SOME STATISTICS for DATA SCIENCE Nov 1st '21

Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
  - We'll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
  - These are the techniques we'll leverage for Machine Learning and Prediction

Examples of Business Questions

- **Simple (descriptive) Stats**
  - "Who are the most profitable customers?"
- **Hypothesis Testing**
  - "Is there a difference in value to the company of these customers?"
- **Segmentation/Classification**
  - What are the common characteristics of these customers?
- **Prediction**
  - Will this new customer become a profitable customer? If so, how profitable?

Distributions

- Normal
- Binomial
- Poisson

Normal Distributions, Mean, Variance

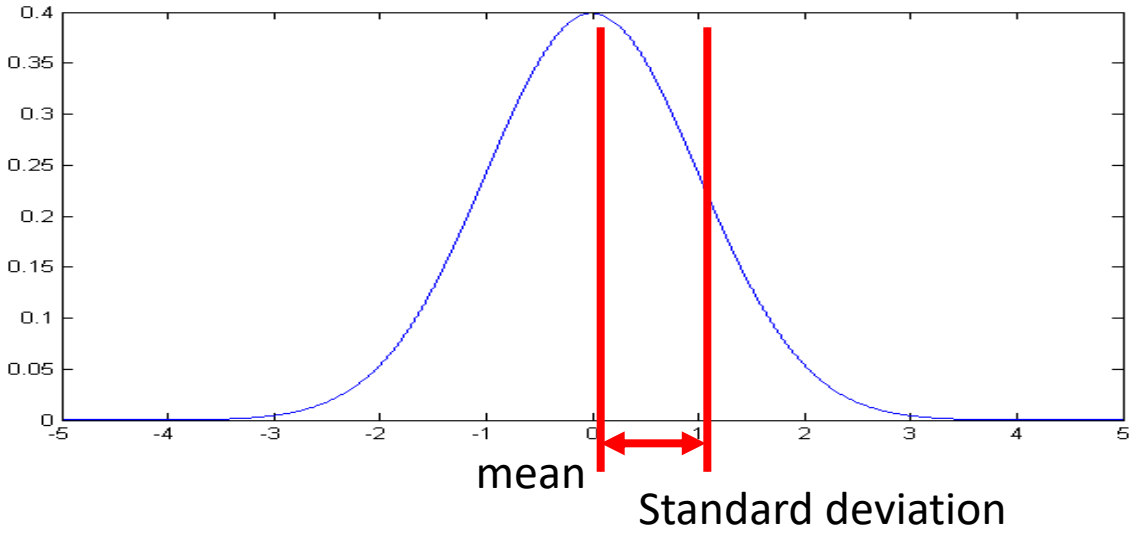
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

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

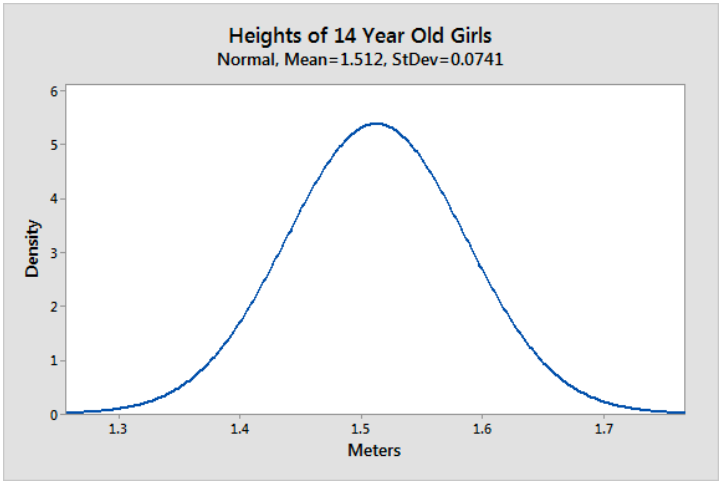
The standard deviation is the square root of variance.

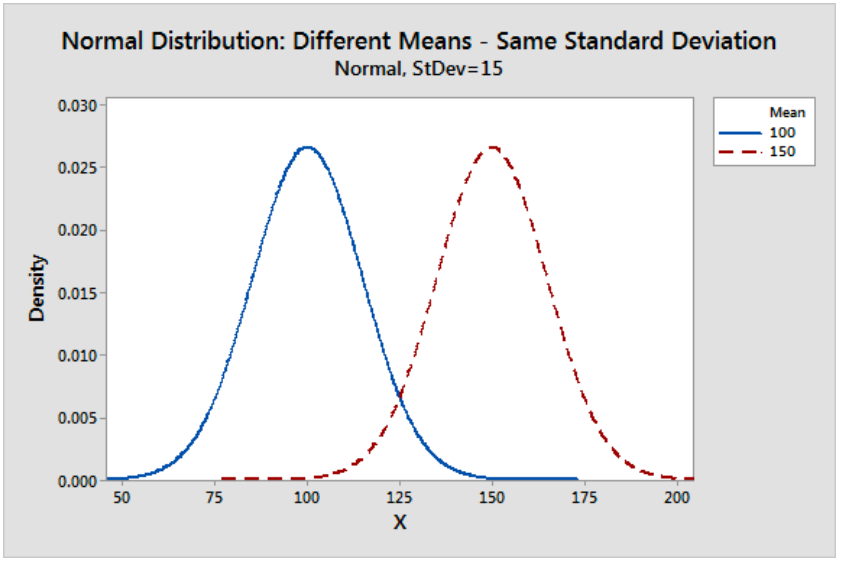
The normal distribution is completed characterized by mean and variance.

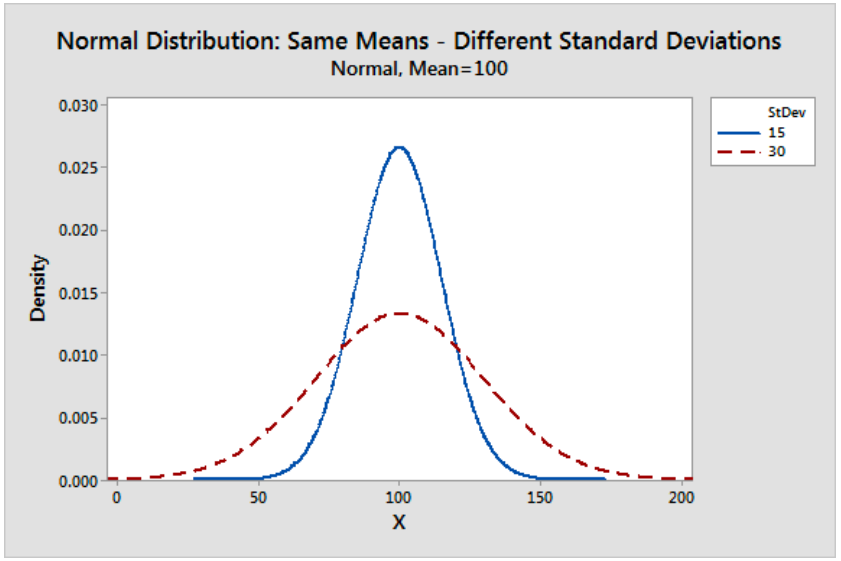


Normal Distribution

- Symmetric
- Heights, BP, error, IQ

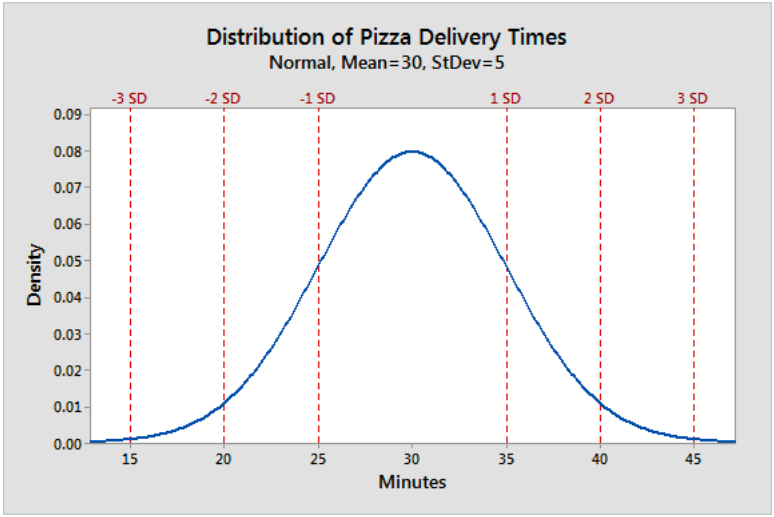






Some Properties

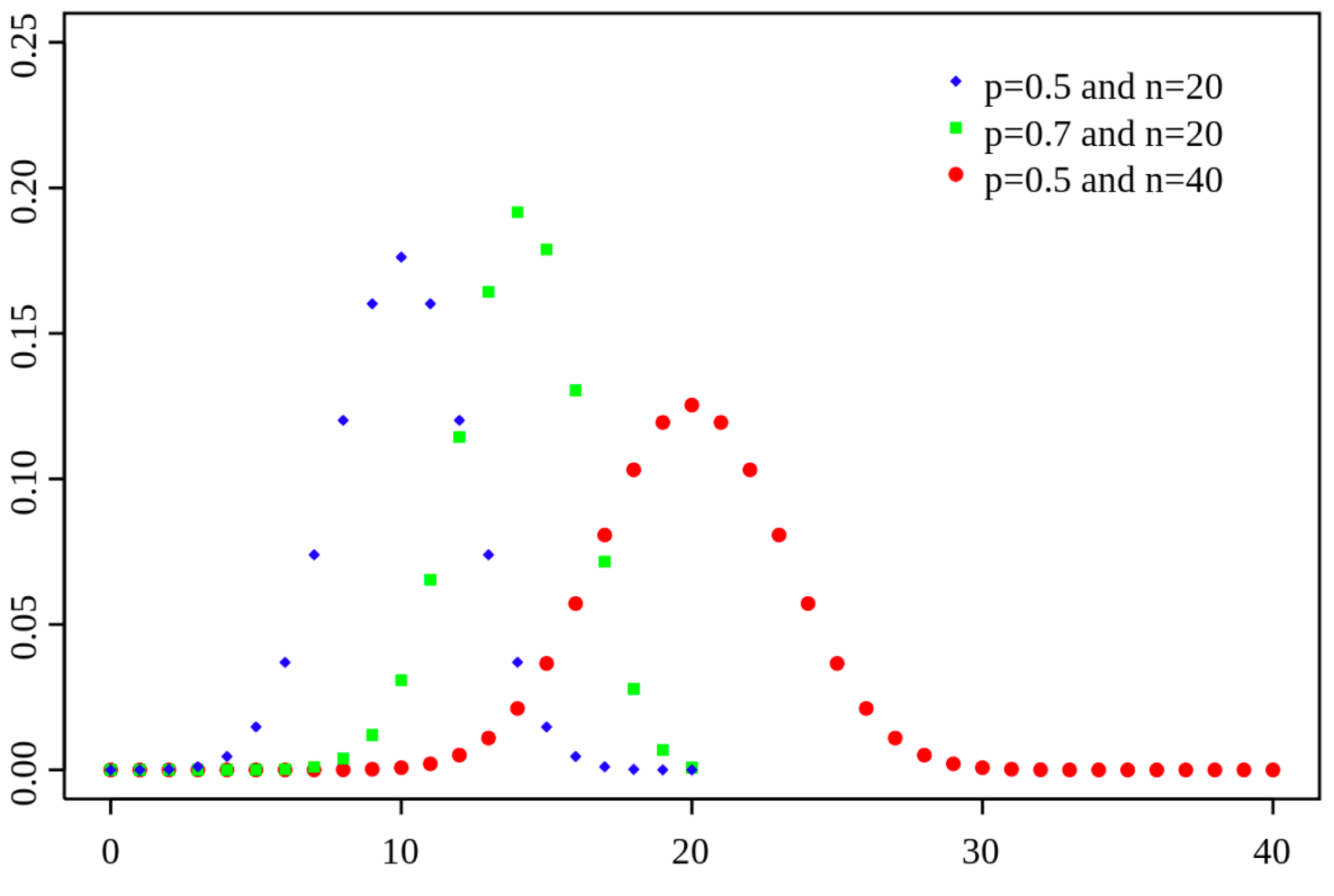
- They’re all symmetric bell curves. The Gaussian distribution cannot model skewed distributions.
- The mean, median, and mode all equal.
- Half of the population is less than the mean and half is greater than the mean.
- The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean. More on this below!
  - 68-95-99.7



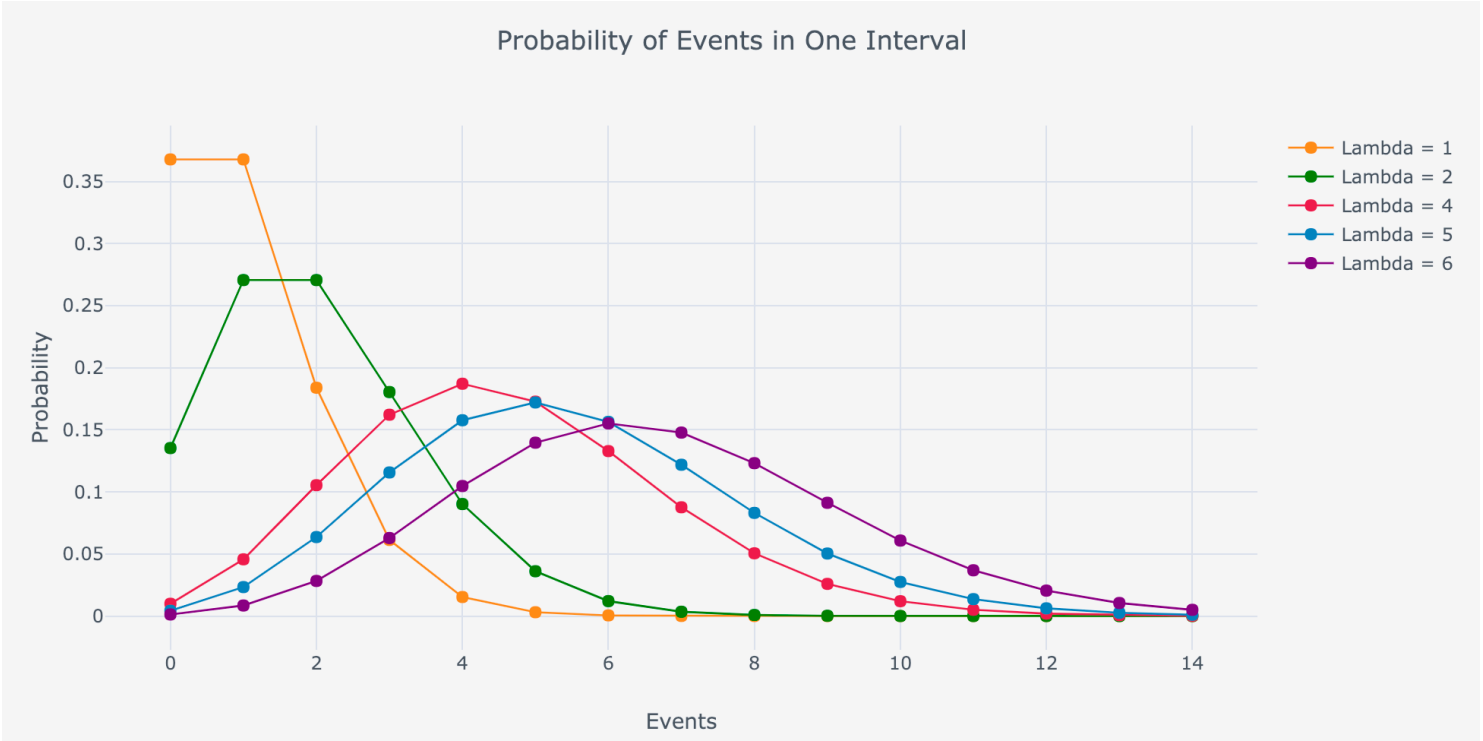
- The **Z-score**

$$Z = \frac{x - \mu}{\sigma}$$

Binomial Distribution



Poisson



$$P(k \text{ events in time period}) = e^{-\frac{\text{events}}{\text{time}} * \text{time period}} * \frac{(\frac{\text{events}}{\text{time}} * \text{time period})^k}{k!}$$

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

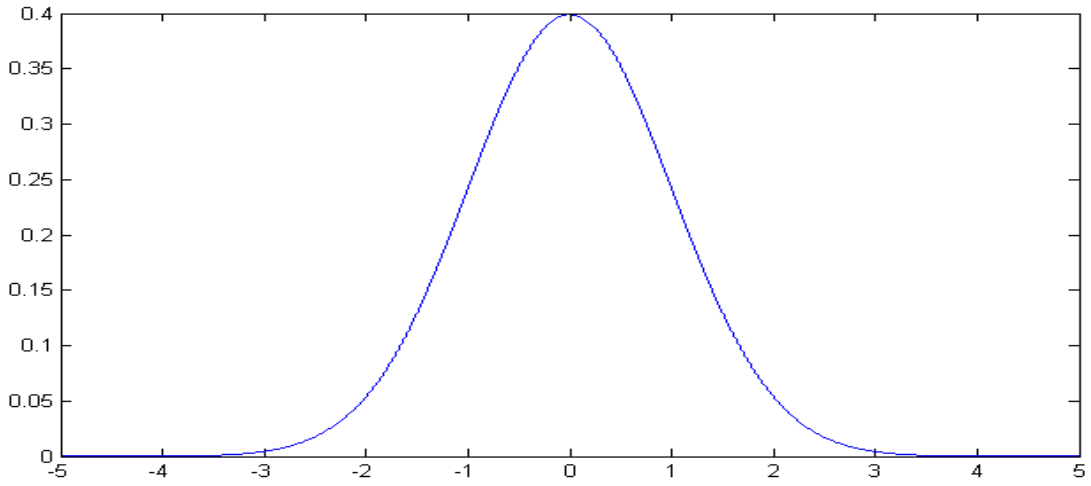
- Examples
  - Web site clicks (per minute), Calls per hour, Network failures per hour

Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

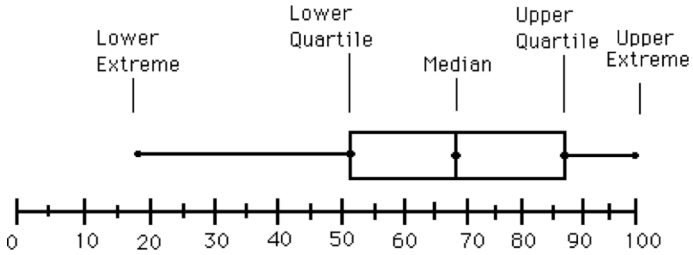
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



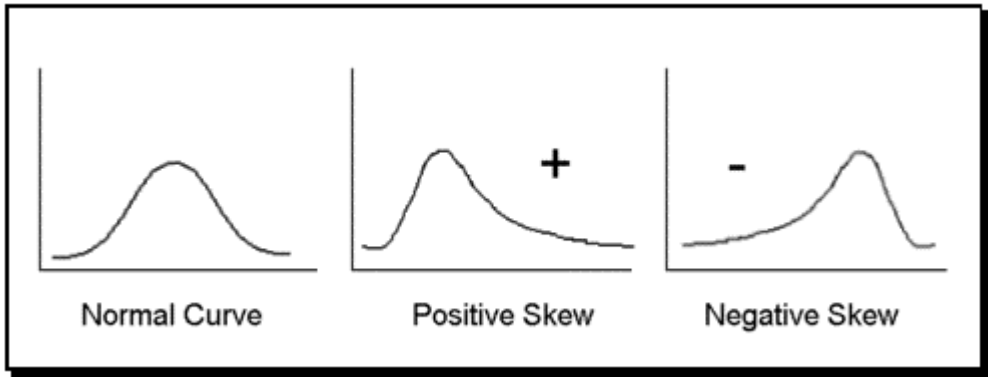
Correcting Distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

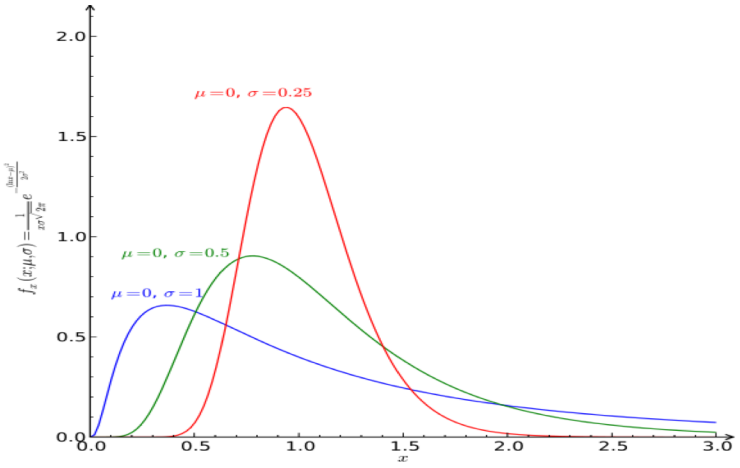


Correcting Distributions

In many cases these distribution can be corrected before any other processing.

Examples:

- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X poisson with mean k and sdev. sqrt(k). Then sqrt(X) is approximately normally distributed with sdev 1.

Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards.

Q: what's wrong with his conclusion?

Rhine Paradox

He called back the “psychic” subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

Hypothesis Testing

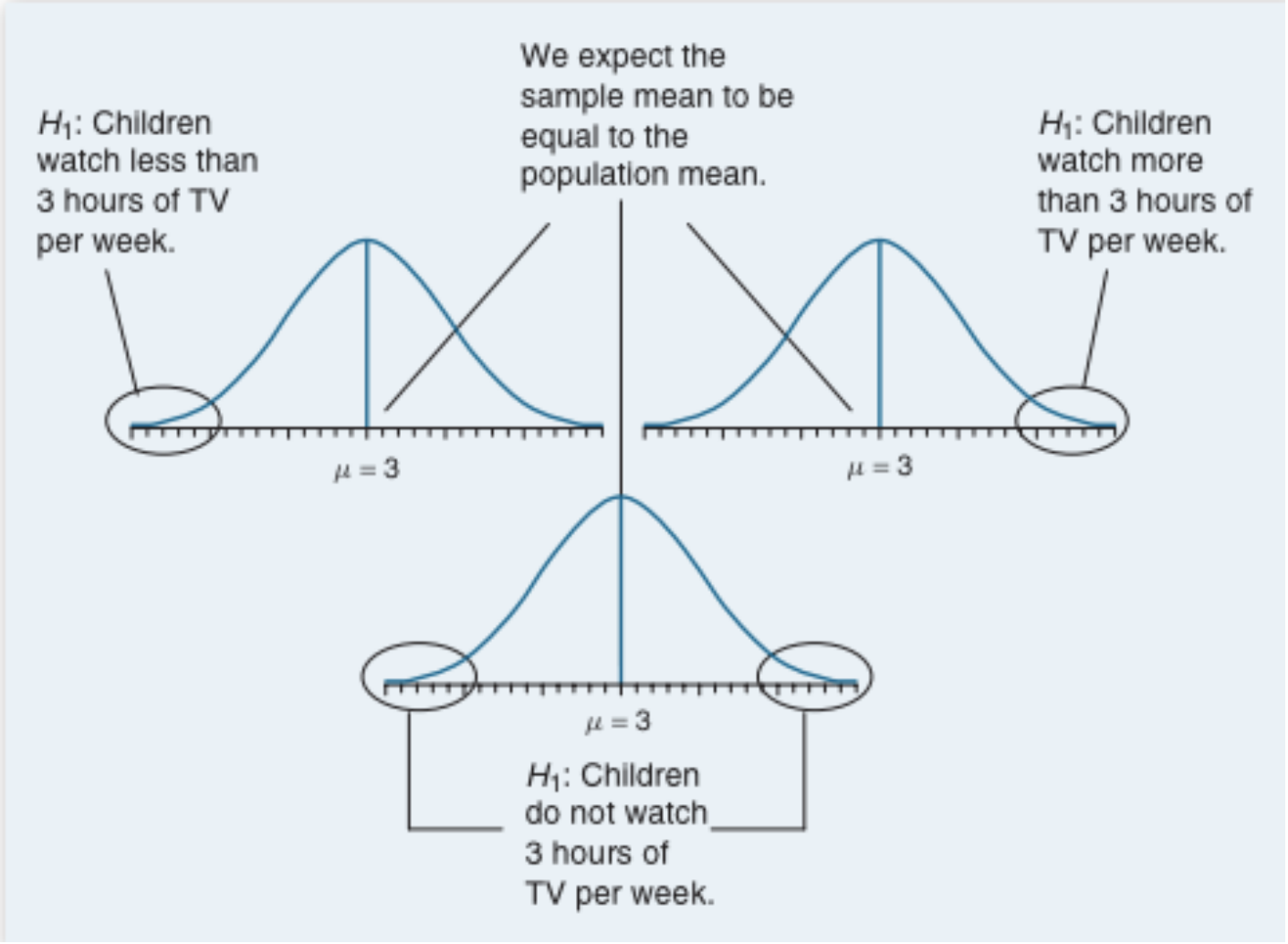
- We want to prove a hypothesis H_A, but its hard so we try to **disprove a null hypothesis H_0**.
- A **test statistic** is some measurement we can make on the data which is likely to be **big under H_A** but **small under H_0**.
- We chose a test statistic whose distribution we know if H_0 is true: e.g.
  - Two samples a and b, normally distributed, from A and B.
  - H_0 hypothesis that mean(A) = mean(B), test statistic is: s = mean(a) – mean(b).
  - s has mean zero and is normally distributed under H_0.
  - But its “large” if the two means are different.

Hypothesis Testing – contd.

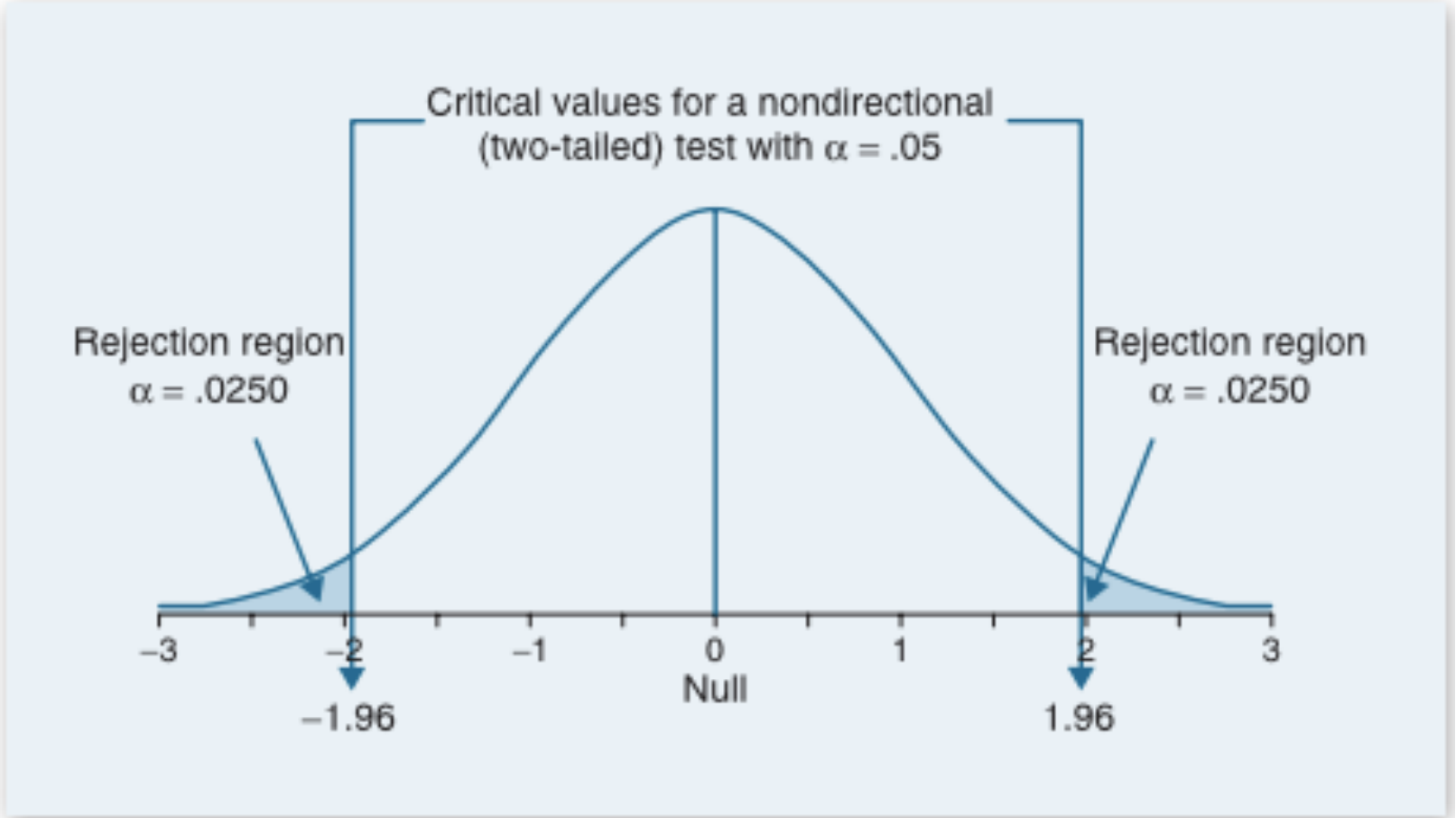
- s = mean(a) – mean(b) is our test statistic,
  H_0 the hypothesis that mean(A)=mean(B)
  - We reject if $Pr(x > s \mid H_0) < p$
  - p is a suitable “small” probability, say 0.05.

- This threshold probability is called a p-value.
  - P directly controls the false positive rate (rate at which we expect to observe large s even if is H_0 true).
  - As we make p smaller, the false negative rate increase – situations where mean(A), mean(B) differ but the test fails.
  - Common values 0.05, 0.02, 0.01, 0.005, 0.001

P-value


We expect the sample mean to be equal to the population mean.
H_1: Children watch less than 3 hours of TV per week.
H_1: Children watch more than 3 hours of TV per week.
$\mu = 3$
$\mu = 3$
$\mu = 3$
H_1: Children do not watch 3 hours of TV per week.

Two-tailed Significance



When the p value is less than 5% ($p < .05$), we reject the null hypothesis

Three important tests

- **T-test:** compare two groups, or two interventions on one group.

- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table".

- **ANOVA:** compare outcomes under several discrete interventions.

T-test

Single-sample: Compute the test statistic:

$$t = \frac{\bar{X}}{\bar{\sigma}}$$

where $\bar{X}$ is the sample mean and $\bar{\sigma}$ is the sample standard deviation, which is the square root of the sample variance Var(X).

If X is normally distributed, t is **almost** normally distributed, but not quite because of the presence of $\bar{\sigma}$.

You use the single-sample test for **one group** of individuals in **two conditions**. Just subtract the two measurements for each person, and use the difference for the single sample t-test.

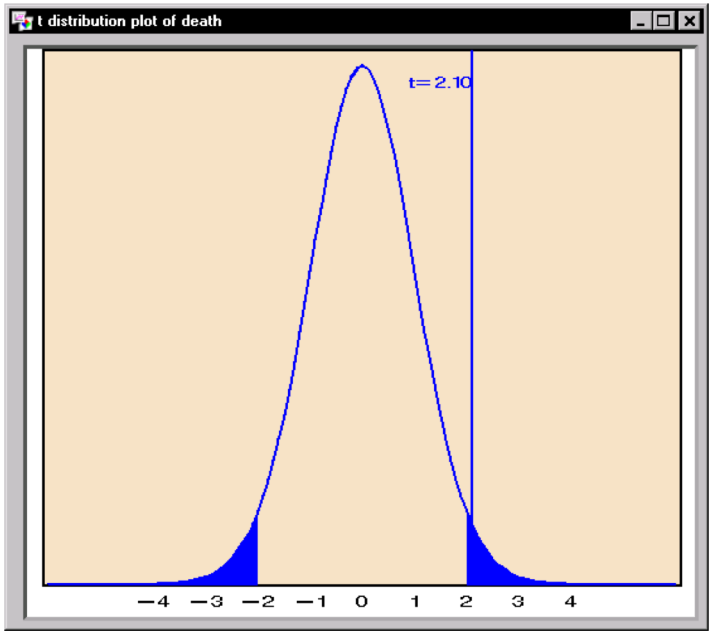
This is called a **within-subjects** design.

T-statistic and T-distribution

- We use the t-statistic from the last slide to test whether the mean of our sample could be zero.
- If the underlying population has mean zero, the t-distribution should be distributed like this:

- The area of the tail beyond our measurement tells us how likely it is under the null hypothesis.

- If that probability is low (say < 0.05) we reject the null hypothesis.



Two sample T-test

In this test, there are **two samples** X_1 and X_2. A t statistic is constructed from their sample means and sample standard deviations:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where: $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$

This design is called a **between-subjects** test.

Chi-squared test

Often you will be faced with discrete (count) data. Given a table like this:

	Prob(X)	Count(X)
X=0	0.3	10
X=1	0.7	50

Where Prob(X) is part of a null hypothesis about the data (e.g. that a coin is fair).

The CHI-squared statistic lets you test whether an observation is consistent with the data:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i is an observed count, and E_i is the expected value of that count. It has a chi-squared distribution, whose p-values you compute to do the test.

ANOVA

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

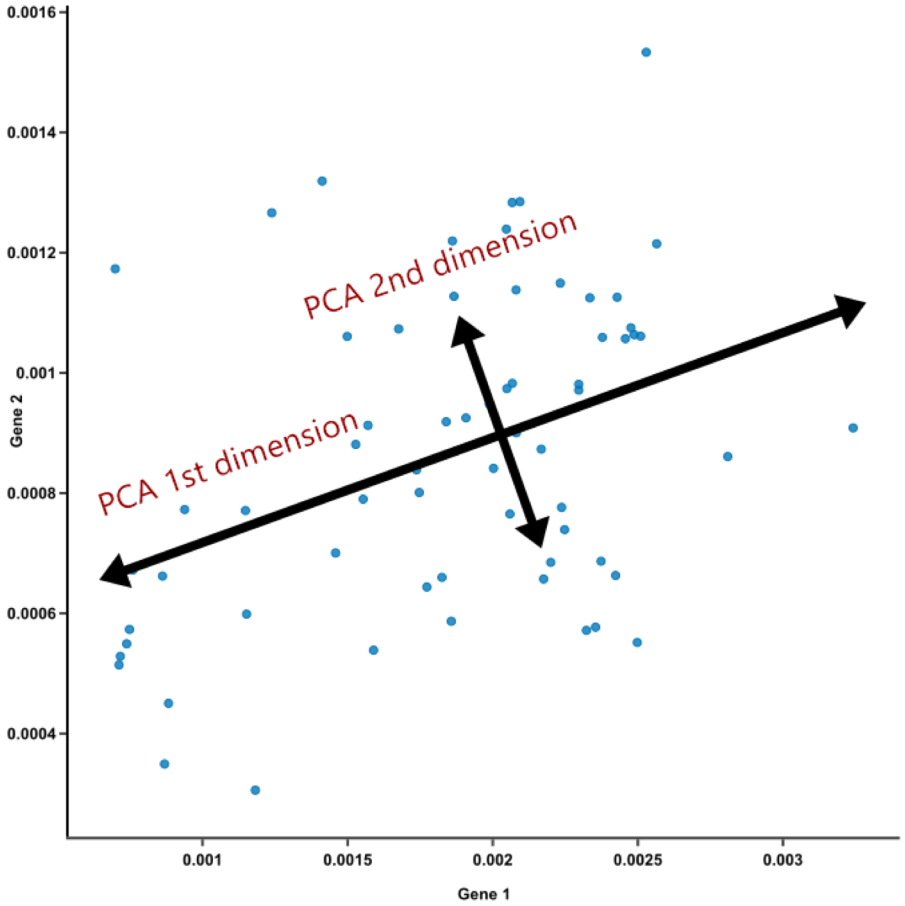
$$F = \frac{VAR_{between}}{VAR_{within}}$$

The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.

We can look up the F-statistic value in a cumulative F-distribution (similar to the other statistics) to get the p-value.

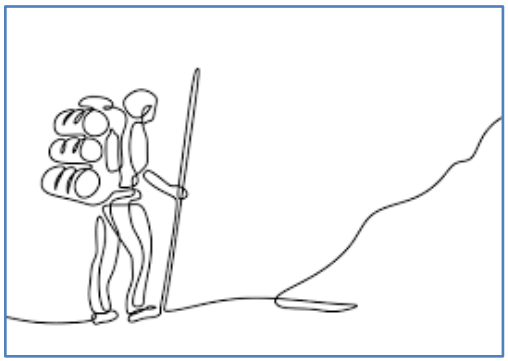
ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

PRINCIPAL COMPONENT ANALYSIS



https://setosa.io/ev/principal-component-analysis/

MACHINE LEARNING FRAMEWORKS

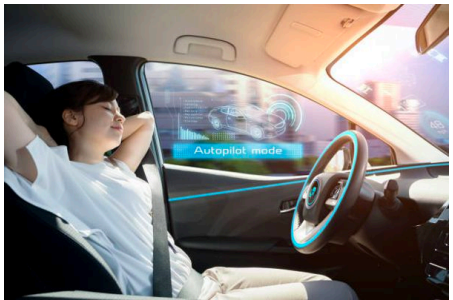


Statistical analysis on paper

Spreadsheets

Statistical analysis with computer tools
R, SPSS, ...

Machine Learning
Weka, SciKit, TensorFlow, PyTorch, ...



***Auto** ML*

Google Cloud AutoML,
AzureML, AWS SageMaker