Statistical Techniques for Data Science & **Robotics**

Weeks 3


QUIZ
TIME!

6 min. quiz

Question 1: How would you explain a connection between MLE and MSE?

Objectives (for today)

- Recap about interval estimates, confidence interval
- to know about **hypothesis testing** and basic tests
  - t-tests
  - F-tests
  - Goodness of Fit test (GoF test)
- to understand connections to ML via **Decision Theory**

Application to Estimation of Error / Variability

(1) Bernoulli:

  (1) a trial with “0” or “1” outcomes

  (2) p is a parameter (a probability of success)

(2) We know that parameters are $\mu = p; \sigma^2 = p(1-p)$

(3) But can we estimate the p as a function of sample?

  (1) $\hat{p} = ?$…. well, it is just mean,

  (2) and we can find $\hat{\sigma}^2$…

Application of the CLT

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Application of the CLT

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to CLT (approximately) $Z \sim \mathcal{N}(0, 1)$:

$$Z = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

Application of the CLT:

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to CLT (approximately) $Z \sim \mathcal{N}(0,1)$:

$$Z = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

- Derive basic 95% Confidence Interval for p using CLT

$$\bar{X} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \bar{X} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Estimation result

(1) So, just by tossing a coin many (n) times we can build upper and lower bounds for the probability of HEAD:

$$\bar{X} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \bar{X} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

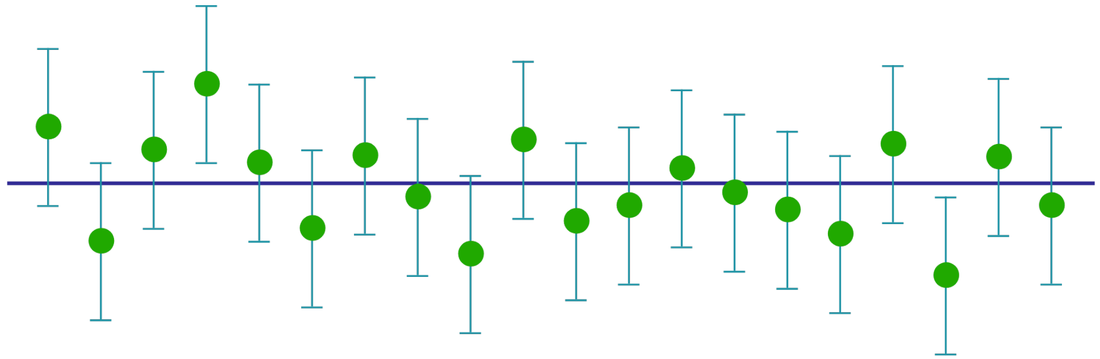
(1)

The lower and upper bounds are **RANDOM VARIABLES** !
(as they depend on a random sample)

Interval Estimates

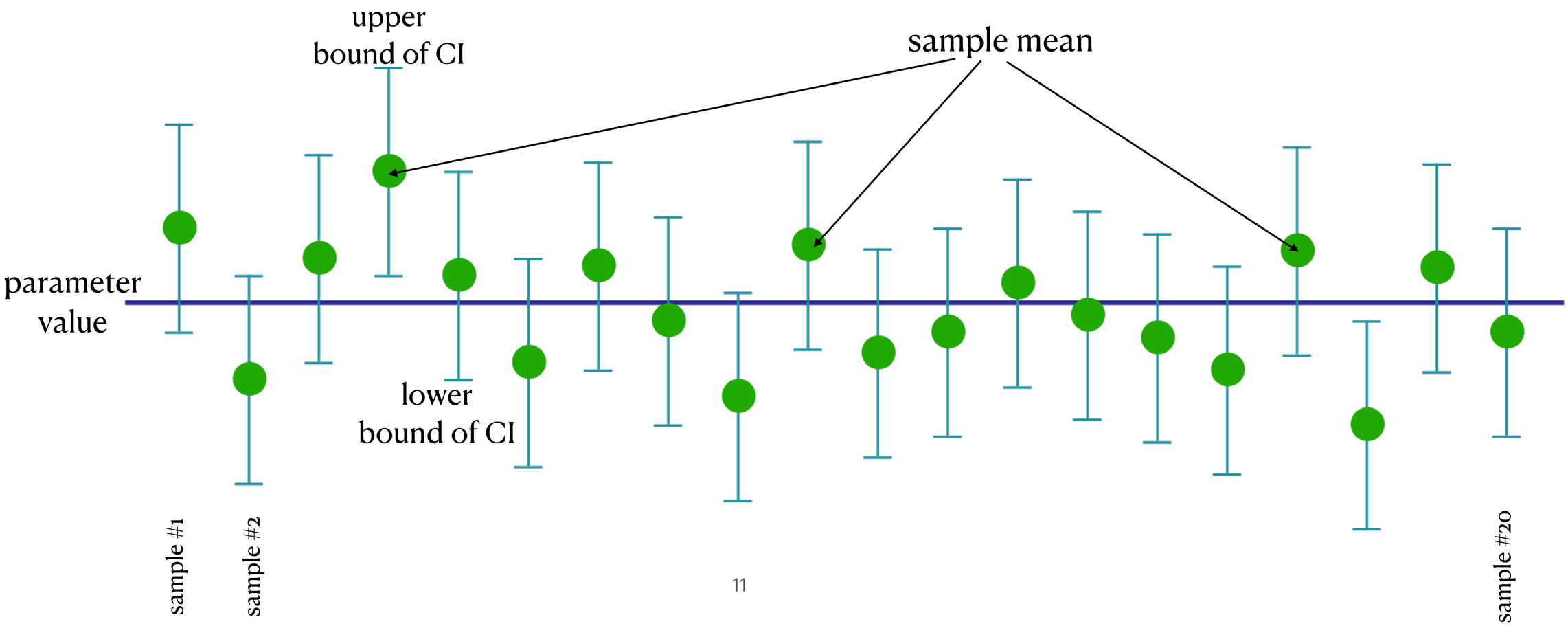
(1) A **confidence interval** contains the true value of the corresponding parameter with the specified probability

(2) Informally, if you run 100 experiments a 95%-confidence interval will contain the value of parameter 95 times (we will discuss it later in the course)

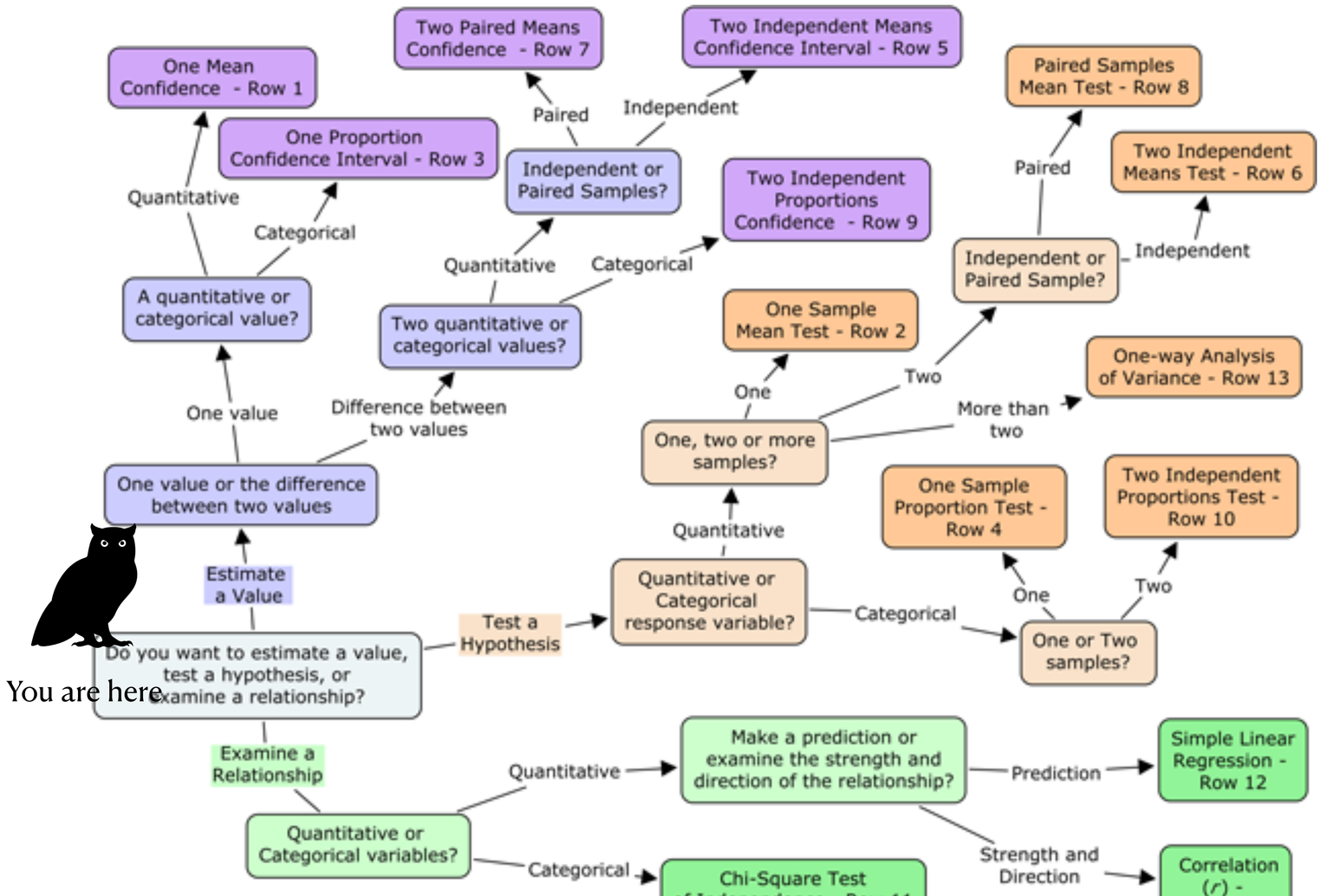


Confidence interval

What is random here? Where is the probability?



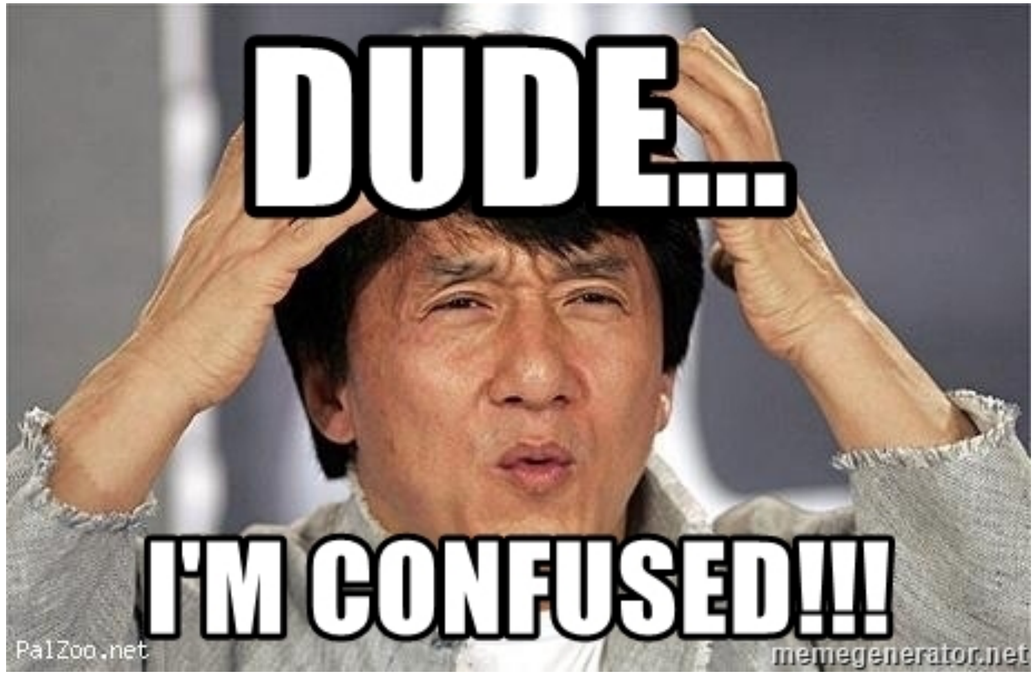
Roadmap of Statistical Tests

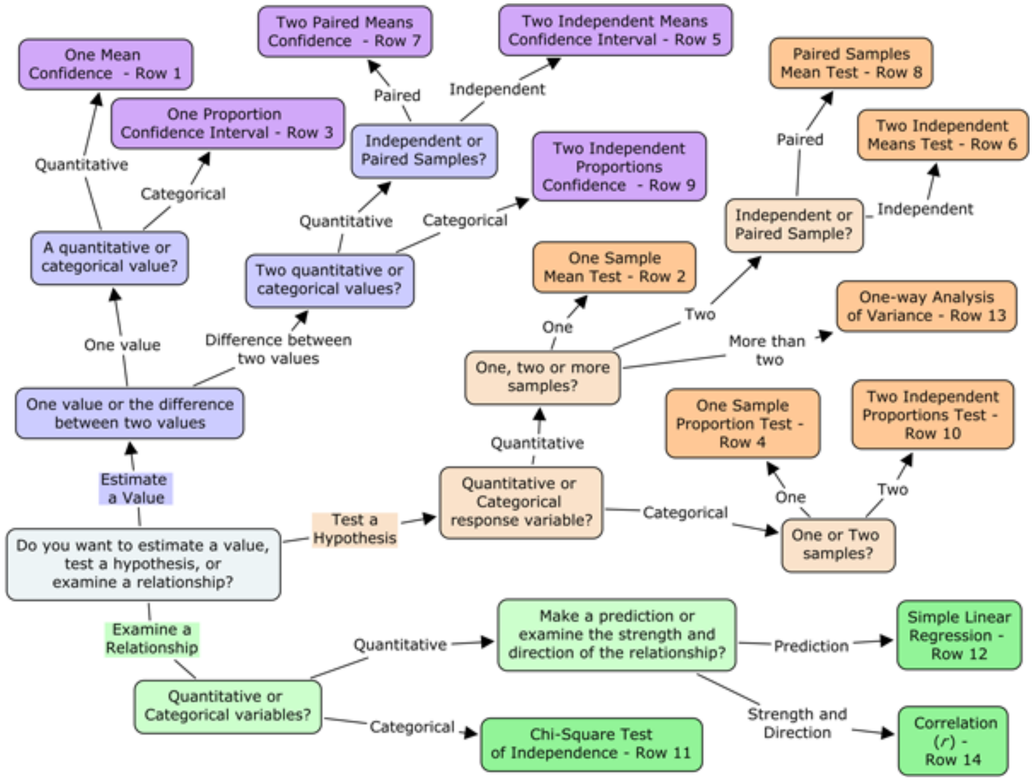


Roadmap :)

to Hell

- https://online.stat.psu.edu/stat500/lesson/12/12.1
- https://online.stat.psu.edu/stat500/lesson/12/12.2





Hypothesis Testing

Hypothesis Testing

Intro

- We assume that **H_o is true** (most likely it is just the `status quo')
- We look for evidence (data) that can help us **reject H_o**
- If we failed to reject **H_o**
  - it does **not** mean that the alternative **H_A** is true or false
- But if can successfully reject **H_o** then its alternative, **H_A**, must be true (no other options)!

- e.g. Either a person has cancer (**H_A**) or not (**H_o**). What would be the data?

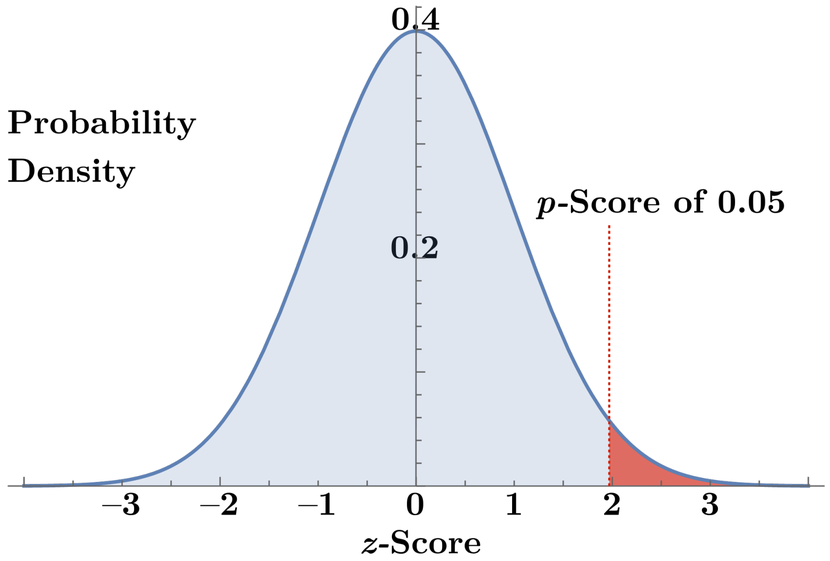
Hypothesis Testing

p-value

- So, having the **data** (result of a test, or a value of a statistic) we give the $\mathbf{H_o}$ a try:
- Based on the null hypothesis we calculate a probability of observing the data

(if $\mathbf{H_o}$ was actually true, or under the null hypothesis)

- p-value = size of the red area
- if p-value is low enough, then the data provides evidence to reject H_0



Type I and Type II errors

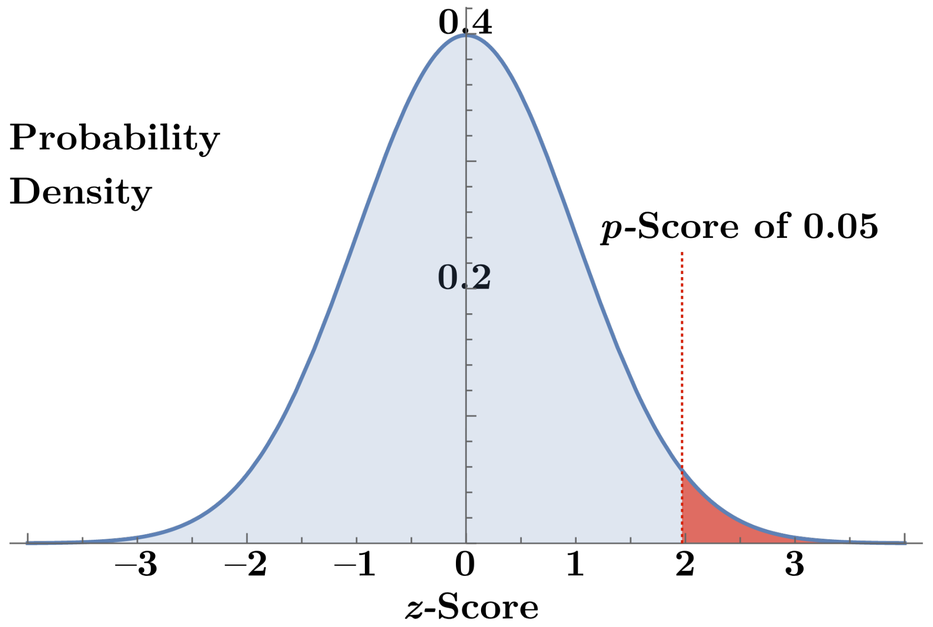
- If we decide that we reject the null hypothesis $\mathbf{H_o}$,
  (and it happend to be true)

e.g. we have a positive test for cancer, when a persion is healthy

What should we expect to happen

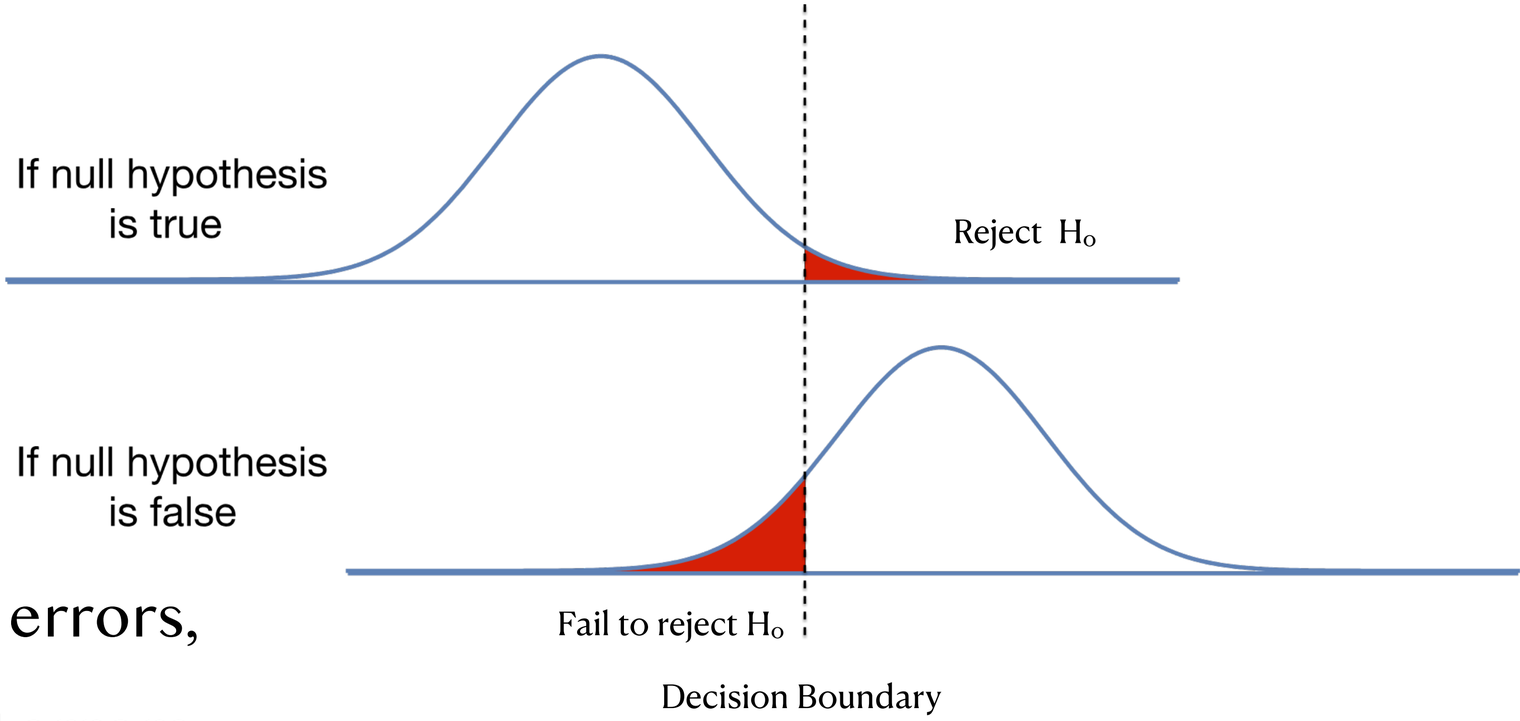
approximately 5% of the time?

- a false positive
- a false negative
- a correct conclusion



Type I and Type II errors

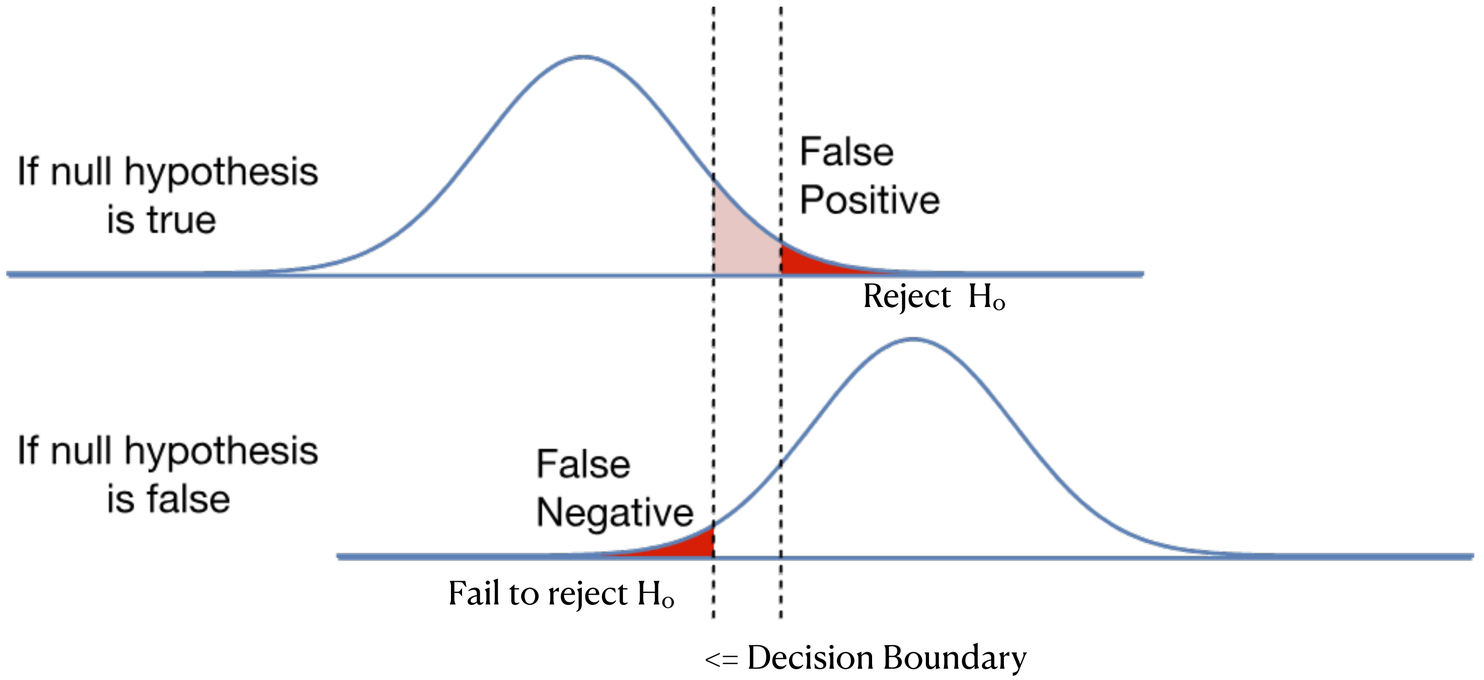
- **Type I**: the null hypothesis is **true**, but we **reject** it.
- **Type II**: the null hypothesis is **false**, but we **fail to reject** it



- Errors are marked in red;
  - the upper graph shows Type I errors,
  - the lower graph shows Type II errors

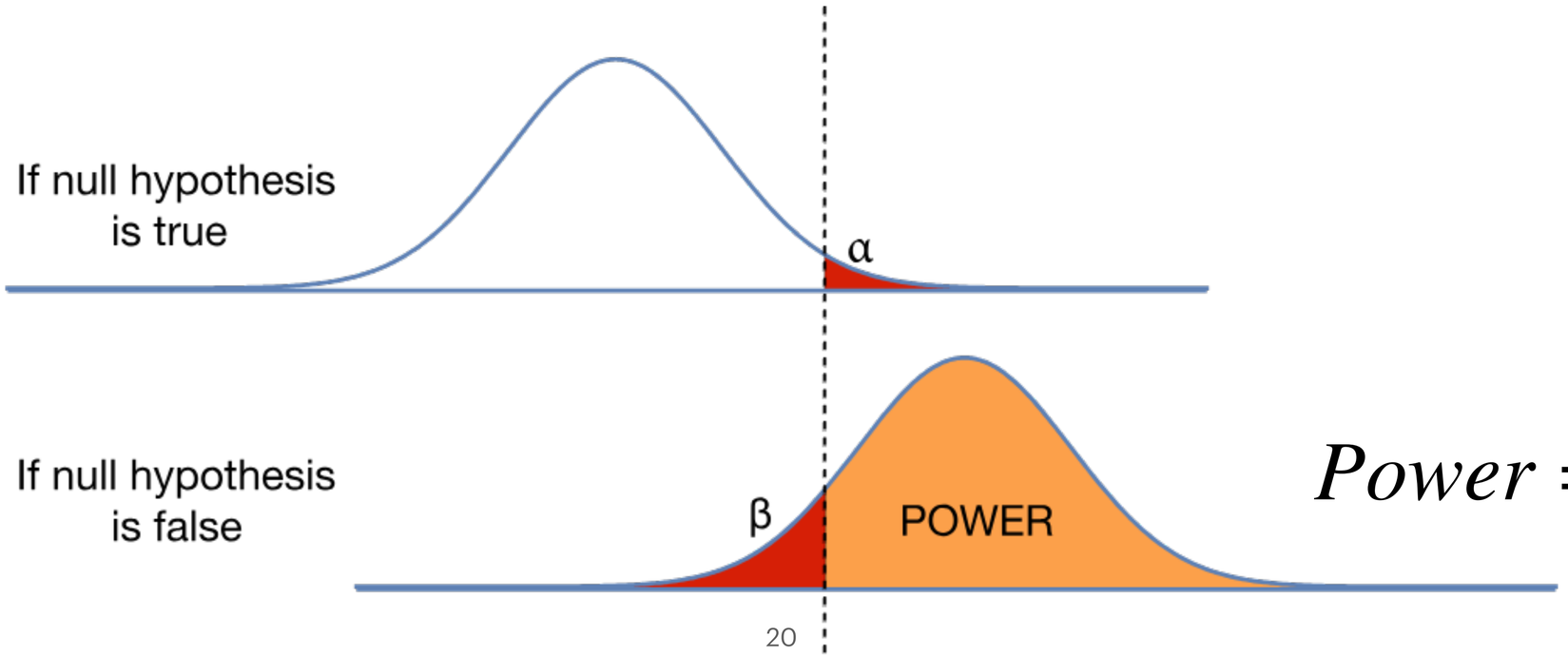
Quick check

- In medical screening, it is sometimes more important to avoid false negatives than false positives. (i. e., we want to be sure to detect a disease if present.)
- How would this affect the p-value we might use for a medical test?



Power of a test

- The "power" of a test is the probability that it will be able to find the effect it is looking for, assuming that the effect is there



$$Power = 1 - \beta$$

Statistical Tests

A/B Testing

- **Basic scenarios:**
  - Testing two news headlines to determine which generates more clicks
  - Testing two prices to determine which yields more net profit in USD
  - Testing two web ads to determine which generates more conversions (number of purchases)
  - Testing two products on their popularity among customers

Practical Problem

- All **subjects** assigned to one of two groups **randomly:**
  - group A **got drug A**;
  - group B is a **control group**

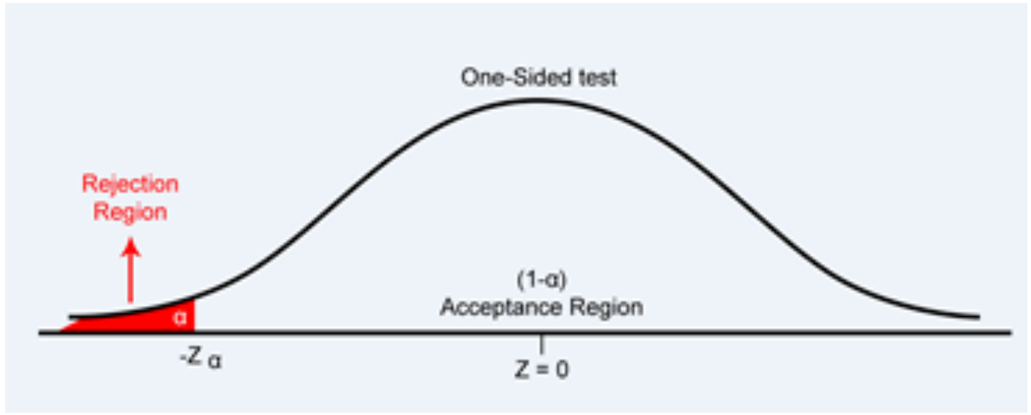
- Test statistic, or metric to compare the 2 groups can be:
  - binary, categorical
  - continuous

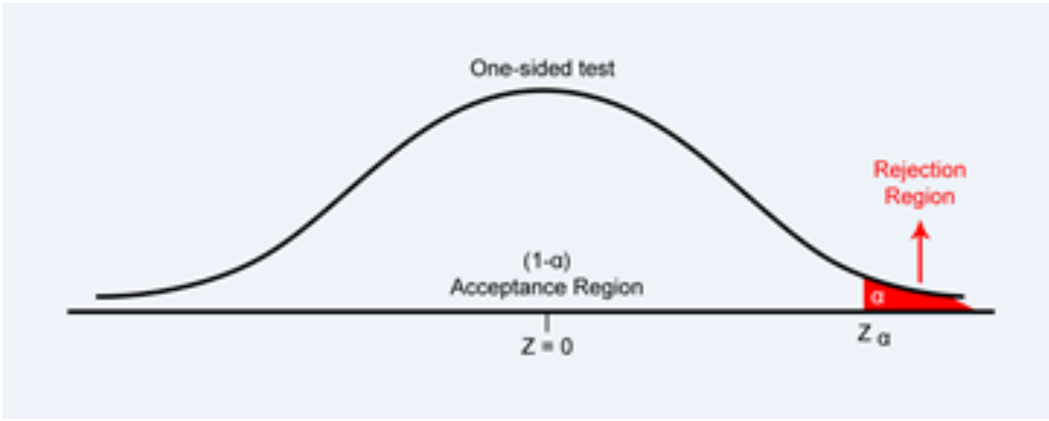
Warning #1. Set up experiment before you collect data

- You should decide on the setting of the experiment BEFORE you run the experiment
- Especially, you should define the **test statistic** AHEAD of time
- You may collect any additional metrics, but the final decision depends on the predefined **test statistic**

One-tailed test

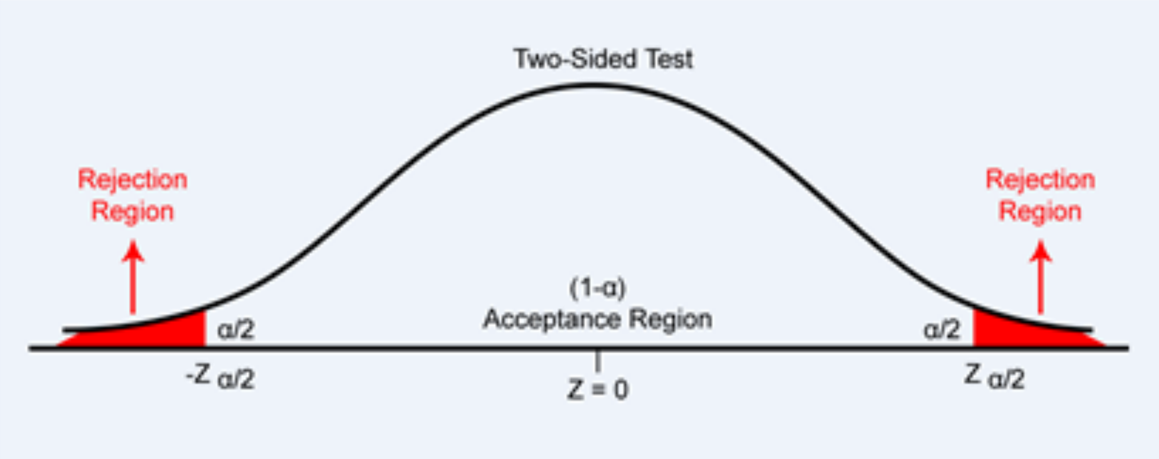
- One-way test is a typical setting if you use A/B testing
  - because you test that something (A) **is better** than control (B)
- It is one-way test:
  - a value of some test statistic in group A is **significantly higher** (or lower) than in group B





Two-tailed, or two-sided test

- Two way is a typical setting
  - if you do not want to be fooled by a baseline (null) chance model in either direction
- The two-way test:
  - a value of some test statistic in group A **is significantly different** from that in group B



Example 1

- **TESTS ABOUT ONE MEAN**
  - simple null hypothesis $\mathbf{H_o}$: $\mu = 50$
    against the **simple** alternative hypothesis $\mathbf{H_1}$: $\mu = 100$.
  - If the alternative hypothesis had been H1: $\mu > 50$, it would be a **composite** hypothesis

Example 2

TESTS OF THE EQUALITY OF TWO MEANS

- A botanist would like
- to test the null hypothesis
  against the alternative hypothesis

$\mathbf{H_o}: \mu_X - \mu_Y = 0$

$\mathbf{H_1}: \mu_X - \mu_Y < 0.$

Example 3

- **TESTS ABOUT PROPORTIONS** (p is a proportion)
  - a frequently used procedure for testing $\mathbf{H_o}$: $p = p_0$,
  - where p_0 is some specified probability of success.

  - Consider the test of $\mathbf{H_o}$: $p = p_0$ against $\mathbf{H_1}$: $p > p_0$ that rejects $\mathbf{H_o}$ and accepts $\mathbf{H_1}$ if and only if
  - $$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_\alpha$$

t-Tests

William Sealy Gosset

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

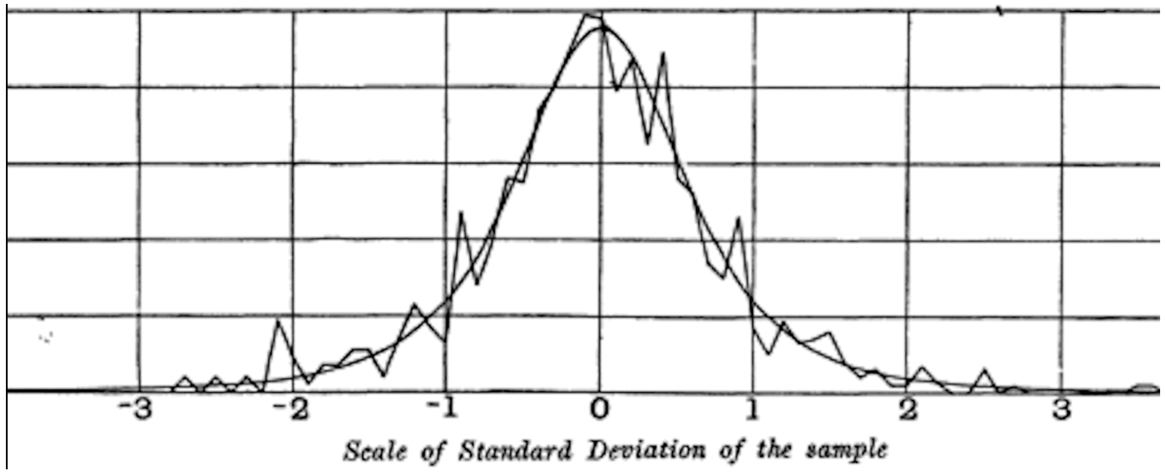
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

osset's 1908 article describing the "Student's *t* test." (Reproduced with permission

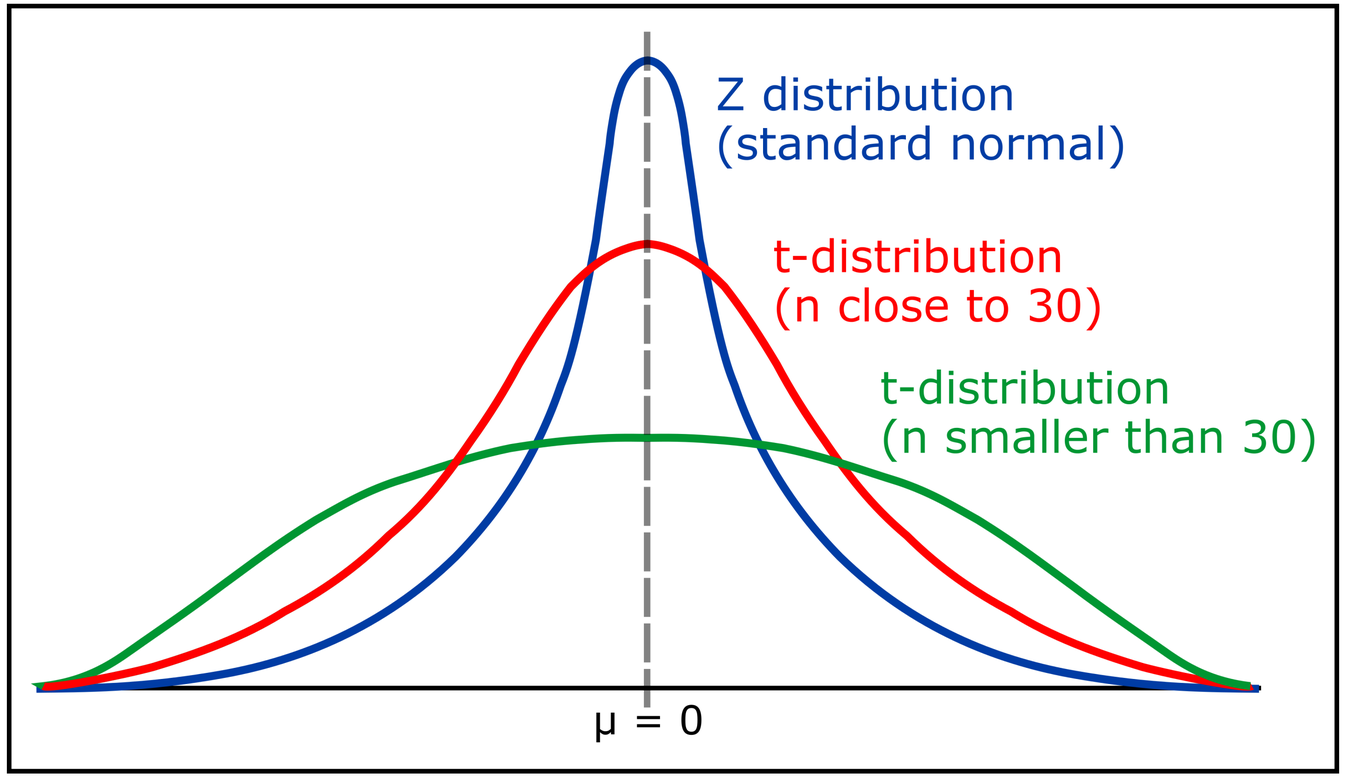
Setting

- Gosset wanted to answer the question:
  - “What is the sampling distribution of the mean of a sample, drawn from a larger population?”
- He started out with a resampling experiment drawing random samples of 4 from a data set of 3,000 measurements
- He plotted the standardised results (the z scores) on the x-axis and the frequency on the y-axis.



Student's t-Distribution

- How this distribution (of a statistic) arises from sampling?



Example of t-Test

- Test for difference between two sample means

```
SAMPLE 1:
     NUMBER OF OBSERVATIONS          = 249
     MEAN                            =  20.14458
     STANDARD DEVIATION              =   6.41470
     STANDARD ERROR OF THE MEAN      =   0.40652

SAMPLE 2:
     NUMBER OF OBSERVATIONS          = 79
     MEAN                            = 30.48101
     STANDARD DEVIATION              =  6.10771
     STANDARD ERROR OF THE MEAN      =  0.68717
```

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$T = -12.62$$

F-tests, ANOVA

Analysis of variance (ANOVA) uses F-tests to statistically assess the equality of means when you have **three or more groups**

One-Way ANOVA (ANalysis Of VAriance)

- Question: How to determine whether a set of means are all equal.
- Instead of an A/B test, we had a comparison of multiple groups, say A-B-C-D, each with numeric data.
- Tests for a statistically significant **difference among the groups is called ANOVA**.

- Whether the **in-group variability** is significantly different from **between-group variability**

ANOVA (ANalysis Of VAriance). Example

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand Avg.				173.75

Single overall omnibus test:

- are there any differences between 4 (groups)?

Decomposition of variance

- Value = Grand Avg. + Group effect + residual

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand Avg.				173.75

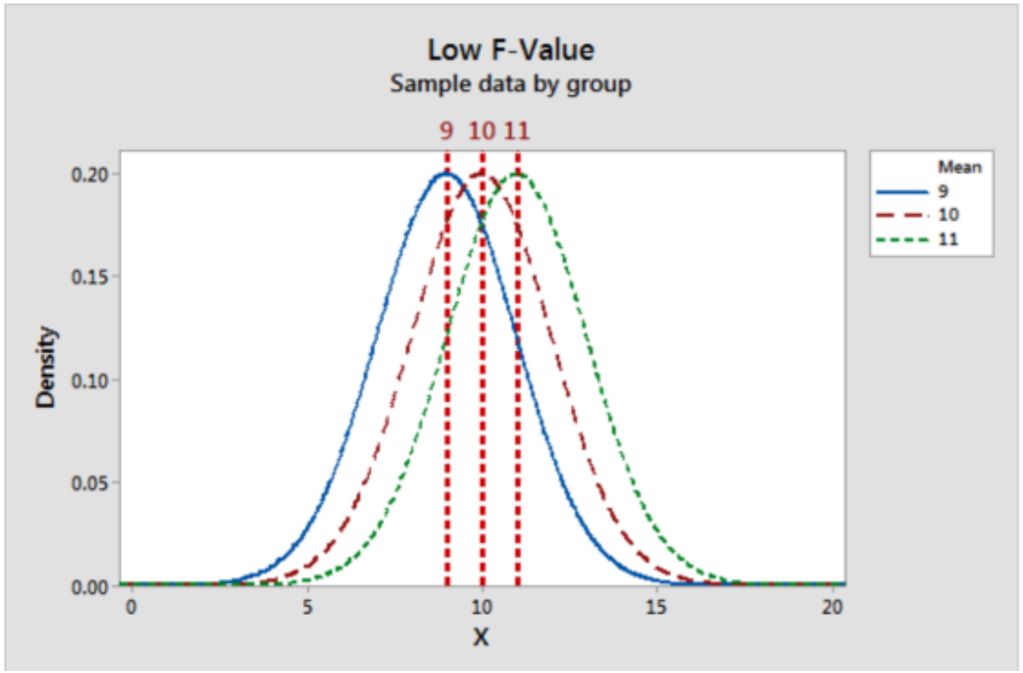
- **164 = 173.75 + (172 - 173.75) + (164 - 172)**

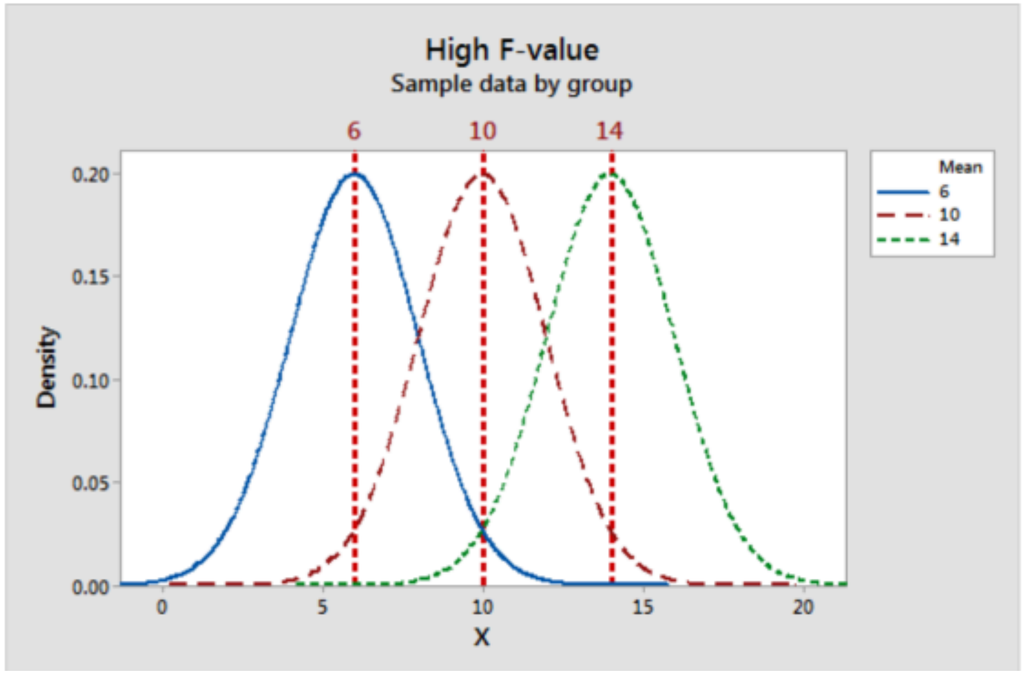
F-statistic

- ANOVA is based on the F-statistic
- The F-statistic
  - the ratio: variance across group means to the variance due to residual error
  - The higher this ratio, the **more statistically significant the result**.

- If the data follows a normal distribution, then statistical theory dictates that the statistic should have a certain distribution (F-distribution).
- So, we can compute **p-value**

F-statistic

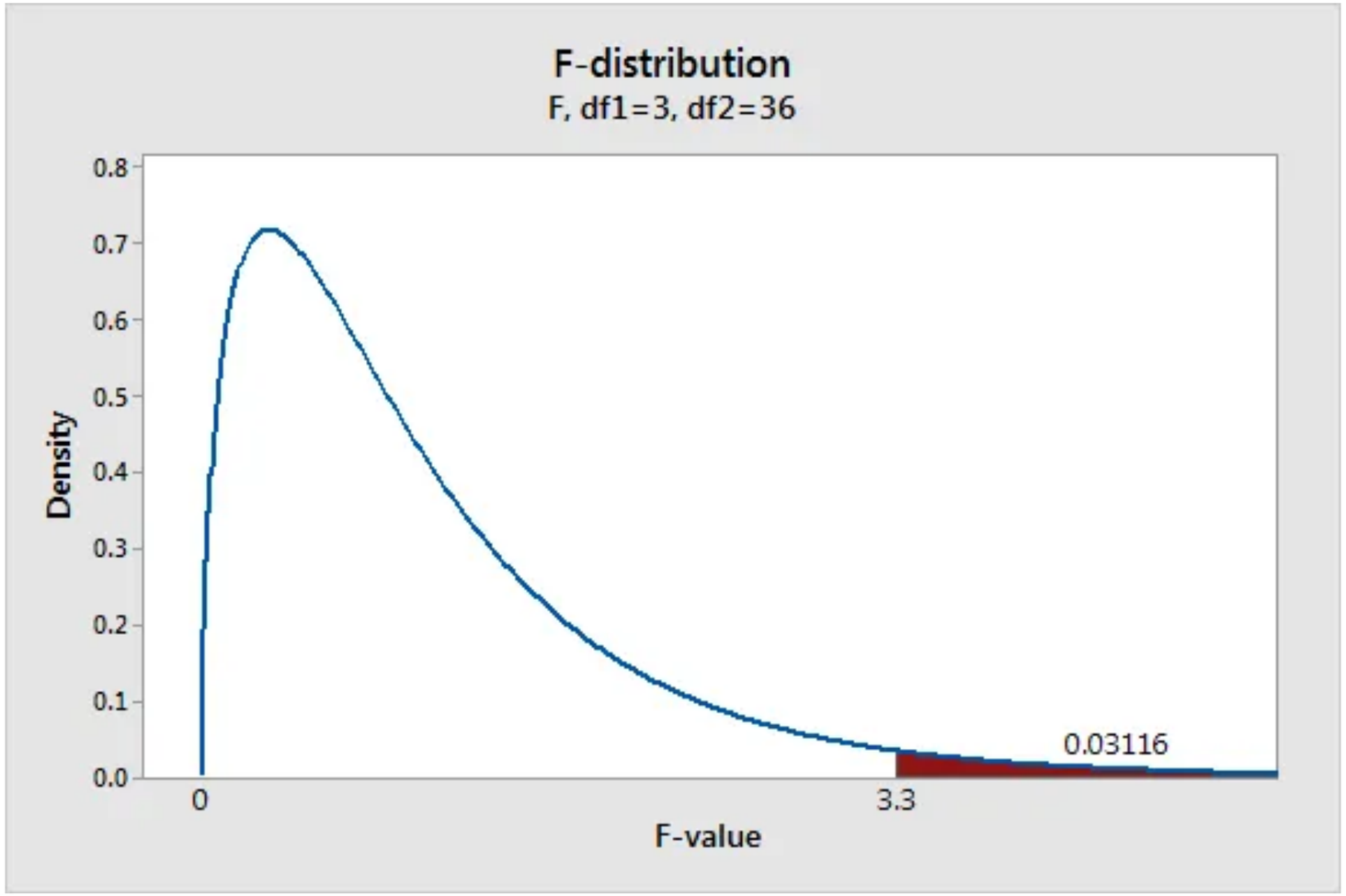




$$F = \frac{'between - groups' \ variance}{'within - group' \ variance}$$

- https://statisticsbyjim.com/anova/f-tests-anova/

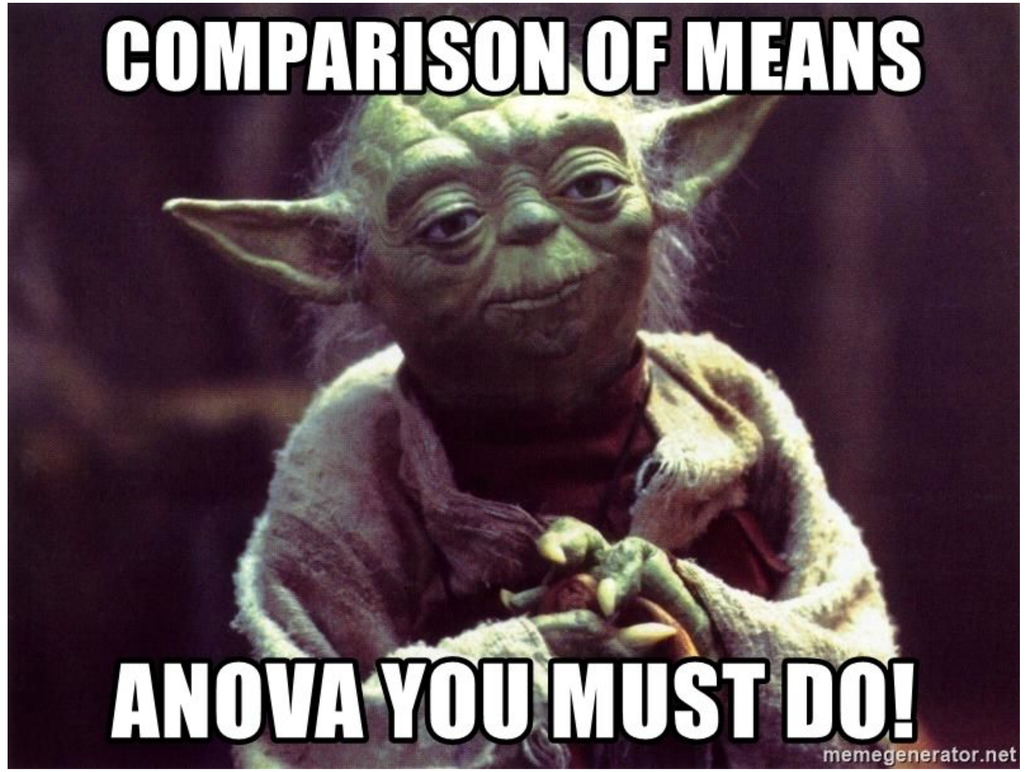
F-statistic


F-distribution
F, df1=3, df2=36
Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.03116
0
3.3
F-value

F-test and ANOVA summary

- ANOVA and F-tests assess the amount of variability between the group means in the context of the variation within groups to determine whether the mean differences are statistically significant.

- ANOVA results indicate that not all means are equal, it doesn't identify which particular differences between pairs of means are significant.



Chi-square Test, χ^2

Chi-square Goodness of Fit Test

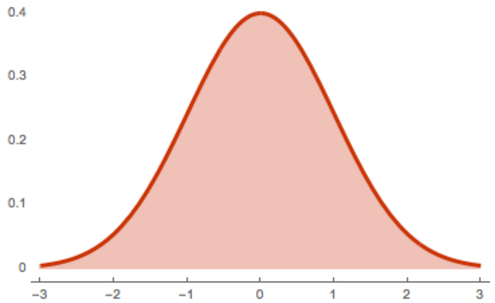
Karl Pearson

Goodness of Fit = GoF

Compares sample* distribution to some target theoretical distribution

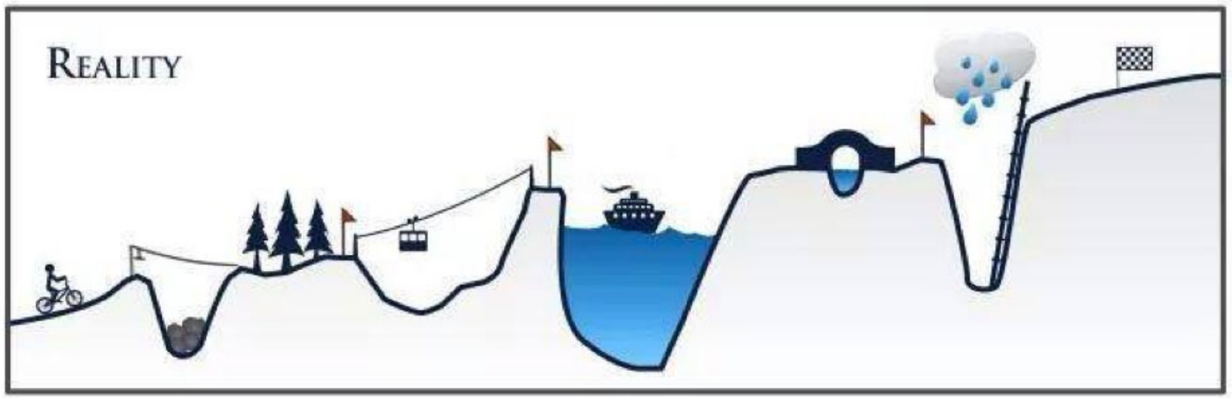
*ONE SAMPLE TEST!

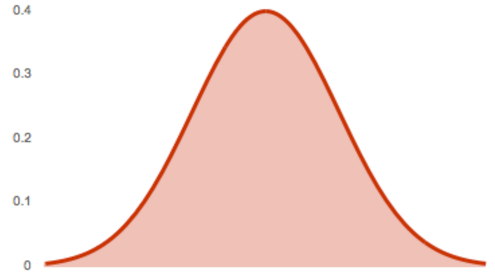
Intuition

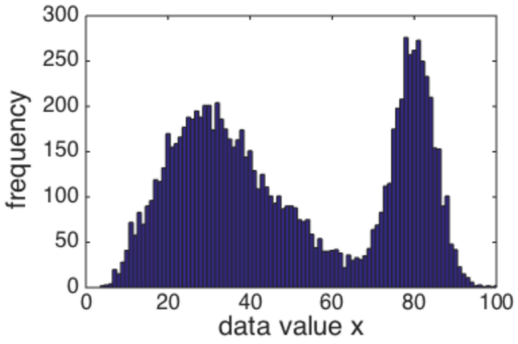


Intuition


YOUR PLAN


REALITY


0.4
0.3
0.2
0.1
0


300
250
200
150
100
50
0
frequency
0
20
40
60
80
100
data value x

Problem

- Given a sample you need to check if this sample came from the a specific distribution (Normal, Poisson,...).

So, we have a sample…

- Sample data fall into intervals and we basically analyze **categorical data**!!!

- the numbers of points that fall into an interval are compared with the expected numbers of points in this interval.

But, what about hypotheses?

Null hypothesis:

In Chi-Square GoF test, the null hypothesis assumes that there is **no significant difference** between the observed and the expected value.

Alternative hypothesis:

In Chi-Square GoF test, the alternative hypothesis assumes that there is **a significant difference** between the observed and the expected value.

Yep, but what about the statistic?

- The chi-square GoF test
- uses χ^2 statistic that is calculated as follows:
- $$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - exp_i)^2}{exp_i}$$
- then it is compared to the statistics values from the table and make a decision.

Example

Candies of different flavors

Observations

we take 10 bags 20 candies in each

Flavor	Number of Pieces of Candy (10 bags)
Apple	180
Lime	250
Cherry	120
Orange	225
Grape	225

Hypothesis

uniform distr. is expected

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy
Apple	180	200
Lime	250	200
Cherry	120	200
Orange	225	200
Grape	225	200

calculations

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected
Apple	180	200	180-200 = -20
Lime	250	200	250-200 = 50
Cherry	120	200	120-200 = -80
Orange	225	200	225-200 = 25
Grape	225	200	225-200 = 25

calculations

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference
Apple	180	200	180-200 = -20	400
Lime	250	200	250-200 = 50	2500
Cherry	120	200	120-200 = -80	6400
Orange	225	200	225-200 = 25	625
Grape	225	200	225-200 = 25	625

Chi-square statistics

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference	Squared Difference / Expected Number
Apple	180	200	180-200 = -20	400	400 / 200 = 2
Lime	250	200	250-200 = 50	2500	2500 / 200 = 12.5
Cherry	120	200	120-200 = -80	6400	6400 / 200 = 32
Orange	225	200	225-200 = 25	625	625 / 200 = 3.125
Grape	225	200	225-200 = 25	625	625 / 200 = 3.125

- We calculate a test statistic. Our test statistic is 52.75

Chi-square statistics

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference	Squared Difference / Expected Number
Apple	180	200	180-200 = -20	400	400 / 200 = 2
Lime	250	200	250-200 = 50	2500	2500 / 200 = 12.5
Cherry	120	200	120-200 = -80	6400	6400 / 200 = 32
Orange	225	200	225-200 = 25	625	625 / 200 = 3.125
Grape	225	200	225-200 = 25	625	625 / 200 = 3.125

- We calculate a test statistic. Our test statistic is 52.75
- Degrees of freedom k-1 (k = number of bins, flavors): df = 4
- p-value:

http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pchisq.html

PAUSE, 5-7 min.

ask questions, please

Decision Theory

Introduction to Decision Theory

- Decision Theory
  - helps us making optimal decisions
  - in case of uncertainty (that is modeled with Prob.Theory)

- **First step**: to infer (or learn) the p(x,t) - a joint distribution of inputs and targets
- **Second step**: use probabilities to make optimal decisions

Example: classification with 2 classes

a medical diagnosis problem

- an X-ray image of a patient,
- we wish to determine whether the patient has cancer or not
  - => we need a decision rule.
- Setup:
  - $\mathbf{x}$ is an input (a vector);
  - $t \in \{0,1\}$;
  - $t = 0 \implies C_1$;
  - $t = 1 \implies C_2$

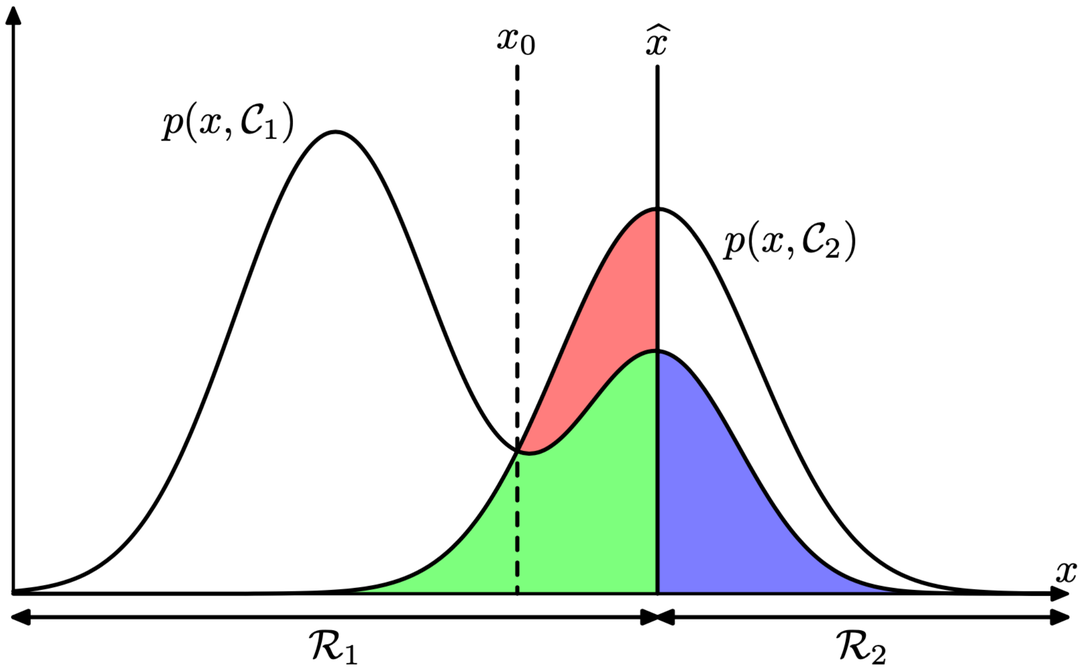
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

k=1,2

Why is the p(x,t) not present here?

Decision regions and boundaries

- R_1, R_2 are decision regions with assignments of **x** to C_1, C_2



- Is $\hat{x}$ the optimal decision boundary?

Minimizing errors of classification

- our rule should minimize:

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$

- R_1, R_2 are regions with **wrong** assignments of $\mathbf{x}$ to C_2, C_1

Inference and decision

3 **distinct** approaches to solving decision problems

- Determine **class-conditional densities** $p(x \mid C_k)$ for each class C_k, then make a decision for the given x

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

  - We get a **generative model**
- Determine **posterior class probabilities** $p(C_k \mid x)$ and predict class for an x
  - This is a **discriminative model**
- Find a **discriminant** function, $f(x)$, that maps x to class label (e.g. a linear clasifier)

Summary