computational Medicine

Prerequisites

- Have R and RStudio (or equivalent) installed
- Understand basics of R and RStudio
    - Terminal, console, script editor, environment windows
- Please feel free to ask questions
    - Even if it is “what does this word mean?”

Libraries

- `install.packages("medicaldata")`
- `install.packages("scatterplot3d")`
- `install.packages("ggplot2")`
- `install.packages("tidyverse")`

Online content

- shorturl.at/gkprL
- https://github.com/diazale/intro_to_r_stats

Core concepts

- R objects and variable types
- Hypothesis testing and analysis of variance (ANOVA)
- Linear regression
- Learn how ***and why*** to use statistical methods

Goals

- Learn core concepts in R and statistical analyses
- “Bread and butter” tools
    - The most commonly used methods
- Basics of quality and interpretable analysis
- Build up to univariate multiple linear regression

Generate + understand this

```
Call:
lm(formula = number12m ~ baseline + treatment + age + sex, data = medicaldata::polyps)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.202 -10.251  -3.595   7.657  23.272 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)   
(Intercept)       51.83751   13.06029   3.969  0.00123 **
baseline           0.05931    0.05441   1.090  0.29289   
treatmentsulindac -26.53350    7.10892  -3.732  0.00200 **
age               -0.70497    0.47822  -1.474  0.16112   
sexmale           -1.37836    8.19318  -0.168  0.86865   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.23 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.579,	Adjusted R-squared:  0.4667 
F-statistic: 5.157 on 4 and 15 DF,  p-value: 0.008127
```

Breakdown

I. R concepts (30 minutes + 15 minutes exercise)

II. Hypothesis tests (45 minutes + 15 minutes exercise)

III. Break (20 minutes)

IV. ANOVA + Regression (Remaining time)

-Overview with slides, in-depth with R Markdown

Structure

- Slides provide overview of concepts
- Some theory to explain math and statistics
  - Try to follow along, but this isn't a math-stats course
- Most work will be in R Markdown files
- Will give you example code to follow along
- Exercises with empty code cells provided

Data types and structures

“Computers are like Old Testament gods;
lots of rules and no mercy.”

–Joseph Campbell

R concepts

- R will do whatever you tell it to do
- You must understand what you are telling it to do
- Do not fish for numbers!
- Learn what you're using and how to get what you need
- Use the `?` command
  - e.g. `?lm` gives the documentation for `lm`
  - `??` will search R documentation instead

R variable types

- Variable types
  - Important ones: Boolean, Numeric, Character, Factor
- Boolean (Logical): {TRUE/FALSE}, {T/F}, {1/0}
- Numeric: Integers, real numbers (“double”)
- Character: Strings of text
- Factor: Categorical, Indicator, Boolean

Pitfalls to avoid

- Some characters have special meanings. **DO NOT USE THEM.**
    - c, g, t, C, D, F, I, T
- R often imports data as factors — check to make sure!
    - Object structure: `str()` command
    - Variable type: `typeof()` command
    - Class type: `class()` command
- Be careful coding categories as numbers
    - Make sure they are factors and not numeric

R objects

- Vectors: `c()`
  - Only one variable type (e.g. all numeric)
- Lists: `list()`
  - Can have multiple variable types
- Matrices
  - Multi-dimensional array of one variable type
- Data frames
  - Subtly different from tibbles

R objects

- Convert between types with `as.[type]`

```
x <- c(1,0,1)

typeof(x)

x <- as.logical(x)

typeof(x)
```

Data frames

- Data is usually stored in a **data frame**
- Observations are rows, variables are columns
- Created with `as.data.frame()`
  - Many functions import external data as data frames
  - Useful functions:

    `names(), colnames(), rownames(), subset()`
  - Conversion functions still apply

Data frames

- Examine with `str()`
- Subset with `[,]` or `subset()`
- Lots of ways to manipulate them, you can get clever
- Beyond the scope of this workshop
- For now, our data is already in good shape

`$` and `Summary()`

- Use `summary()` to retrieve summaries of results
  - e.g. ANOVA, regression
- Use `$` to retrieve parts of R objects
  - e.g. t-test, data frame columns

Hypothesis testing

Overview

- You want to compare data
- Are sets of data different?
  - How “different” are they? “Kinda” different? “Very” different?
  - Test statistics are a formal way of asking this
- We consider two types of tests:
  - t-test
  - ANalysis Of VAriance (ANOVA)

Statistical significance

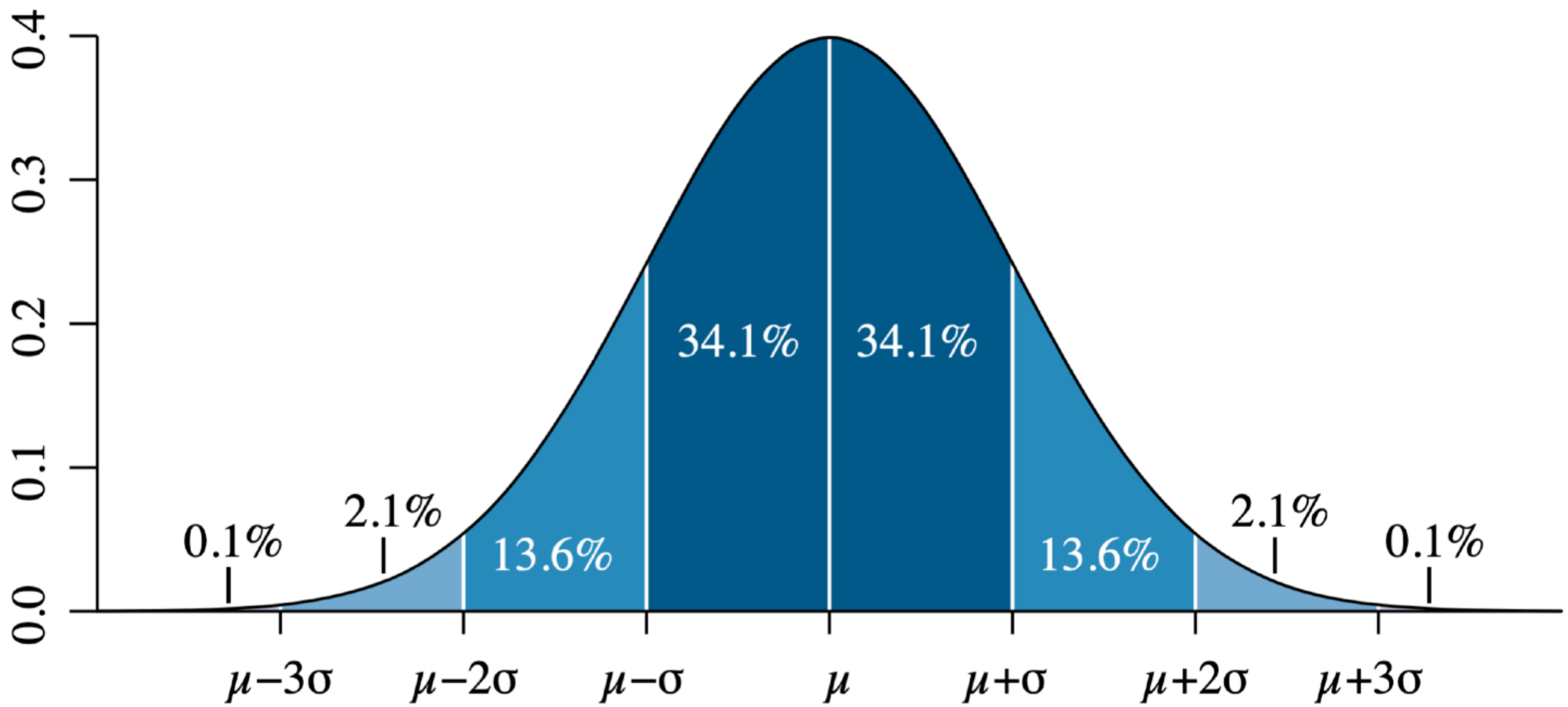
- “This measurement is unlikely to have happened by chance”
- Does <u>**not**</u> mean any of the following:
    - Logically sound
    - Scientifically meaningful
    - Causally linked
- All statistical tests and methods are just fancy calculations
- You, the scientist, make the decisions and interpretations
- <u>**Never choose a test based on whether it gives a smaller p-value**</u>

Quick math-stats review

- Random variables are defined by their distributions
- Distributions defined by PDFs (probability density functions)*
- We look at:
  - Normal
  - t-distribution
  - Chi-squared (sort of, as an intermediate)
  - F-distribution

*actually defined by many things, including PDFs, beyond scope of today

Normal distribution



Credit: Wikipedia

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

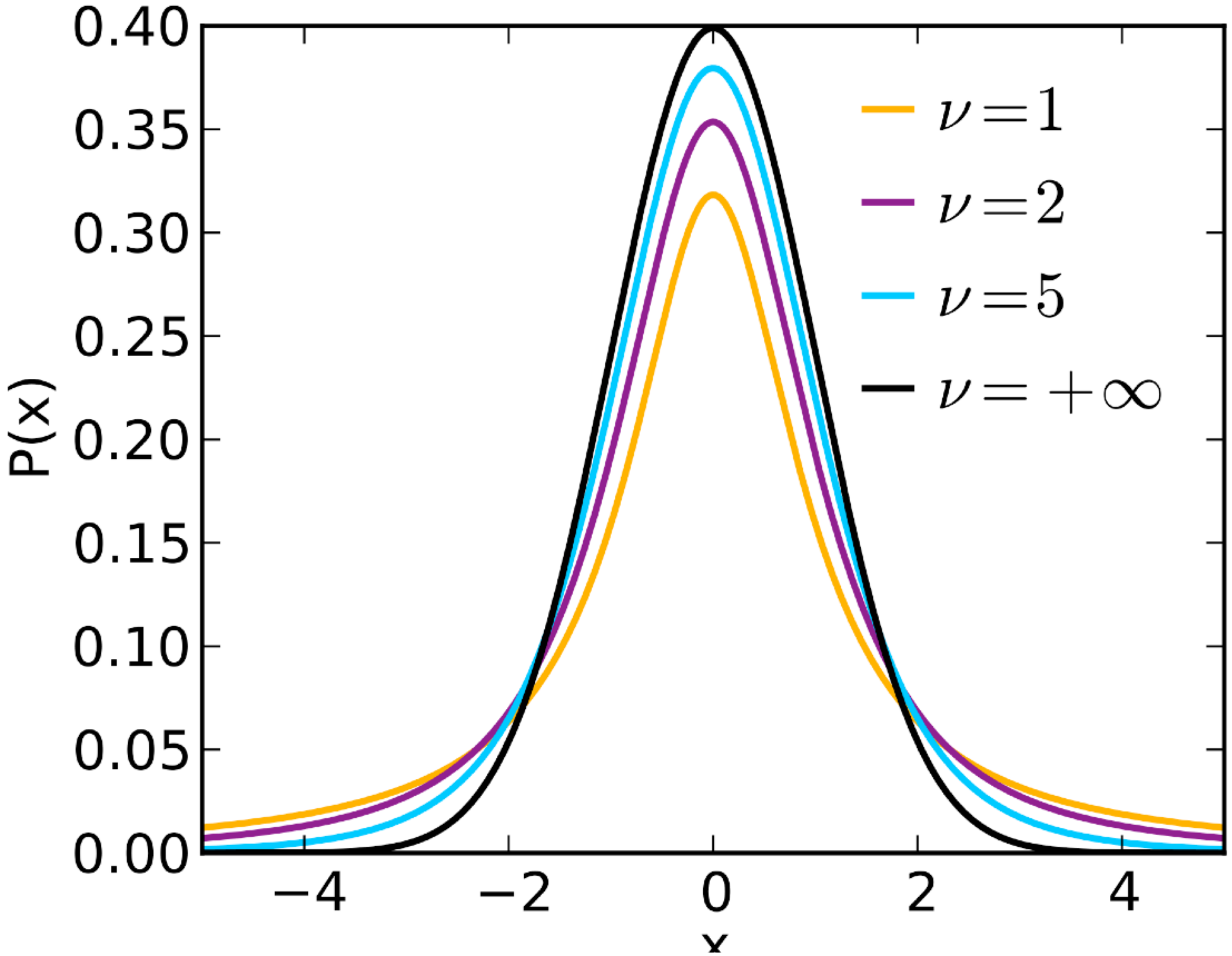
Normal distribution

- Extremely common, also called the Gaussian distribution
- Known for its bell curve
- Lots of nice properties
- Central Limit Theorem
  - For many distributions, their normalized sum is Normal
- Can be standardized to a N(0,1)

Properties of the Normal

- Still Normal when
    - Subtracting a constant, dividing by a non-zero constant
- Standardization:
    - Subtract the mean, divide by standard deviation
    - If $X \sim N(\mu, \sigma^2)$ then $\dfrac{X - \mu}{\sigma} \sim N(0,1)$
- If $X \sim N(0,1)$ then $X^2 \sim \chi^2_1$ (Chi-squared with 1 d.f.)
- If $X_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ then $\sum_{i=1}^{n} (X_i)^2 \sim \chi^2_n$

t-distribution



- Defined by degrees of freedom

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Credit: Wikipedia

Deriving the t-test

- Details are mathematically dense, this is a very brief summary
- Sample of *n* identically independently distributed (i.i.d.) values $X_i \sim N(\mu, \sigma^2)$

CORE ASSUMPTION

- We don't know the *true values* of μ or σ
- Test hypothesis about μ using the sample mean $\bar{X}$
- Estimate σ as $\hat{\sigma}$
- Our test statistic is $t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$, which follows a **t-distribution**
- t-distribution is defined by **degrees of freedom**. We have ***n - 1 (or*** $\nu - 1$***)***

Deriving the t-test

- We want to test whether $\mu = \mu_0$ for some value μ_0 **(Null hypothesis)**
- Often assume $\mu_0 = 0$ and test this assumption
  - This comes up in regression!
  - This test statistic is $t = \frac{Z}{s} = \frac{\bar{X} - 0}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{X}}{\hat{\sigma}/\sqrt{n}}$
- What is the probability that we get this statistic?
- This is the ***p-value***

Assumptions of t-test

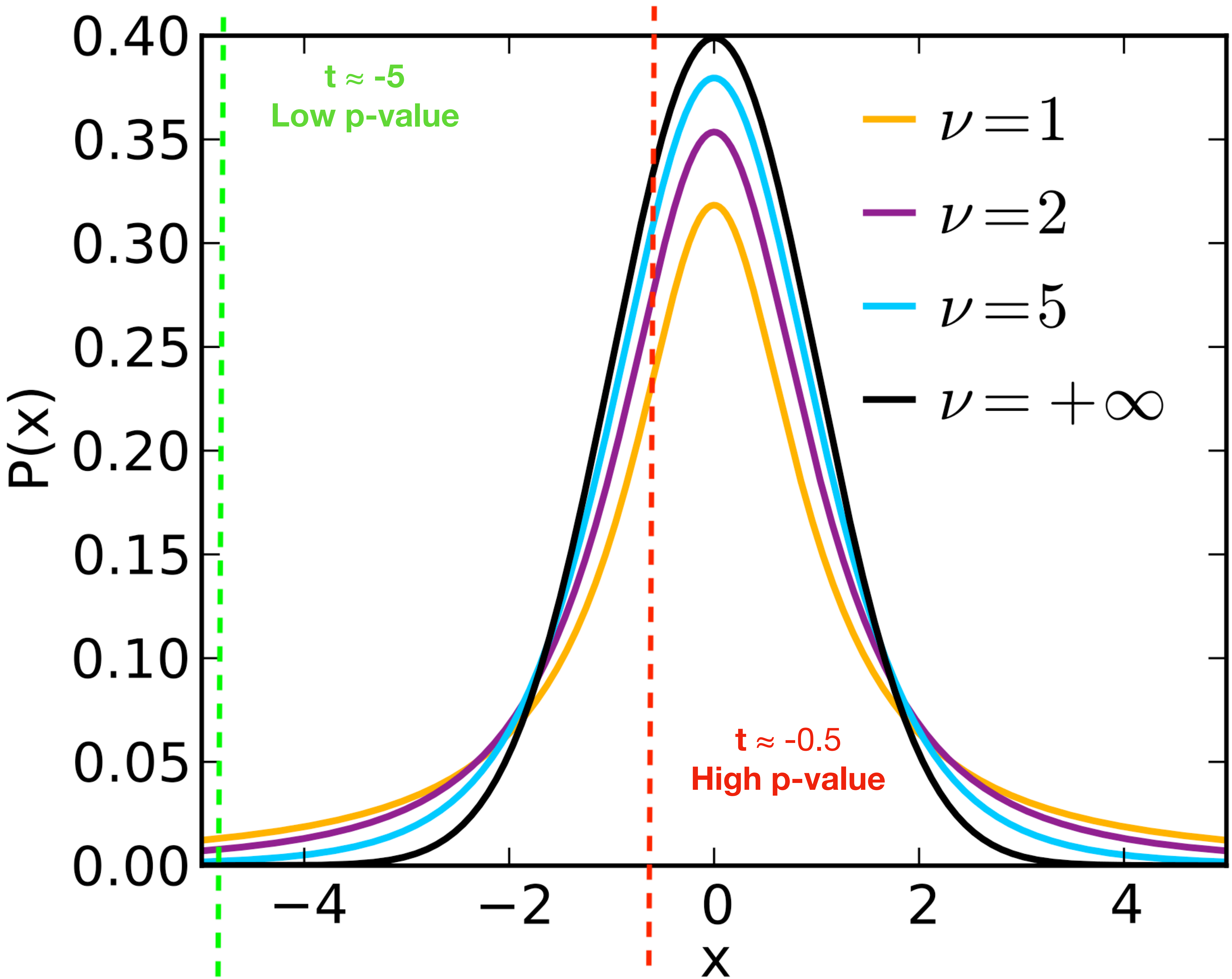
- Observations are <u>independent</u>
    - Don’t use the same observation in two groups
- No massive outliers
    - Look at your data!
- Data are approximately normal
- Groups should have approximately equal variances
    - Welch’s t-test is used if this assumption is violated

Two-sample t-test

- Test whether two population means are the same
  - Instead of testing whether one mean is a specific value
  - e.g. do two populations have the same mean height?
  - Similar derivation
  - $t = \dfrac{\bar{X}_1 - \bar{X}_2}{s}$, where s is a very long formula for s.e.

Using the t-test

- Software does most of the work for us, thank god
- We calculate t (our **t-statistic** or **test statistic**)
- What is the probability that we get this number?
- We find the area under the curve of the distribution
- The probability we get is the **p-value**
- If $p <$ threshold, we say it is “significant”
- A common threshold is 0.05



t ≈ -5
Low p-value
t ≈ -0.5
High p-value
$\nu = 1$
$\nu = 2$
$\nu = 5$
$\nu = +\infty$
P(x)
x
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4

Interpreting the t-test

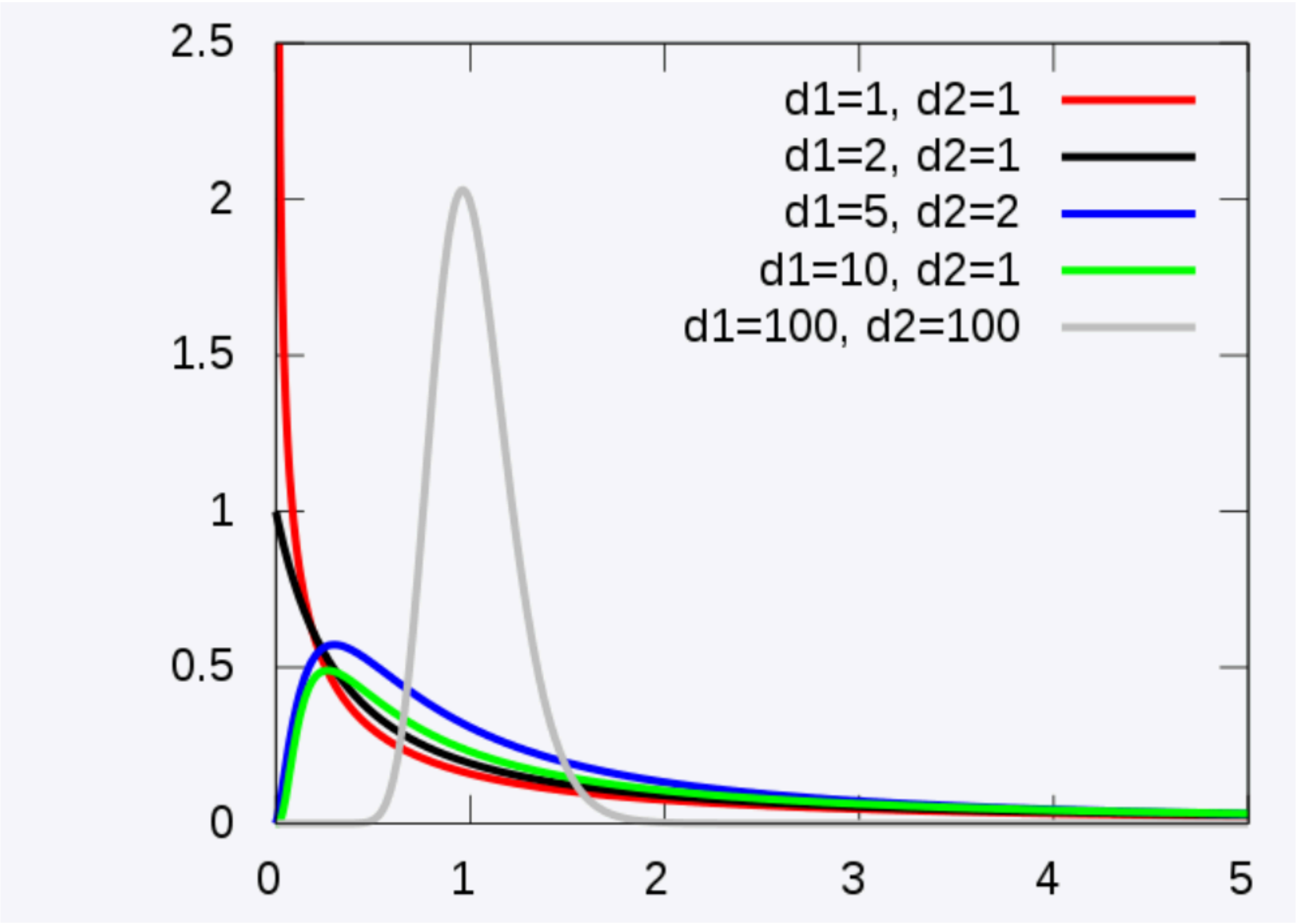
- <u>**Remember your hypothesis!**</u>
- Usually one of two:
  - The mean is equal to [some value]
  - The means of two populations are the same
- Can test if mean is bigger/smaller than [some value]
  - This is the one-tailed test
- A low p-value means the hypothesis is unlikely
- We <u>**reject the null hypothesis**</u>

Statistical significance

- Big t-statistic will give small p-value
- Recall: $t = \frac{Z}{s} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
- What makes *t* big?
  - Big numerator ($\bar{X} - \mu_0$); and/or Small denominator (big *n*)
- Huge sample sizes (big *n*) will give small p-values
- May be more interested in **effect size** (numerator)

ANOVA

F-distribution



- Defined by two degrees of freedom

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \, d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

Credit: Wikipedia

F-statistic

- Similar to t-statistic
- Follows Snedecor's F-distribution
- If $S_1 \sim \chi_a^2, S_2 \sim \chi_b^2$ then $\frac{S_1/a}{S_2/b} \sim F_{a,b}$
- The ratio of two Chi-squareds follows the F-distribution
- Null hypothesis: multiple means are <u>**all**</u> equal
- Rejection means: At least one group is different

F-statistic interpretation

- Null hypothesis: All groups are equal
- Logical opposite of this is:
    - **Not** all groups are equal
    - **At least one** group is not equal
        - Could have one group different
        - Could have all groups different
        - Could have anything in between

What do we check?

- Look at your data
- **<u>Look at your data</u>**
    - It’s the quickest and easiest way
- Pairwise comparisons with correction factor
    - Tukey’s HSD (`TukeyHSD()` in R)
    - Pairwise t-tests with Bonferroni correction

Motivation

- ANOVA generalizes the t-test by partitioning the variance
- We want to test the means of 3+ groups

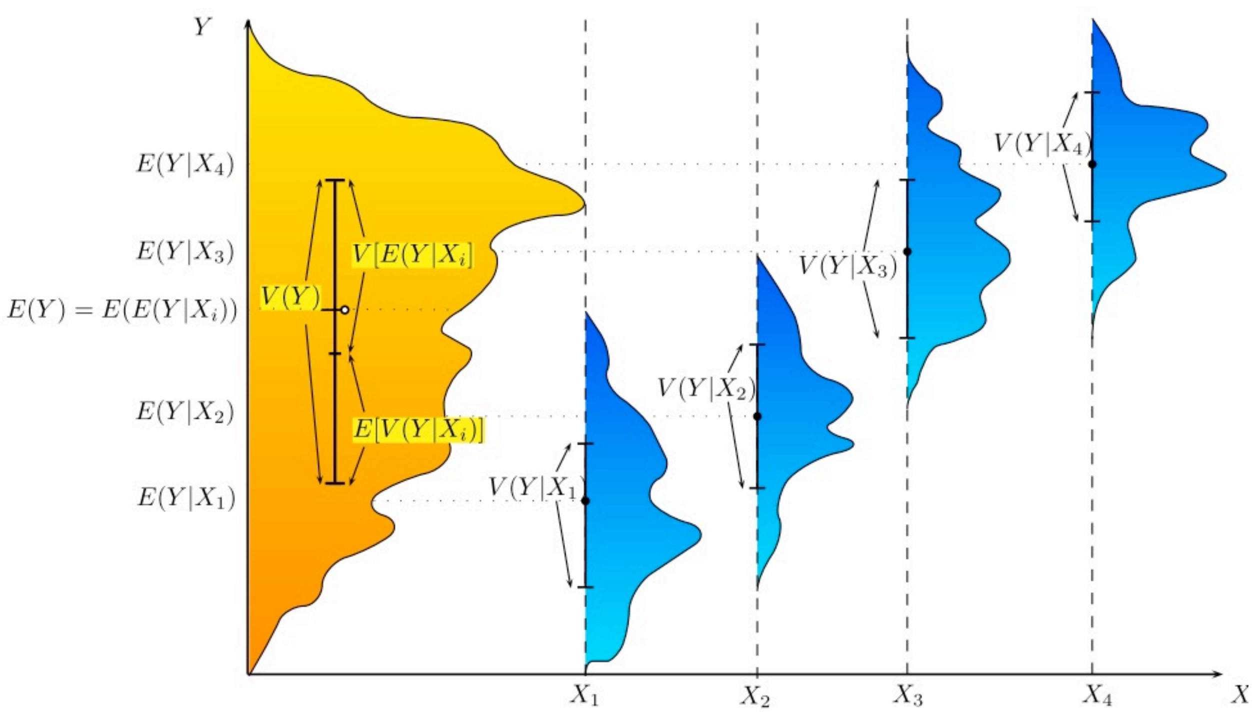


Figure 1: ANOVA : Fair fit

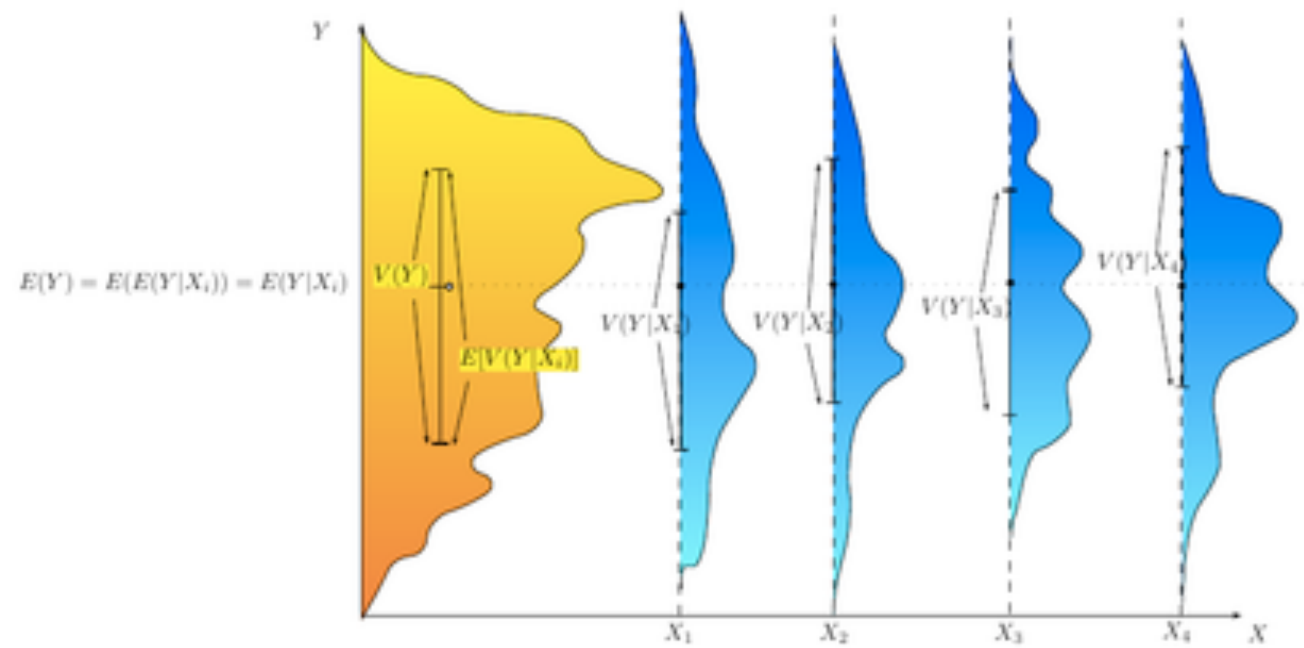


Figure 2: ANOVA : No fit

Credit: Wikipedia article for ANOVA

Motivation

- Interested in the relationship between variables
- Categories and continuous (factors and numerics)
- Examples:
  - Differences between population centres
  - Differences between treatment types
  - Differences between diets

Formulation

- $y_{ij} = \mu + \tau_j + \epsilon_{ij}, \epsilon_{ij} \sim N(0,\sigma^2)$
- Observations *i=1…I*, treatments (groups) *j=1,…,J*
    - Total mean μ
- How much does a measurement vary by group?
    - What is the effect of τ_j?
- $H_0 : \tau_1 = \ldots = \tau_J = 0$ vs. $H_A : \tau_i \neq \tau_j$ for some $(i, j), i \neq j$
    - All treatments are equal vs at least one treatment is not

Formulation

- Equivalent formulation:
  $y_{ij} = \mu_j + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0,\sigma^2)$

- How much variation is in our data?

  $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

  (sum of squares, total)

  $$= SS(Model) + SS(Residuals)$$
  $$= SS(Treatments) + SS(Errors)$$

- How much variation explained by our treatments/model?

Formulation

ANOVA table for fixed model, single factor, fully randomized experiment

Source of variation	Sums of squares Explanatory SS	Sums of squares Computational SS	Degrees of freedom DF	Mean square MS	F
Treatments	$\sum_{Treatments} I_j(m_j - m)^2$	$\sum_j \frac{(\sum_i y_{ij})^2}{I_j} - \frac{(\sum_j \sum_i y_{ij})^2}{I}$	$J - 1$	$\frac{SS_{Treatment}}{DF_{Treatment}}$	$\frac{MS_{Treatment}}{MS_{Error}}$
Error	$\sum_{Treatments} (I_j - 1)s_j^2$	$\sum_j \sum_i y_{ij}^2 - \sum_j \frac{(\sum_i y_{ij})^2}{I_j}$	$I - J$	$\frac{SS_{Error}}{DF_{Error}}$	
Total	$\sum_{Observations} (y_{ij} - m)^2$	$\sum_j \sum_i y_{ij}^2 - \frac{(\sum_j \sum_i y_{ij})^2}{I}$	$I - 1$		

MS_{Error} is the estimate of variance corresponding to σ^2 of the model.

Credit: Wikipedia article for one-way ANOVA

- Degrees of freedom are (J-1), (N-J)
- Ratio of (variance between groups) to (variance within groups)
- If there is high variance between groups, we get a low p-value
- R does all the calculations via `aov()`

Extensions

- Two-way ANOVA (interactions)
- Three-way ANOVA
- Blocking (grouping, confounding, etc)
- Experimental design in general

Linear regression

Formulation

- We have two continuous variables
- Want to measure their relationship, assume it's linear
- Draw a line of best fit between them

  Remember from high school: $y = mx + b$

  What if we had *y* and *x* and needed to estimate *m*,*b*?

- We would also need to account for error in estimates

Motivation

- Simple but powerful concept: line of best fit
- Measure relationship between variables
- Simplest form: correlation between continuous variables
  - Simple single linear regression
- Very powerful, robust, flexible tool
- Can predict (extrapolate) and infer (interpolate)

Assumptions

- Outcome is **continuous** and a **linear combination of predictors**
    - Predictors must not be perfectly correlated
- Errors are $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$. For every observation *i* the error is:
    - **Normally distributed**
    - **Mean zero**
    - **Homoskedastic** (same variance as other observations)
    - **Independent** (not correlated)
- Patterns in your errors are <u>**very bad**</u> and <u>**should be examined!!!!!**</u>

Mathematical form

(linear model)

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon, \epsilon \sim N(0, \sigma^2)$
- Outcome (y) is a linear combination of measurements (x)
    - Values of y and x are **known**
- Related to each other by intercept (β_0) and slopes (β_j)
    - Not known, so we **estimate** them
- Errors are <u>**random**</u>, on average <u>**zero**</u>, with <u>**same variance**</u>
    - **On average**, model won't over/underestimate (mean zero, unbiased)
    - Errors do not grow/shrink with data (constant variance)

Connecting concepts

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon, \epsilon \sim N(0, \sigma^2)$
- ANOVA is a special case of linear regression
  - Categorical variables rather than continuous
- t-tests are used to test individual slopes in regression

  $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$
- F-tests are used to test all slopes in regression

  $H_0 : \beta_1 = \ldots = \beta_p = 0$ vs $H_A : \beta_j \neq 0$ for some j
- Confidence intervals are built around normal variables

Connecting concepts

regression equation

```
Call:
lm(formula = number12m ~ baseline + treatment + age + sex, data = medicaldata::polyps)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.202 -10.251  -3.595   7.657  23.272 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)   
(Intercept)        51.83751   13.06029   3.969  0.00123 **
baseline            0.05931    0.05441   1.090  0.29289   
treatmentsulindac -26.53350    7.10892  -3.732  0.00200 **
age                -0.70497    0.47822  -1.474  0.16112   
sexmale            -1.37836    8.19318  -0.168  0.86865   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.23 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.579,	Adjusted R-squared:  0.4667 
F-statistic: 5.157 on 4 and 15 DF,  p-value: 0.008127
```

categorical data

t-statistics

variance explained + F-statistics

Extensions

- Interaction effects
  - Outcomes depend on interactions of (e.g.) age*sex
- What if we have other types of data?
  - Multivariate regression (multiple, correlated outcomes)
  - Logistic regression (Binary outcomes like lived/died)
  - Multinomial regression (Count data)
  - ARIMA (Time series—data correlated over time)

Importance

- Regression is a bedrock of modern science
- Understand that software is a calculator
- You are ultimately responsible for interpretation
    - Statistics will guide your decision-making

Resources

- Introduction to regression modeling (Bovas Abraham and Johannes Ledolter, 2006)
  - More mathematical but fairly concise
- Wikipedia pretty reliable for equations, summaries
- University lecture notes and course topics:
  - Regression
  - Experimental design