Analysis of variance continued
and
inference for regression

Overview

Review and continuation of one-way analysis of variance (ANOVA)

- Another example and insight into what the F-statistic is calculating
- Brief mention: two-way ANOVA

Inference for regression

- Resampling methods
- Parametric results

Announcement

Final exam review session

- Tuesday December 9th from 1-2:15pm
- In this classroom (Marsh)

Final exam:

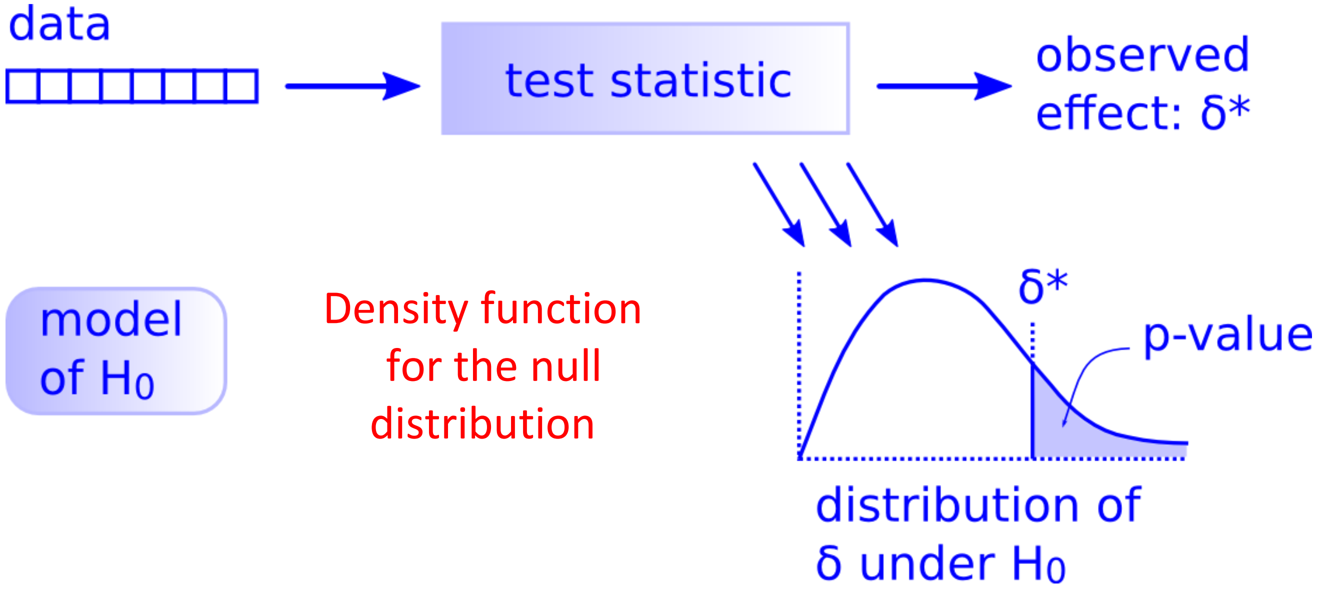
- Dec 15th (Monday) at 2pm
- Location…

Final project due Sunday Dec 7th

Review and continuation of one-way analysis of variance (ANOVA)

One test to rule them all

There is only one hypothesis test!



data

test statistic

observed effect: δ*

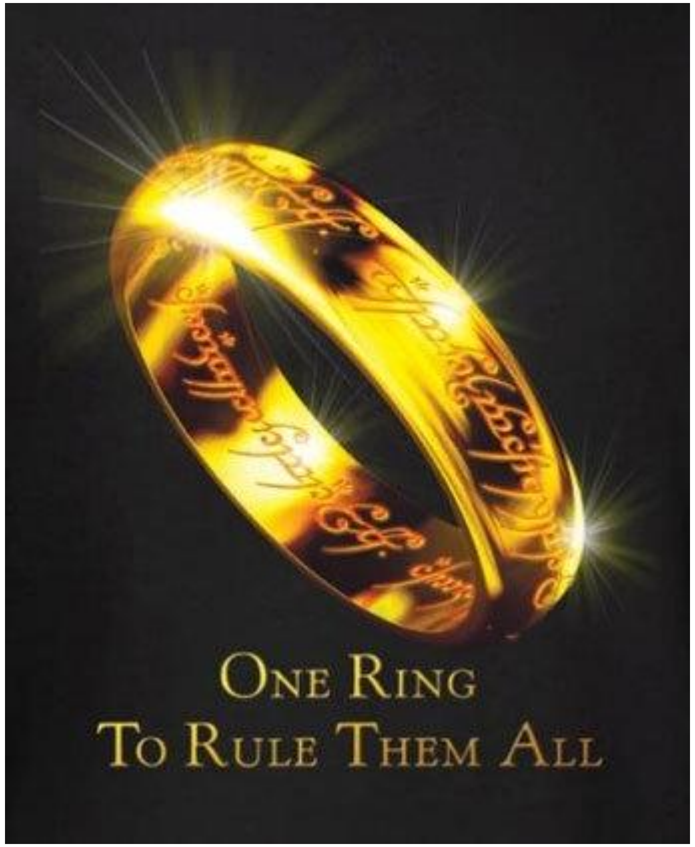
model of H_0

Density function for the null distribution

δ*

p-value

distribution of δ under H_0



Just follow the 5 hypothesis tests steps!

One-way ANOVA

An Analysis of Variance (ANOVA) is a parametric hypothesis test that can be used to examine if a set of means are all the same

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

H_A: $\mu_i \neq \mu_j$ for some i, j

The statistic we use for a one-way ANOVA is the F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

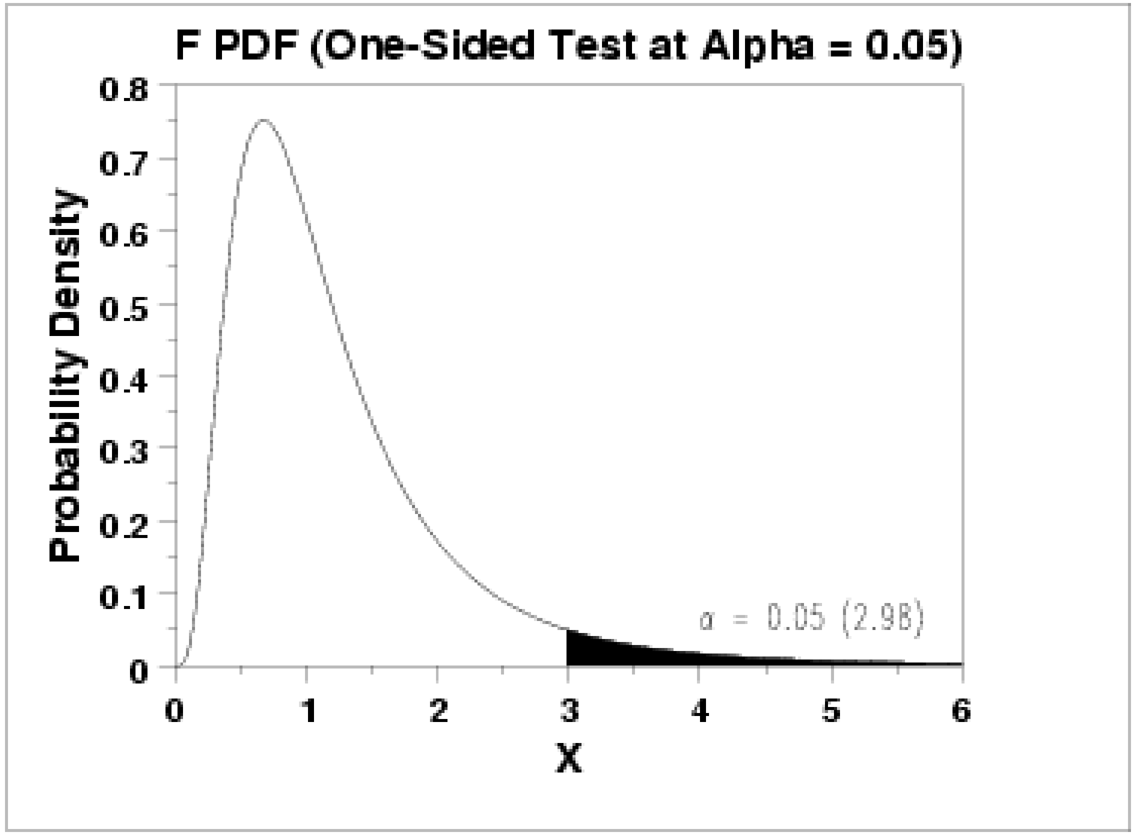
One-way ANOVA – the central idea

If H_0 is true, the F-statistic will come from an F distribution with parameters

- $df_1 = K – 1$
- $df_2 = N - K$

The F-distribution is valid if these conditions are met:

- The data in each group should follow a normal distribution
- The variances in each group should be approximately equal

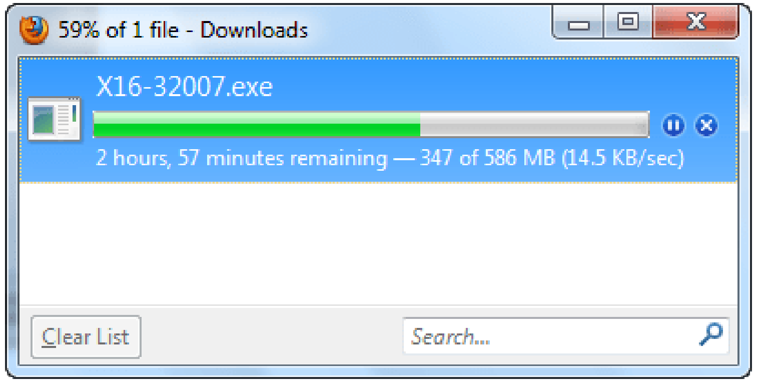


How does the time of the day affect download speeds?

To see whether the time of day affected the download speeds, a college sophomore performed an experiment

He placed a file on a remote server and then proceeded to download it at three different time periods of the day (7AM, 5PM, 12AM)

He downloaded the file 48 times in all, 16 times at each time of day, and recorded the time in seconds that the download took



59% of 1 file - Downloads

X16-32007.exe

2 hours, 57 minutes remaining — 347 of 586 MB (14.5 KB/sec)

1. State the null and alternative hypotheses

H$_0$: $\mu_{7AM} = \mu_{5PM} = \mu_{12AM}$

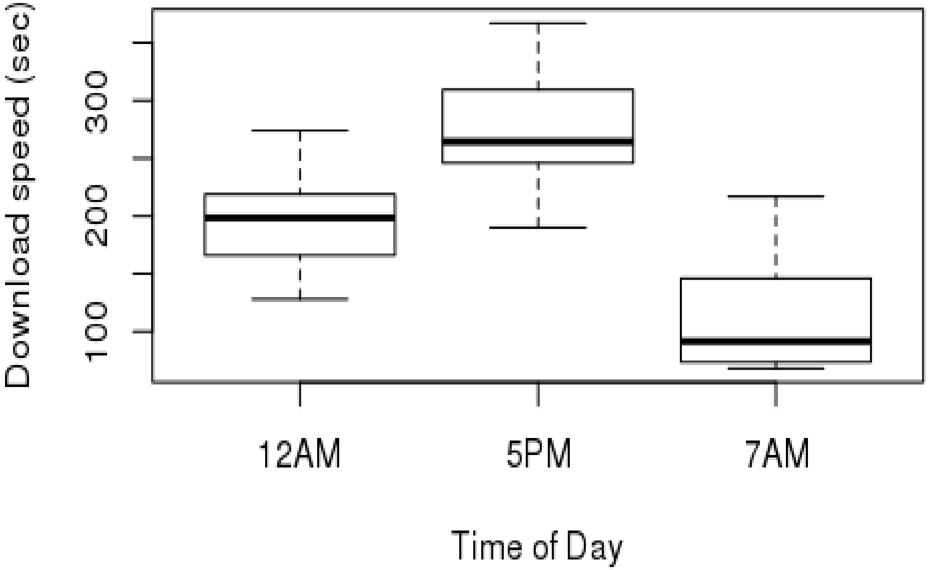
H$_A$: $\mu_i \neq \mu_j$ for one pair of the times of day

Let’s check if the ANOVA conditions are met first…

Checking ANOVA conditions (‘assumptions’)

We can check if the data in each group is relatively normal by creating boxplots and seeing:

- Is the data very skewed?
- Are there are many outliers?



We can check the equal variance condition by seeing if the ratio of the largest to smallest standard deviation is greater than 2

- $s_{max}/s_{min} < 2$

$s_{7AM} = 47.6$ $s_{12AM} = 40.9$

$s_{5PM} = 52.2$

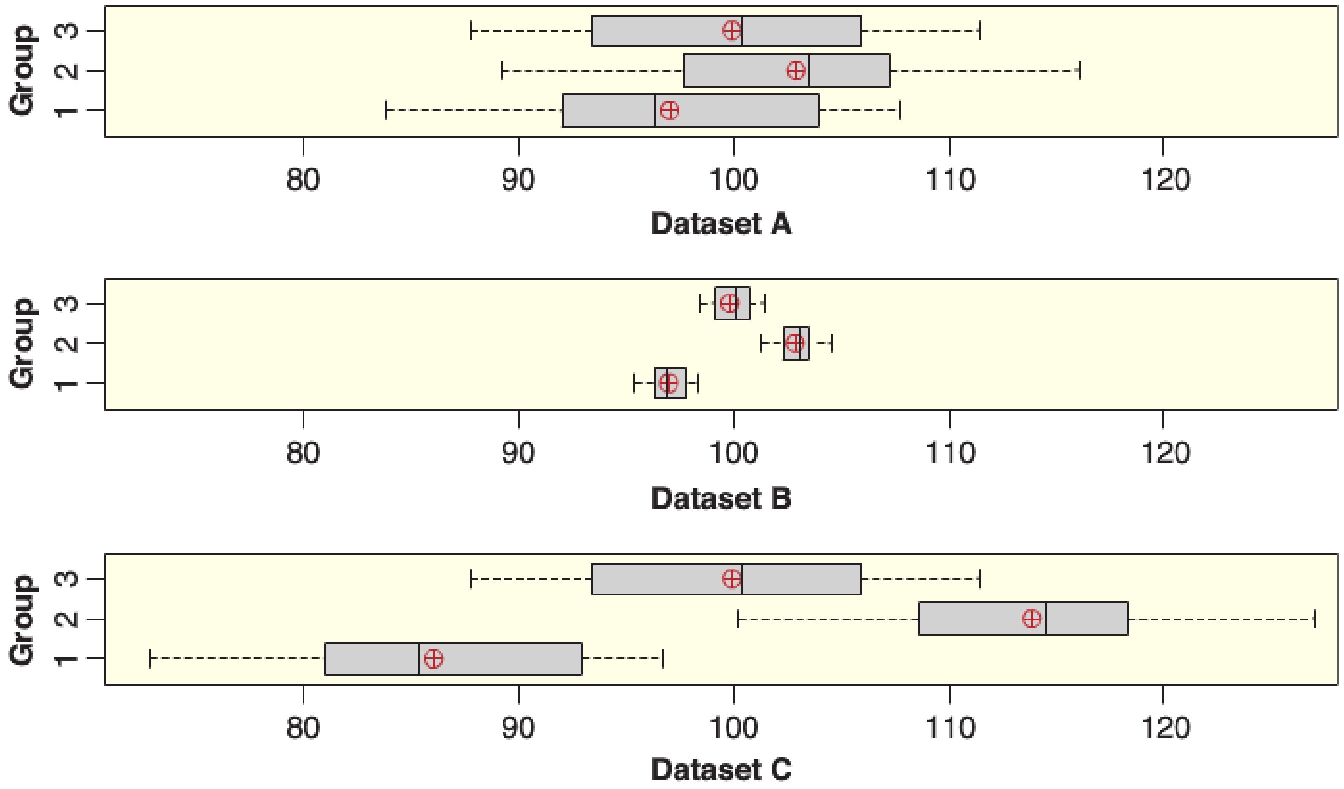
2. Calculating the observed F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

K: the number of groups	K = 3 different times of day
N: total number of points	N = 48 total downloads (16 * 3)
$\bar{x}_{tot}$: the mean across all the data	$\bar{x}_{tot}$: the mean speed across all data
$\bar{x}_i$: the mean of group i	$\bar{x}_i$: the means for the ith time of day
n_i: the number of points in group i	n_i = 16 downloads for each time of day
x_{ij} : the jth data point from group i	x_{ij} : the jth download at the ith time of day

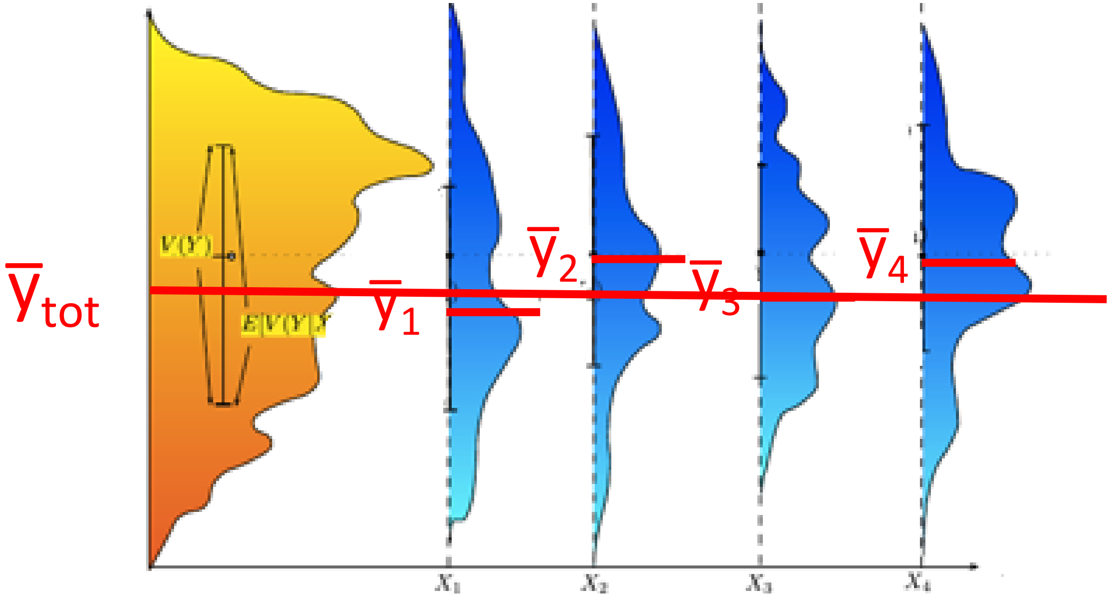
Why use the F-Statistic?

Which dataset gives the strongest evidence that there is a difference in population means?



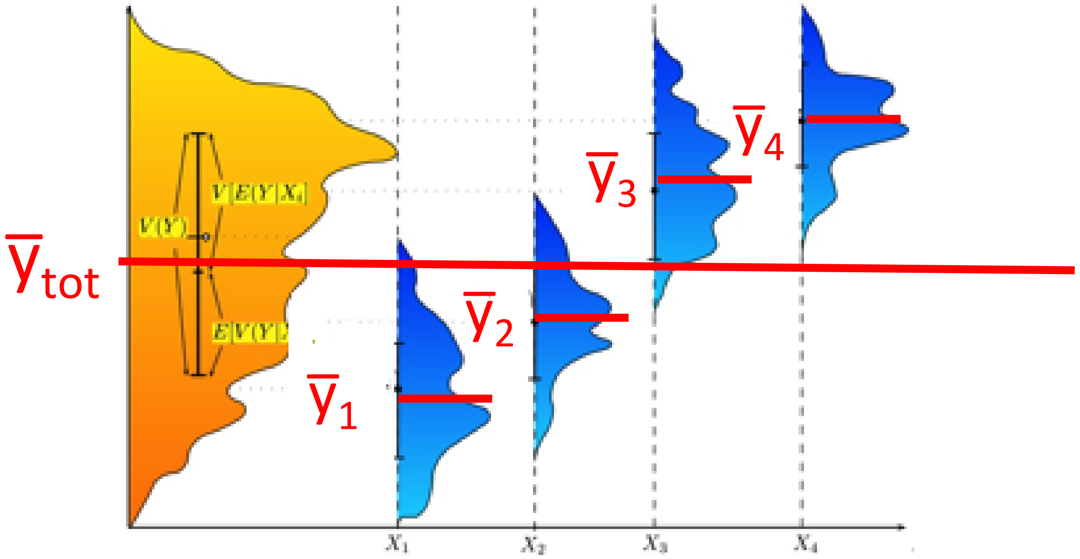
The F-Statistic

If H_0 is true, the data from all groups have the same means



- Similar means $\bar{y}_i$
- Similar spread s_i

If H_0 is not true, the data from all groups do not have the same mean



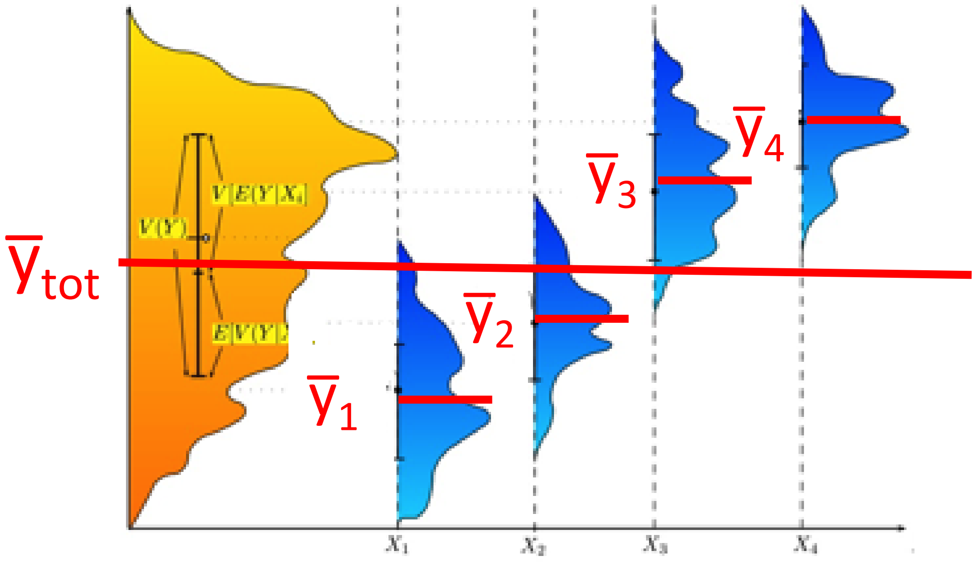
- Different means $\bar{y}_i$
- Smaller spreads s_i

The F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\text{variability within each group}}$$



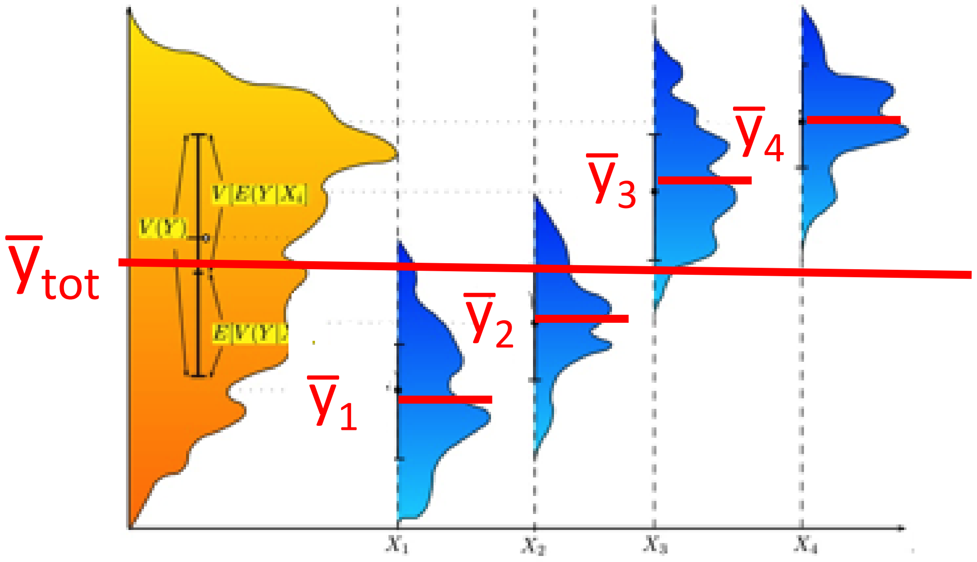
The F-statistic

Sum of Squares Group (SSG)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\text{variability within each group}}$$



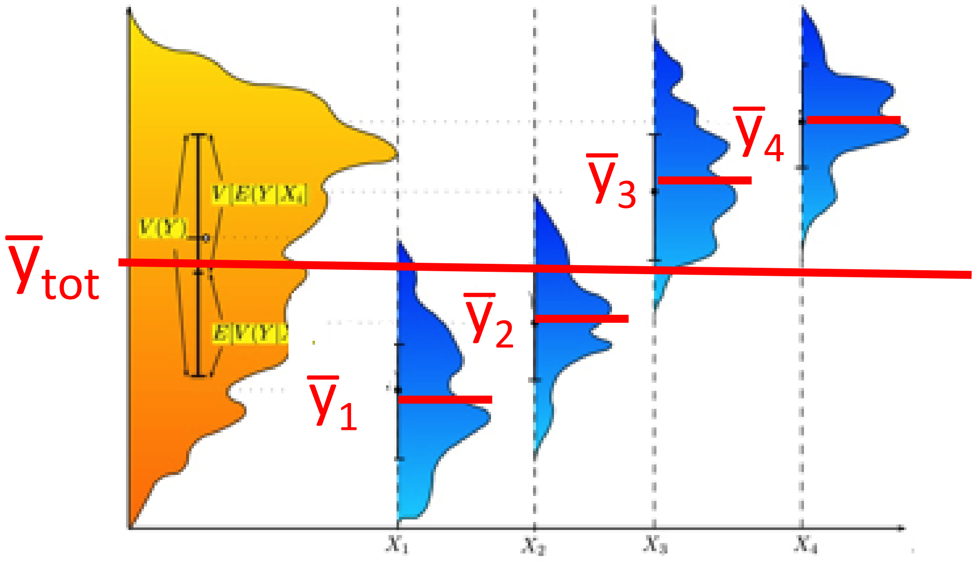
The F-statistic

Mean Squares Group (MSG)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\text{variability within each group}}$$



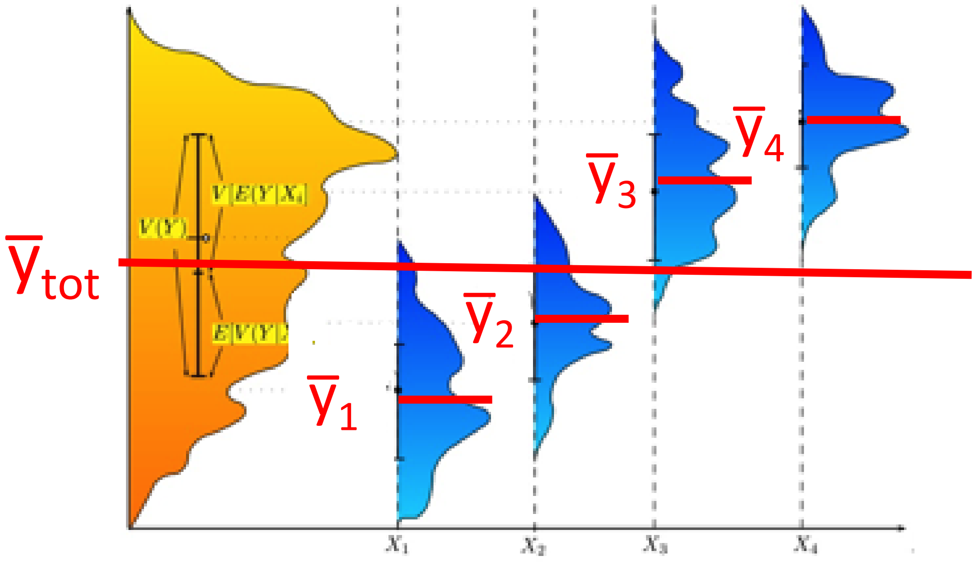
The F-statistic

Sum of Squares Error (SSE)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{variability within each group}}$$



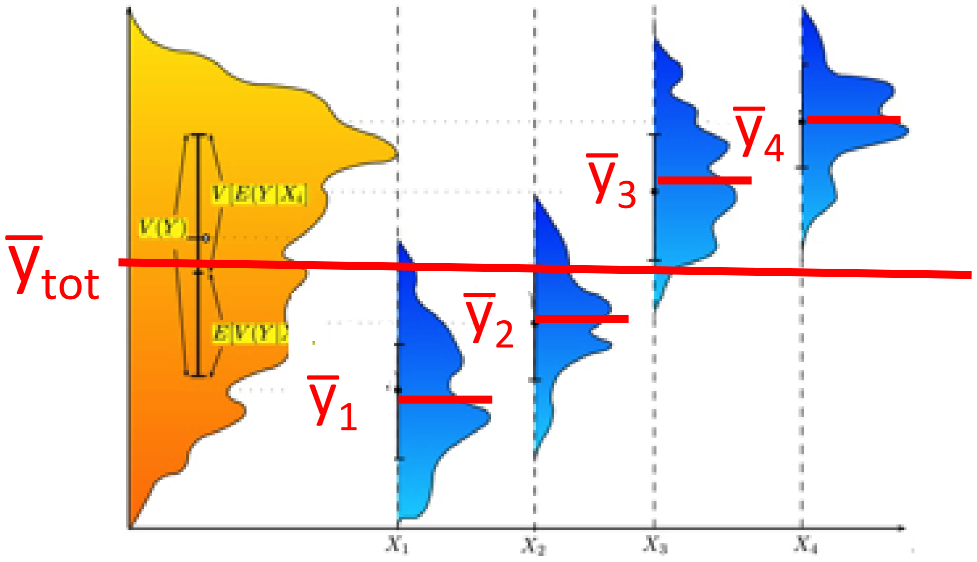
The F-statistic

Mean of Squares Error (MSE)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{variability within each group}}$$



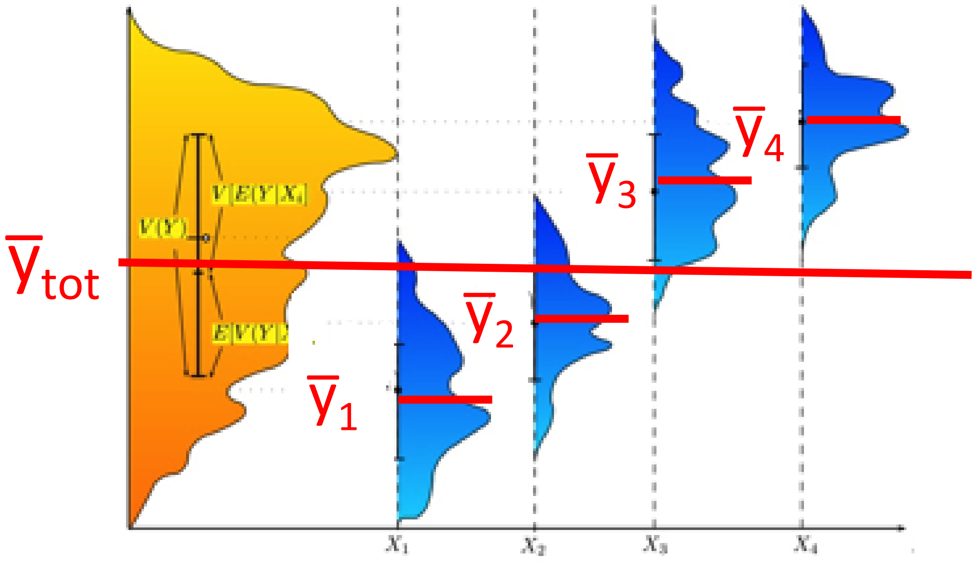
The F-statistic

Mean of Squares Error (MSE)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{Mean of Squares Error (MSE)}}$$



The F-statistic

Mean of Squares Error (MSE)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\boxed{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}}$$

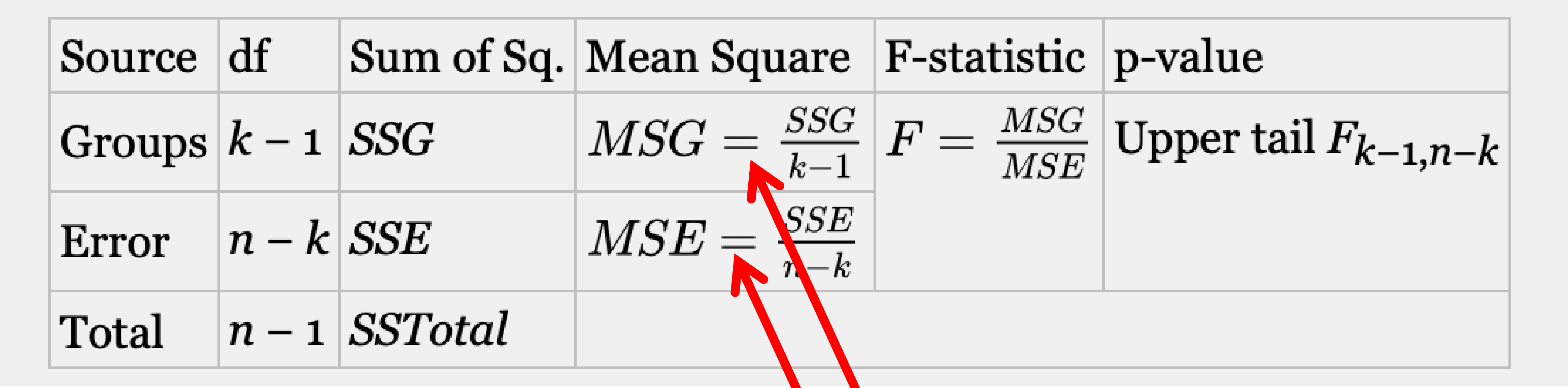
The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{Mean of Squares Error (MSE)}}$$

If the null hypothesis is true, the F-statistic will be around 1

Larger values, are stronger evidence against the null

ANOVA table



Source	df	Sum of Sq.	Mean Square	F-statistic	p-value
Groups	$k-1$	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	Upper tail $F_{k-1,n-k}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	$SSTotal$			

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

Let’s complete our analysis of download speeds in R…

Brief mention: Two-way ANOVA

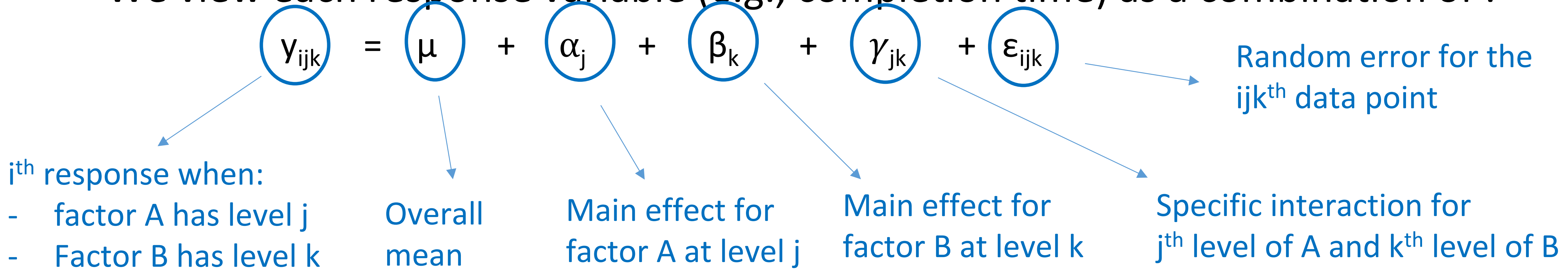
Brief mention: Two-way ANOVA

In a **two-way analysis of variance** (two-way ANOVA) we assess whether a quantitative variable is influenced by two factors

For example, is the time it takes to complete a sudoku influenced by:

- **A students' major area**: applied science, natural science, social science, arts/humanities
- **The year a student is**: first-year, sophomore, junior, senior

We view each response variable (e.g., completion time) as a combination of :



$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

i^{th} response when:
- factor A has level j
- Factor B has level k

Overall mean

Main effect for factor A at level j

Main effect for factor B at level k

Specific interaction for j^{th} level of A and k^{th} level of B

Random error for the ijk^{th} data point

Two-way ANOVA hypotheses

Main effect for A (major doesn't matter)

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_{J-1} = 0$

H_A: $\alpha_j \neq 0$ for some j

Main effect for B (year in doesn't matter)

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$

H_A: $\beta_k \neq 0$ for some k

Interaction effect (exact major-year combo):

H_0: All $\gamma_{jk} = 0$

H_A: $\gamma_{jk} \neq 0$ for some j, k

Where:

α_j: main effect for factor A at level j

β_k: main effect for factor B at level k

γ_{jk} : interaction between level j of factor A, and level k of factor B.

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

Brief mention: Two-way ANOVA

To learn more, take more advanced statistics classes!

- E.g., Data Exploration and Analysis (S&DS 2300)

Inference in regression using simulation methods

Review of regression (class 6 and 7)

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

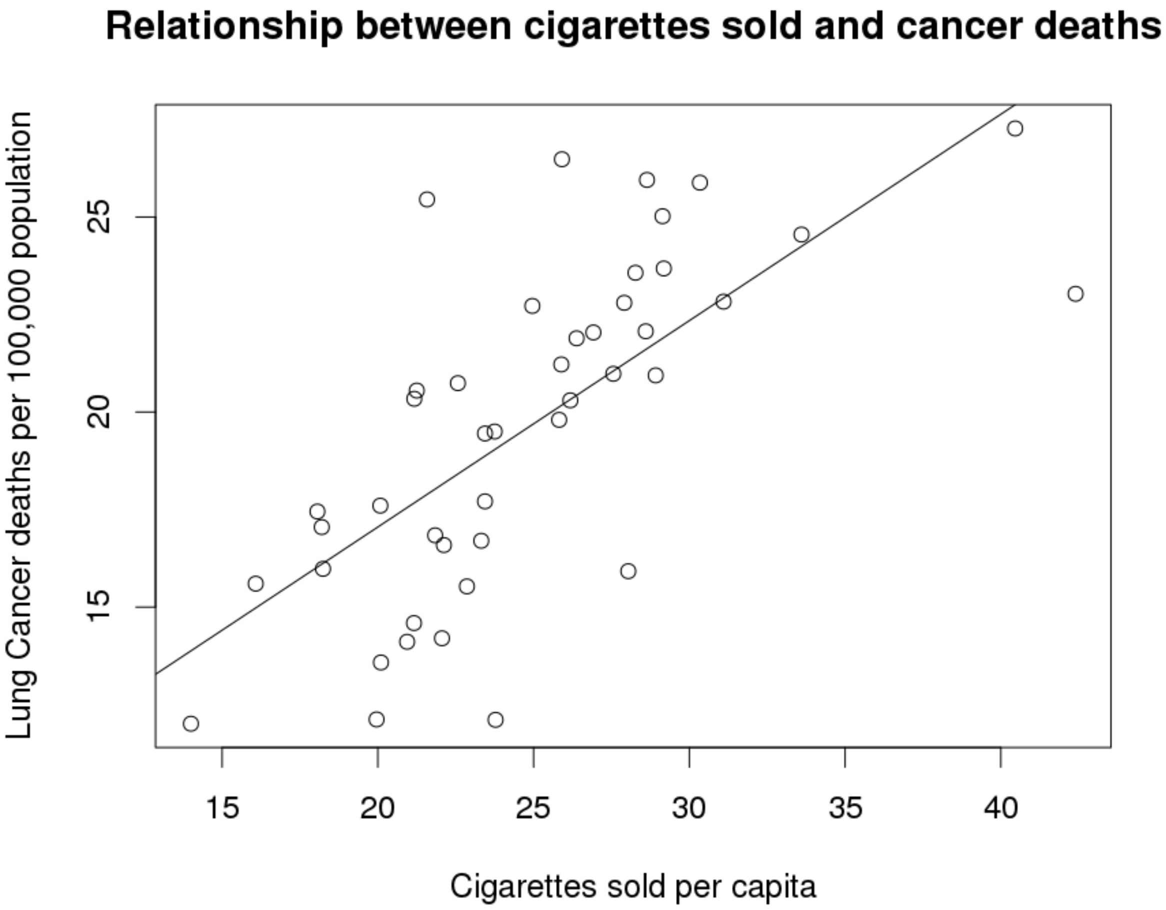
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = a + b \cdot x$$

$$Response = a + b \cdot Explanatory$$

Review cancer smoking regression line



$\hat{y} = a + b \cdot x$

```
R: my_fit <- lm(y ~ x)
   coef(my_fit)
```

$a = 6.47 \quad b = 0.53$

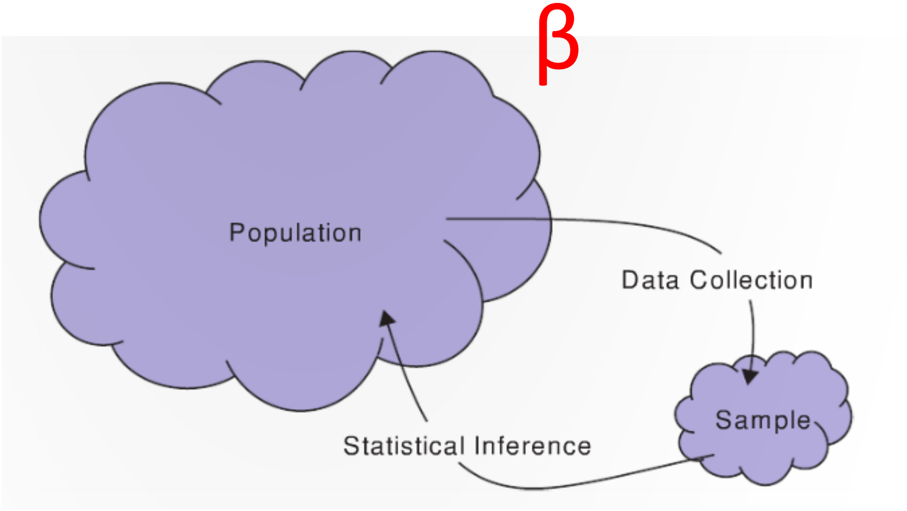
$\hat{y} = 6.47 + .53 \cdot x$

Review regression notation

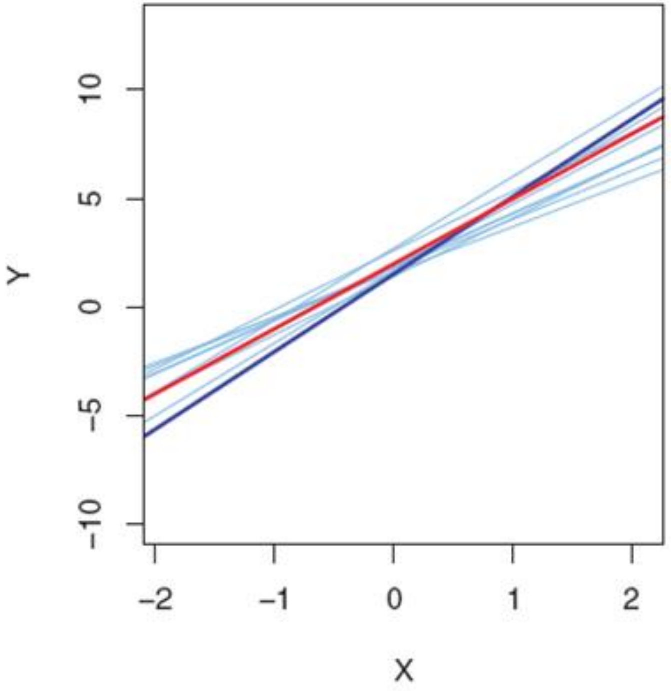
The Greek letter **β** is used to denote the slope of the **population**

The letter **b** is typically used to denote the slope of the **sample**

β



b



Inference for regression

How can we create confidence intervals and run hypothesis tests for the regression slope β?

Any ideas?

Using the bootstrap to create confidence intervals

We could use the bootstrap to create confidence intervals by:

1. Creating a bootstrap sample by sampling with replacement from our *paired data*
   - SDS1000: resample_pairs(v1, v2)
2. Fitting a regression line to our bootstrap sample and extracting the slope b
3. Repeat 10,000 times to get a bootstrap distribution of b's
4. Taking the standard deviation of the bootstrap distribution to get SE*
5. Using our confidence interval formula:

$$Statistic \pm 1.96 \cdot SE^*$$

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Using permutation hypothesis tests

If we wanted to run a hypothesis tests for the regression slope, how would we write the null and alternative hypotheses using symbols?

H_0: $\beta = 0$

H_A: $\beta \neq 0$

Any ideas how to run a permutation test to assess whether H_0: $\beta = 0$?

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Using permutation hypothesis tests

We could use run a permutation test for H_0: $\beta = 0$ by creating a null distribution using:

1. Shuffle one of the columns of data
2. Fitting a regression line to our bootstrap sample and extracting the slope b
3. Repeat 10,000 times to get a null distribution of b's

We can obtain a p-value by seeing how many points in the null distribution are greater than the observed statistic value of b

Let's try it in R!

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Parametric inference for regression

Review of regression (class 6 and 7)

In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = a + b \cdot x$$

$$\textit{Predicted response} = a + b \cdot \textit{Explanatory}$$

Change in notation to be consistent with the Lock5 and what most statisticians use

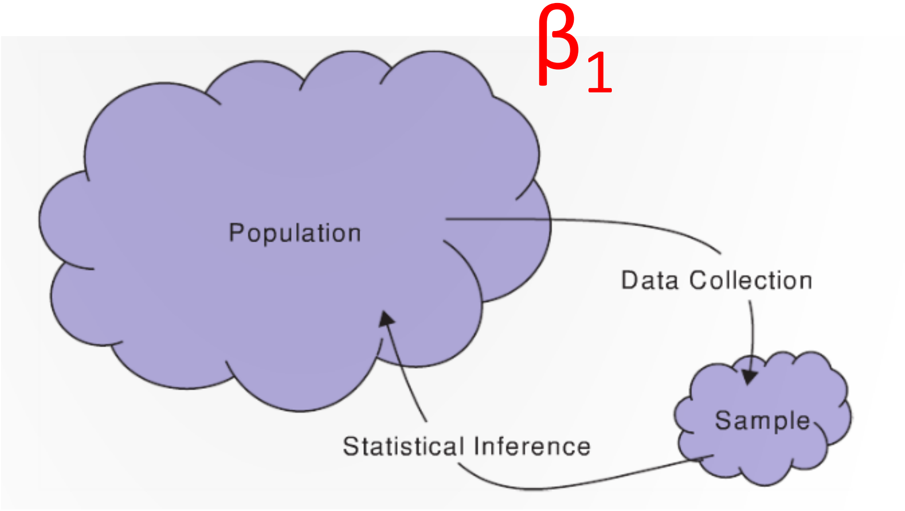
$$\textit{Predicted response} = b_0 + b_1 \cdot \textit{Explanatory}$$

Inference on simple linear regression

The Greek letter β_1 is used to denote the slope of the population

The letter b_1 is typically used to denote the slope of the sample

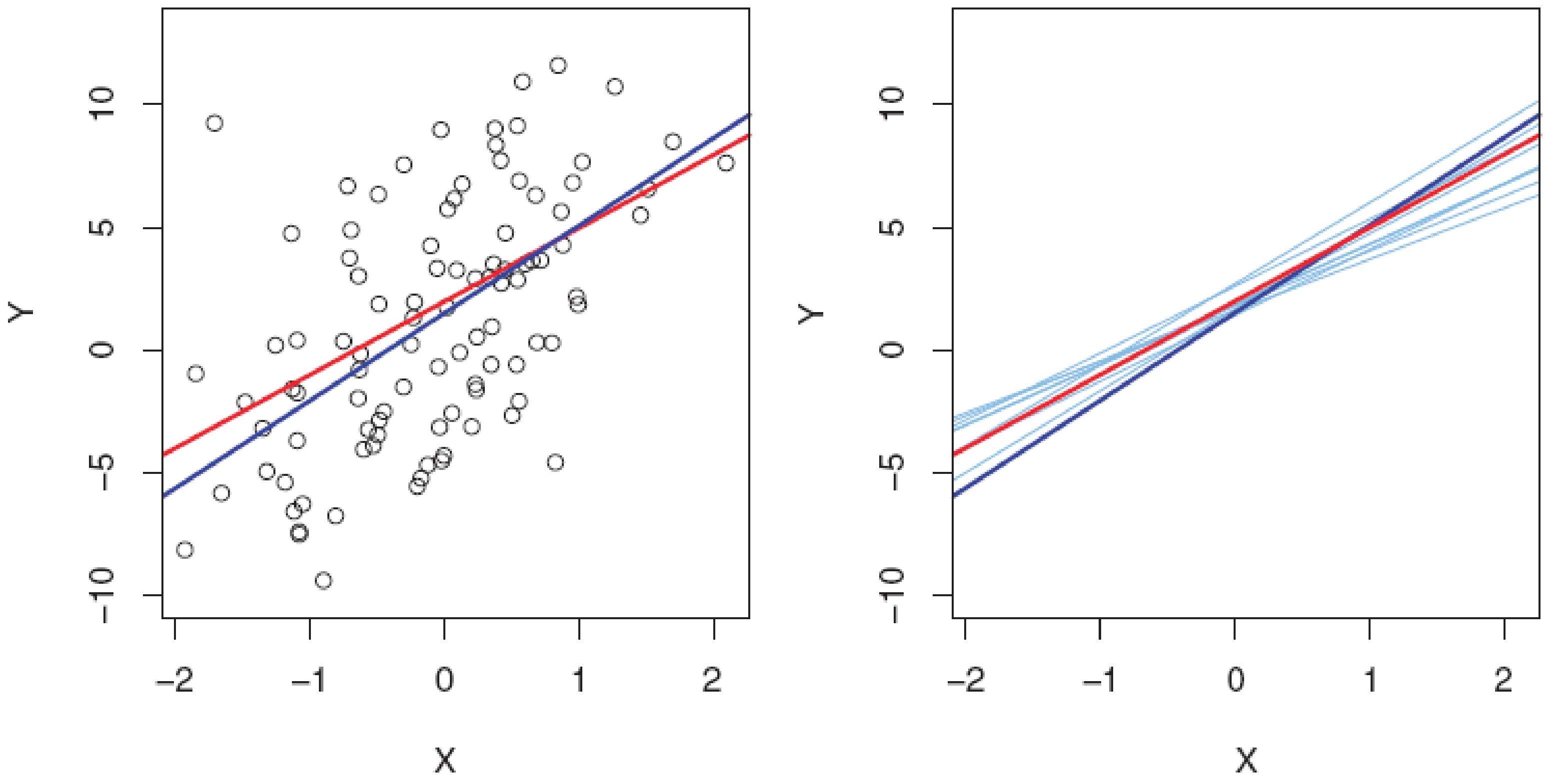
β_1



b_1

Population: β_1

Sample estimates: b_1

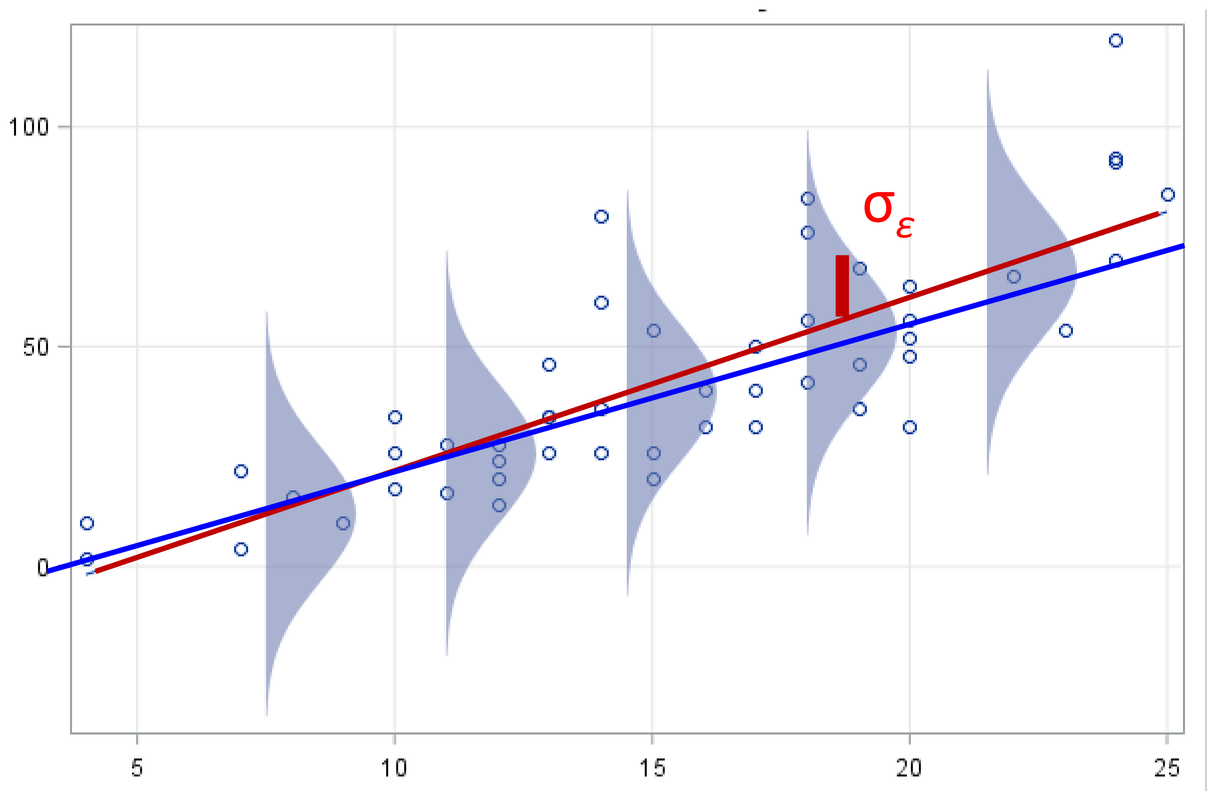


Simple linear regression underlying model

Intercept Slope } *Parameters*

$$Y \approx \beta_0 + \beta_1 x$$

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

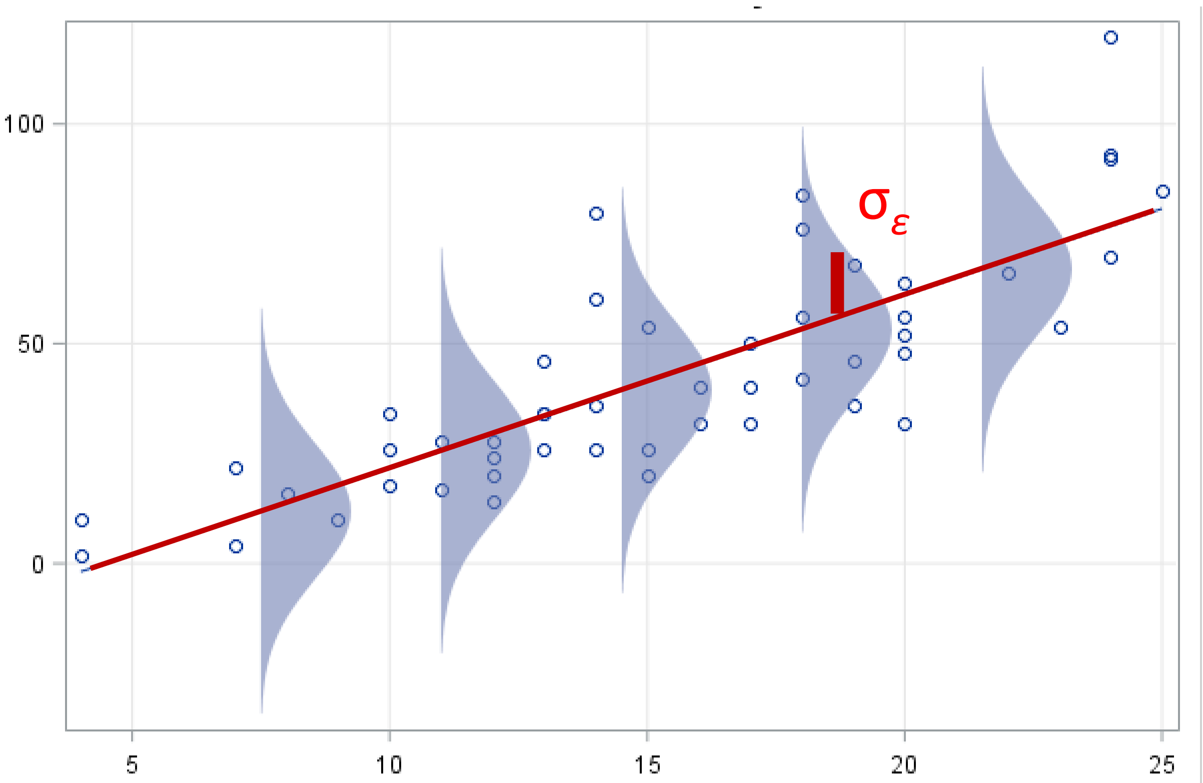


$$\epsilon \sim N(0, \sigma_\epsilon)$$

$$\hat{y} = b_0 + b_1 x$$

Estimating σ$_\varepsilon$

We can also use the **standard deviation of residuals σ$_e$** as an estimate standard deviation of irreducible noise **σ$_\varepsilon$**



$$\sigma_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x))^2}$$

Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

One type of hypothesis test we can run is based on a t-statistic:

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

$$SE_{b_1} = \frac{\sigma_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

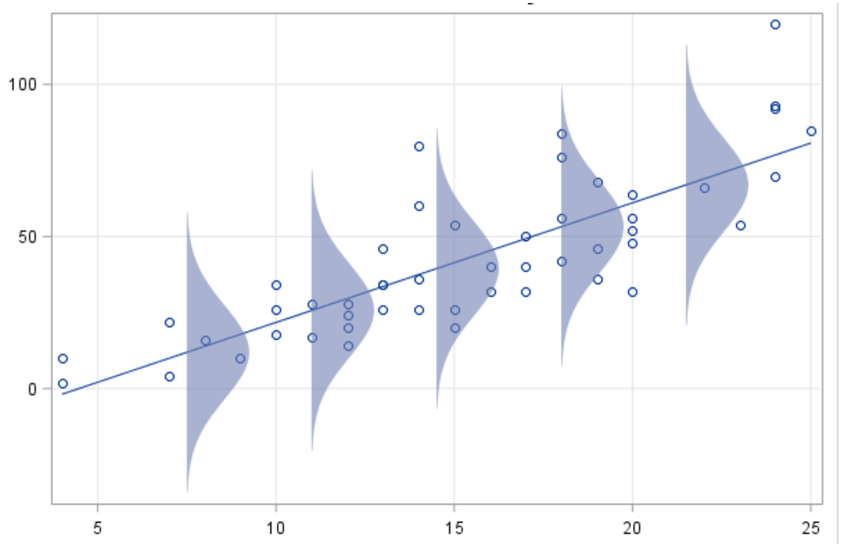
Inference using parametric methods

When using parametric methods, we make the following (LINE) assumptions:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_\epsilon)$$



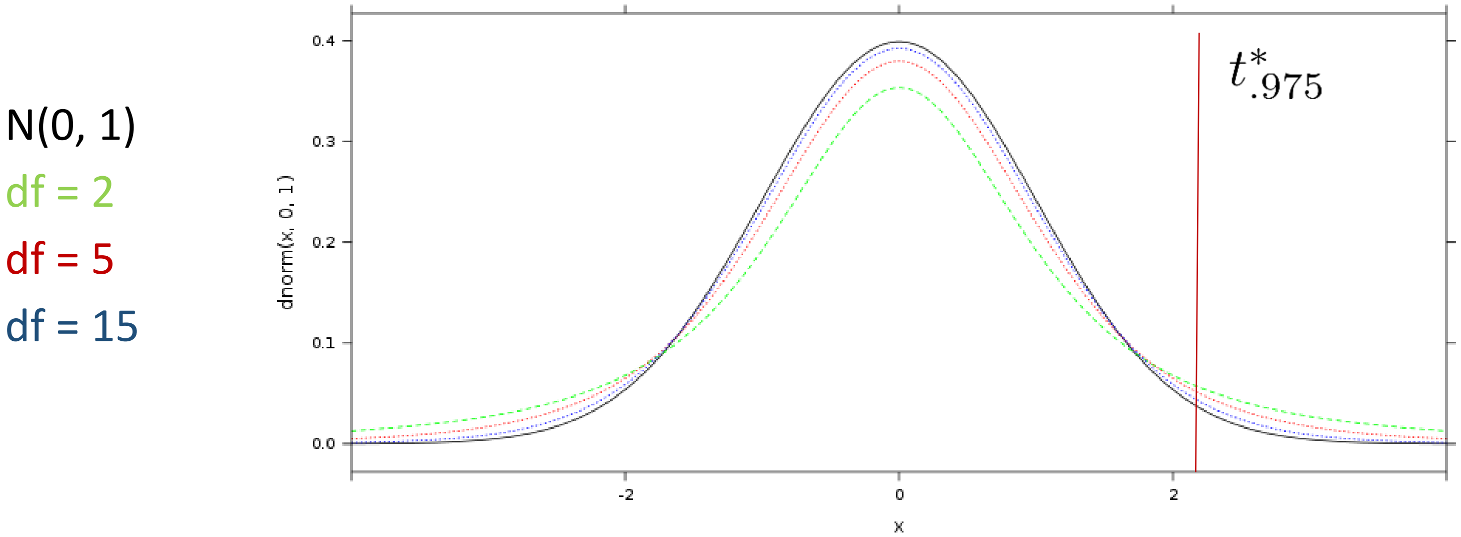
These assumptions are usually checked after the models are fit using ‘regression diagnostic’ plots.

Confidence intervals for regression coefficients

For the slope coefficient , the confidence interval is: $b_1 \pm t^* \cdot SE_{b_1}$

Where: $SE_{b_1} = \dfrac{\sigma_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence level



Let’s try it in R!