Analysis of variance continued
and
inference for regression

Overview

One-way analysis of variance (ANOVA)

- Review of one-way ANOVA
- Insight into the F-statistic

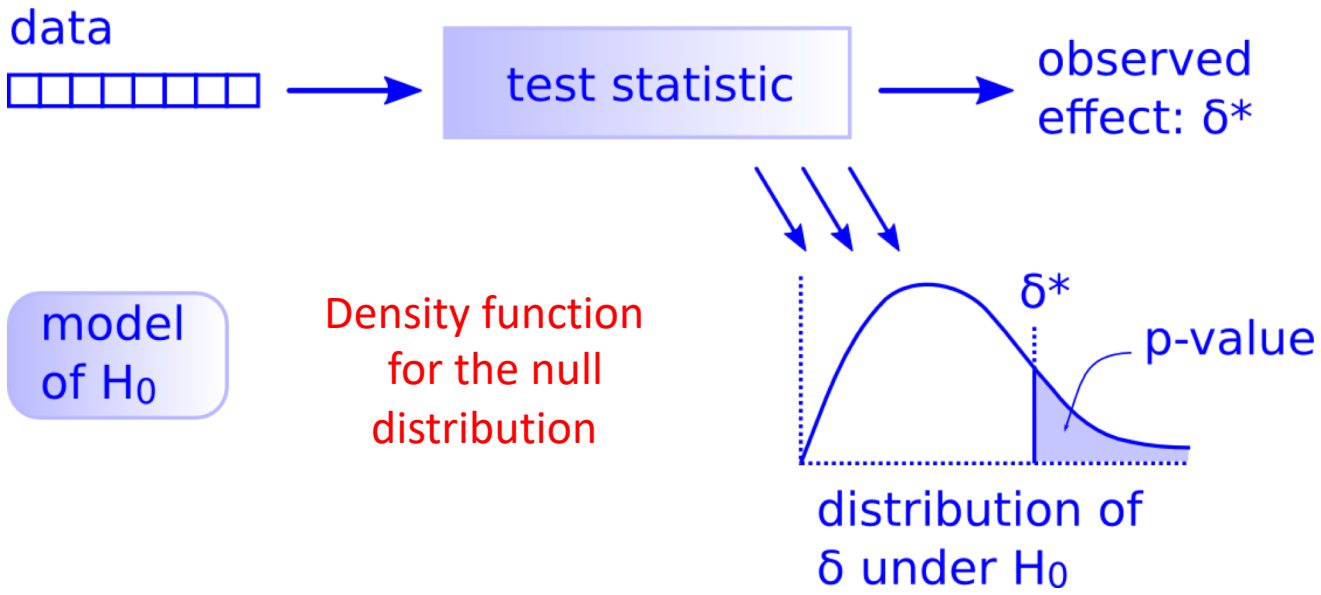
Inference for regression

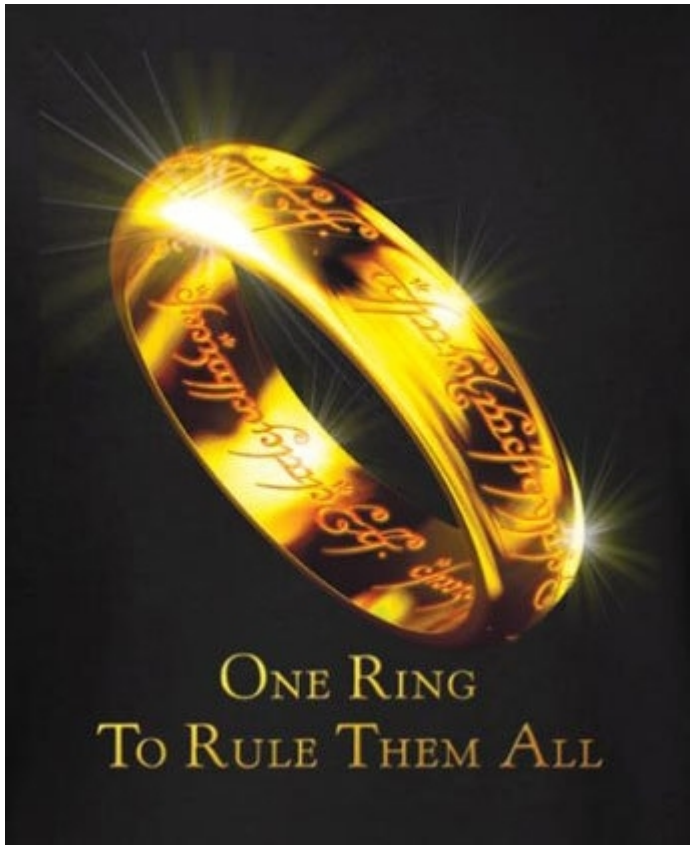
- Resampling methods

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

One test to rule them all

There is only one [hypothesis test](#)!





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

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

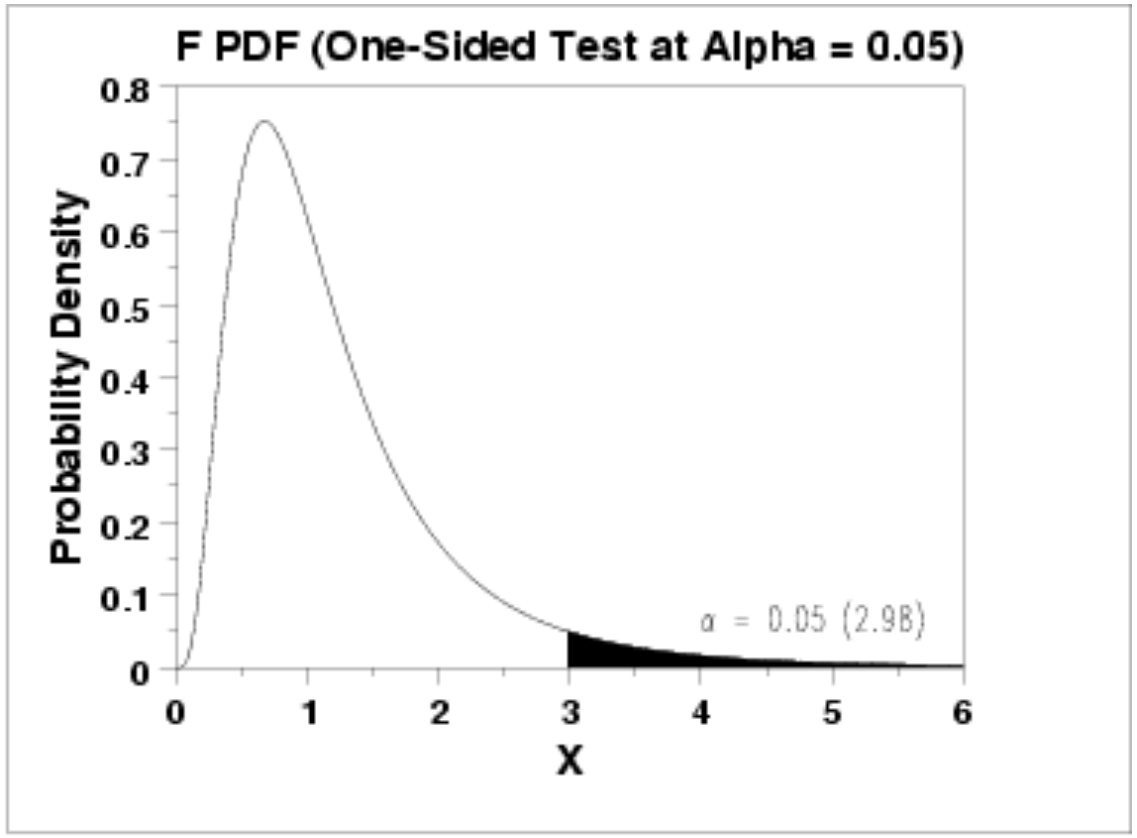
One-way ANOVA – the central idea

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

- $df_1 = K - 1$
- $df_2 = N - K$

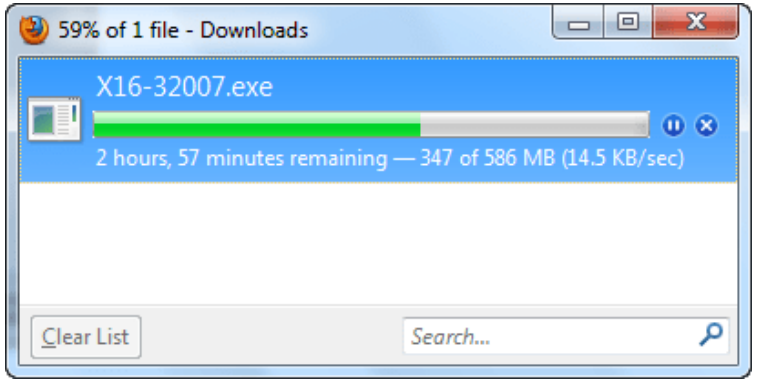
The F-distribution is valid if these conditions are met:

- 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 how much of a difference time of day made on the speed at which he could download files, 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

1. State the null and alternative hypotheses

$\mathbf{H_0}$: $\mu_{7AM} = \mu_{5PM} = \mu_{12AM}$

$\mathbf{H_A}$: $\mu_i \neq \mu_j$ for one pair of the times of day

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

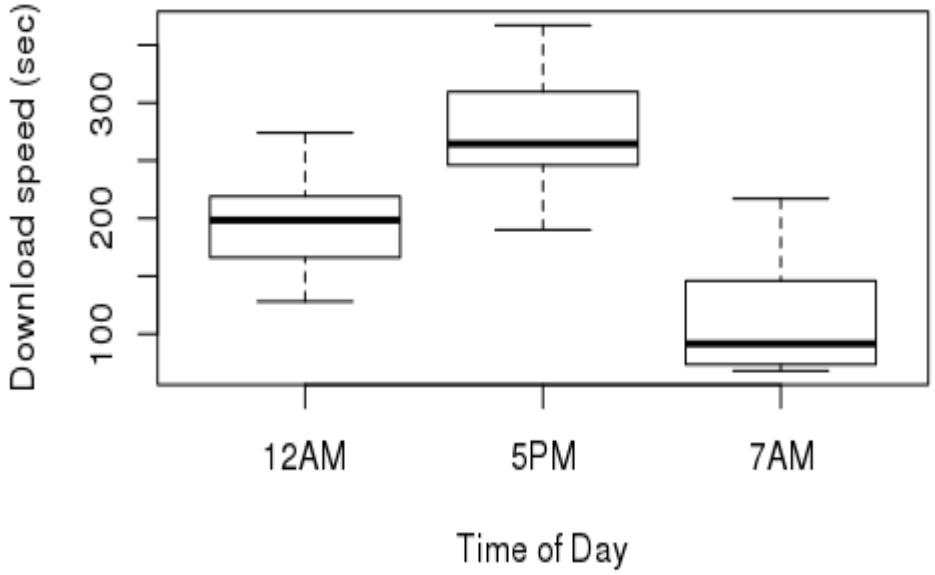
Checking ANOVA conditions ('assumptions')

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

N: total number of points

$\bar{x}_{tot}$: the mean across all the data

$\bar{x}_i$: the mean of group i

n_i: the number of points in group i

x_{ij} : the j^{th} data point from group i

K = 3 different times of day

N = 48 total downloads (16 * 3)

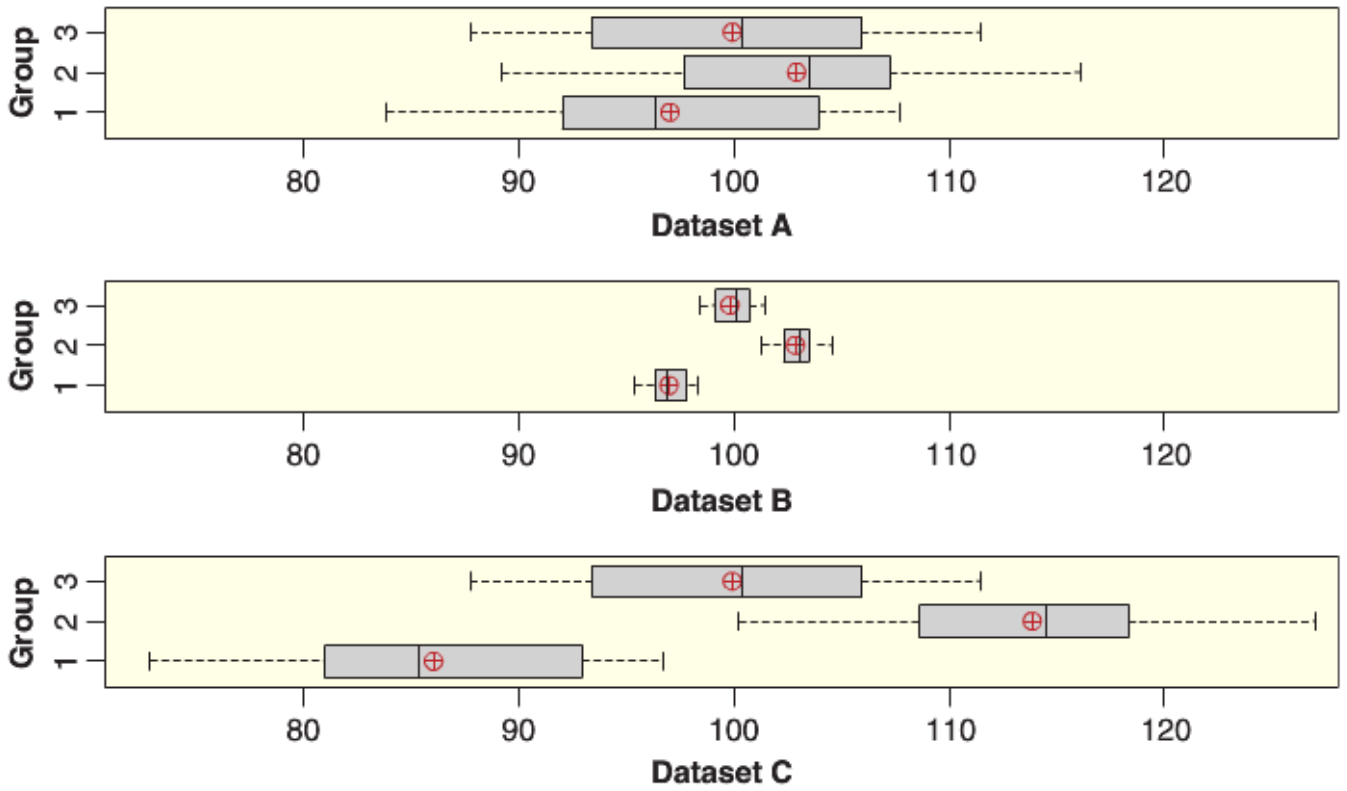
$\bar{x}_{tot}$: the mean speed across all data

$\bar{x}_i$: the means for the i^{th} time of day

n_i = 16 downloads for each time of day

x_{ij} : the j^{th} download at the i^{th} time of day

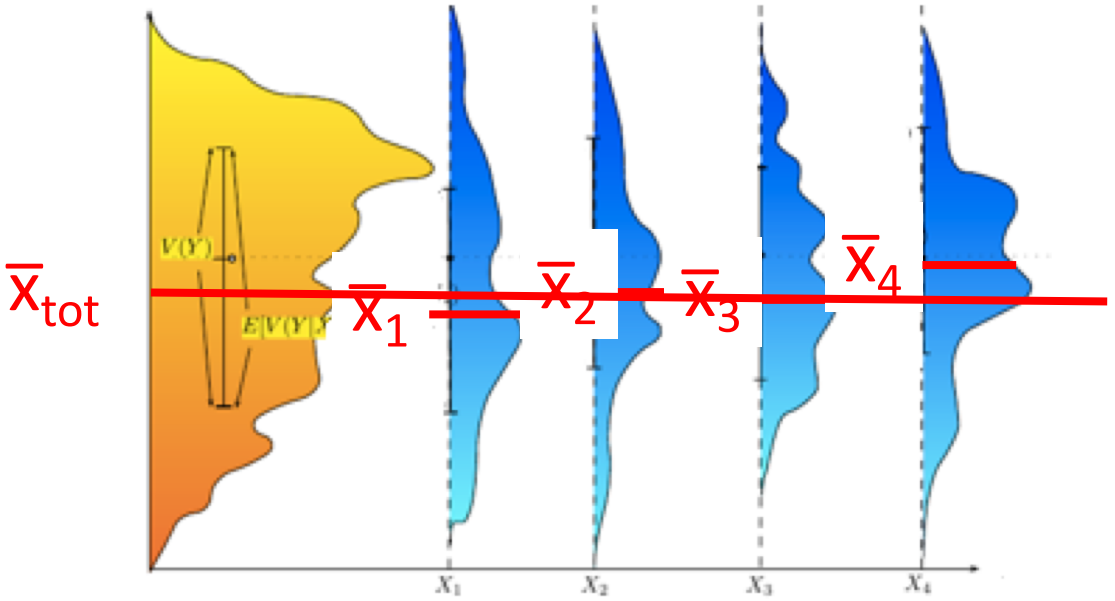
Why use the F-Statistic?



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

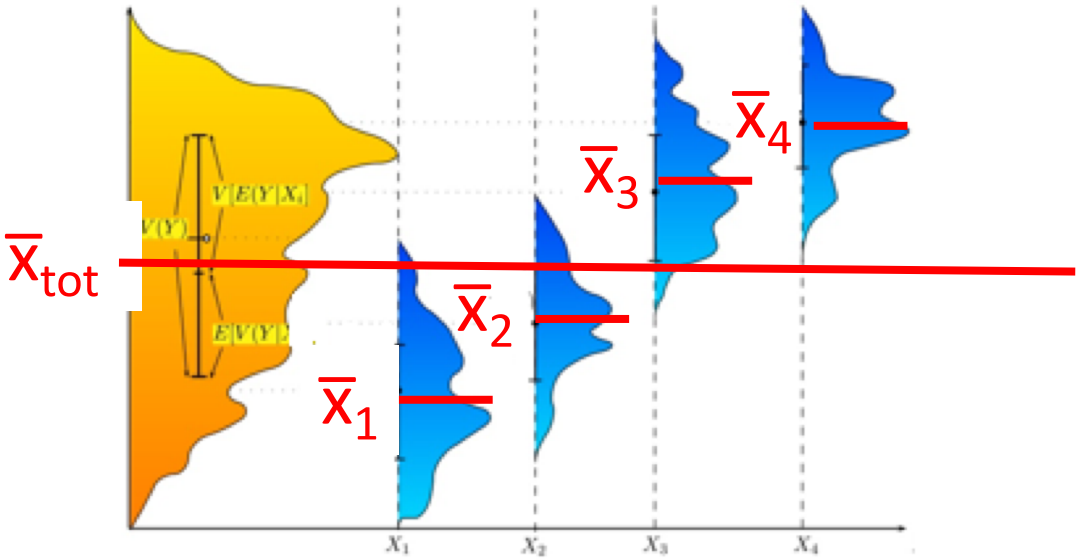
The F-Statistic

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



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

If H_0 is not true, the data from all groups have different means



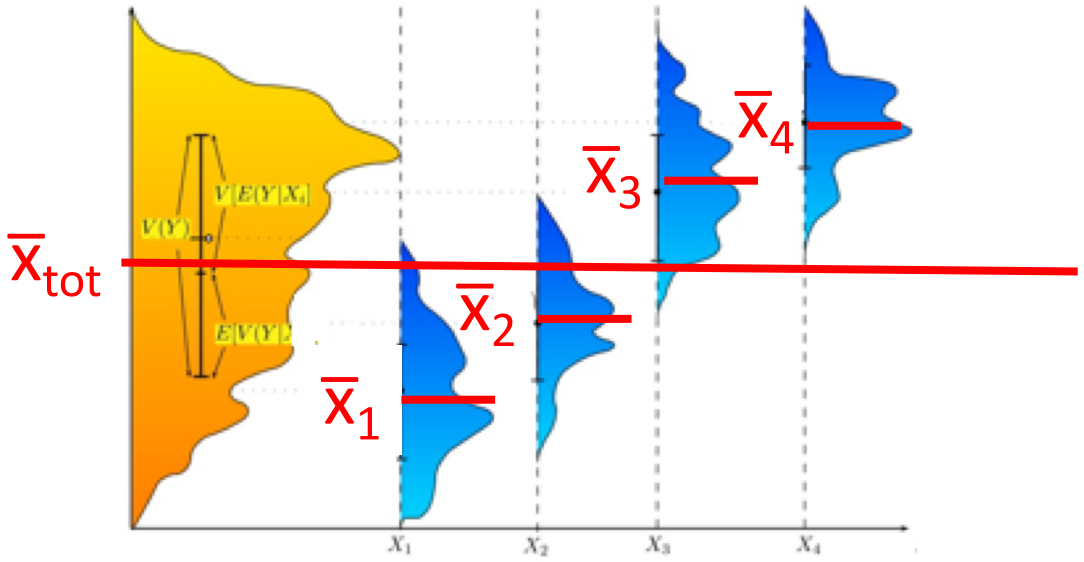
- Different means $\bar{x}_i$
- Smaller spreads s_i

The F-statistic

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

The F statistic measures a fraction of:

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



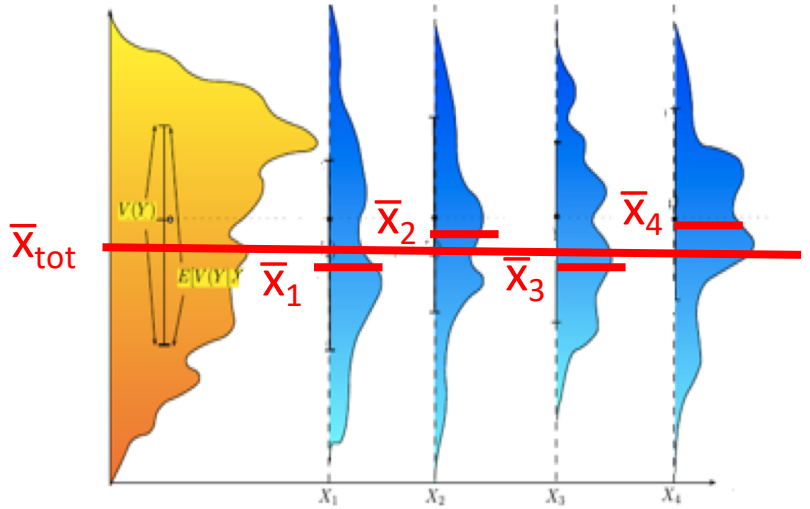
The F-statistic

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

$s_i^2 \approx \sigma^2$

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\approx \sigma^2}$$



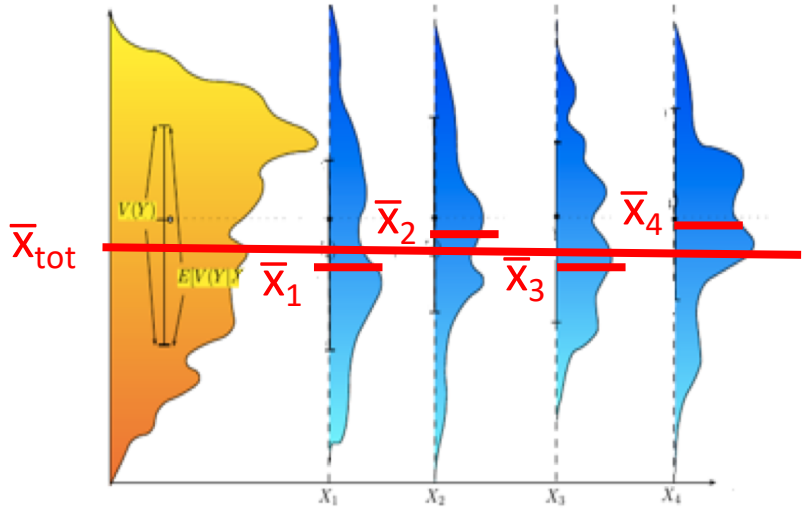
The F-statistic

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

SE^2

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\approx \sigma^2}$$



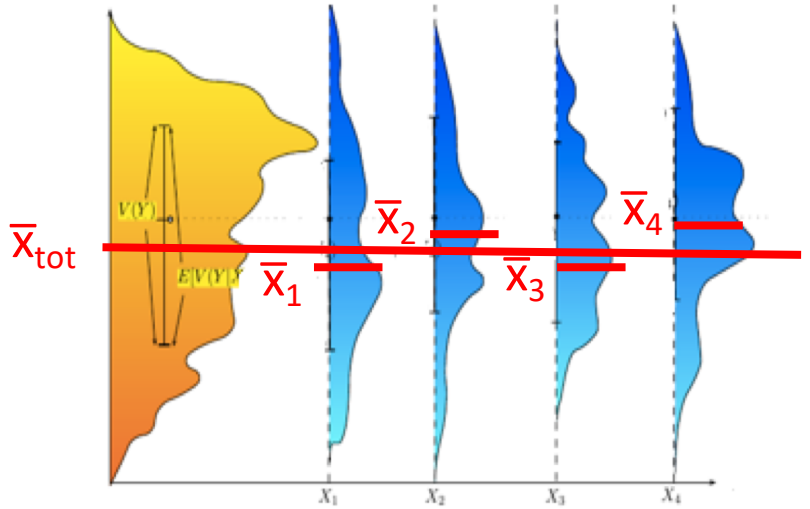
The F-statistic

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

$SE^2 \approx \sigma^2/n$

The F statistic measures a fraction of:

$$F = \frac{\approx \sigma^2}{\approx \sigma^2} \approx 1$$



The F-statistic

Sum of Squares Group (SSG)

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

The F statistic measures a fraction of:

$$F = \frac{\approx \sigma^2}{\approx \sigma^2} \approx 1$$

The F-statistic

Mean Squares Group (MSG)

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

The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\approx \sigma^2} \approx 1$$

The F-statistic

Sum of Squares Error (SSE)

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

The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\approx \sigma^2} \approx 1$$

The F-statistic

Mean of Squares Error (MSE)

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

The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\text{MSE}} \approx 1$$

Interesting fact: SSTotal = SSG + SSE

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	$n-1$	$SSTotal$			

Interesting fact: SSTotal = SSG + SSE

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

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***

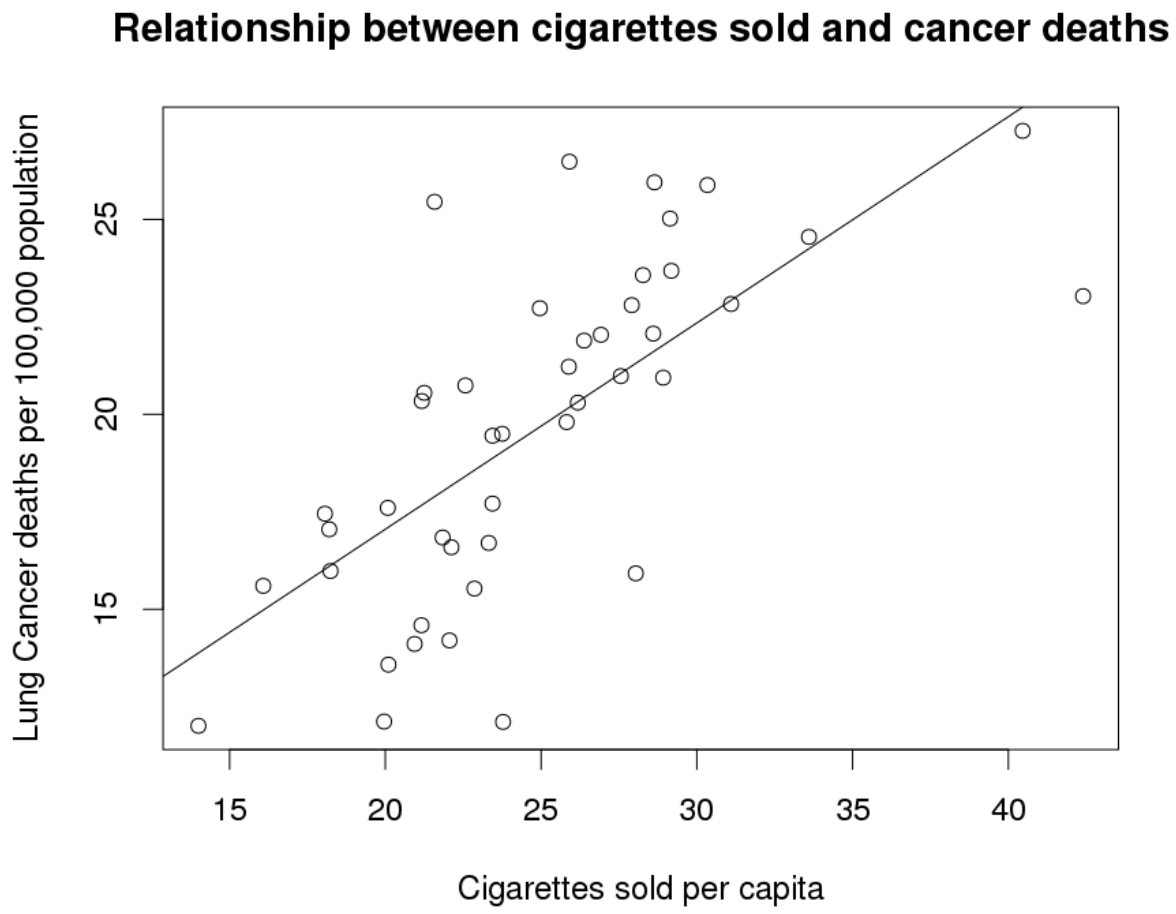
- i.e., $\hat{y} = f(x)$

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)`

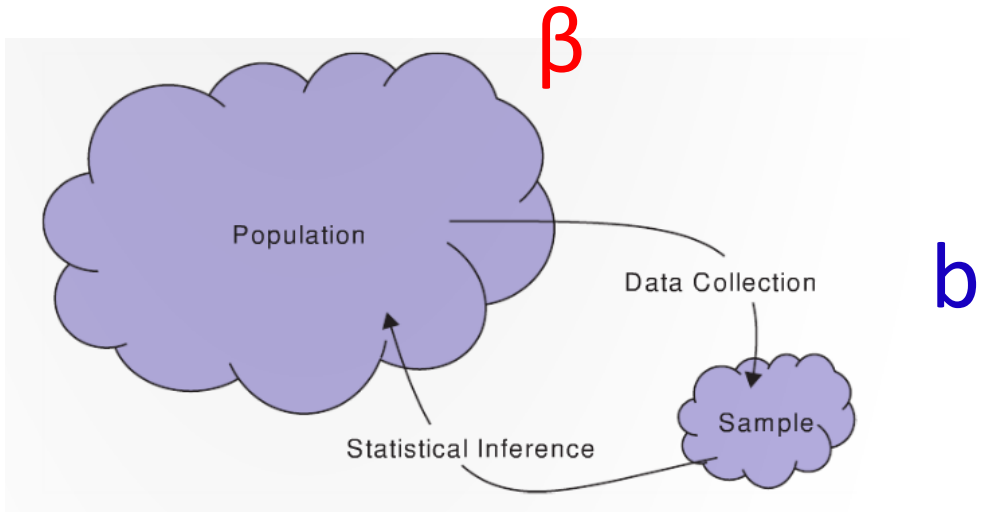
a = 6.47 b = 0.53

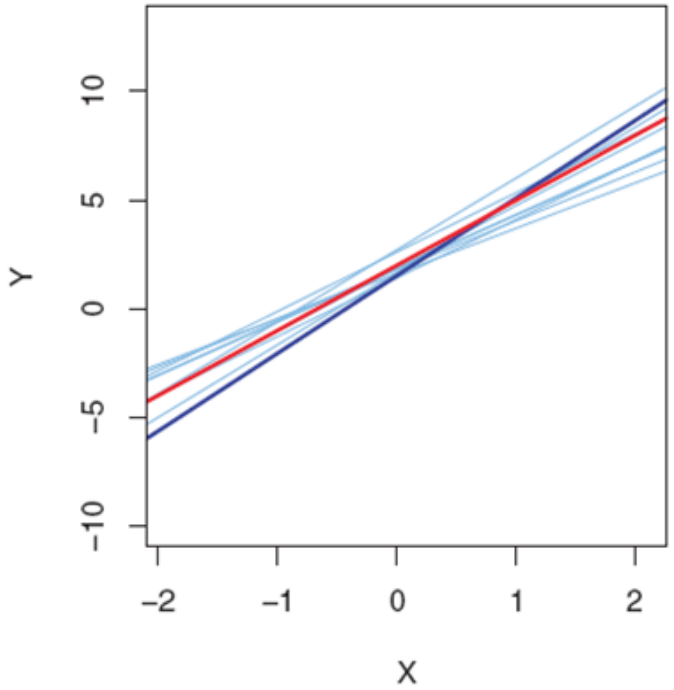
$\hat{y} = 6.47 + .53 \cdot x$

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**





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*
   - ClassTools: 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

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

Let’s try it in R…