Rensselaer

why not change the world?®

Stats Review, Distributions, Hypothesis Testing

Ahmed Eleish

Data Analytics ITWS-4600/6600 CSCI-4600 MGMT-4600/6600 BCBP 4600

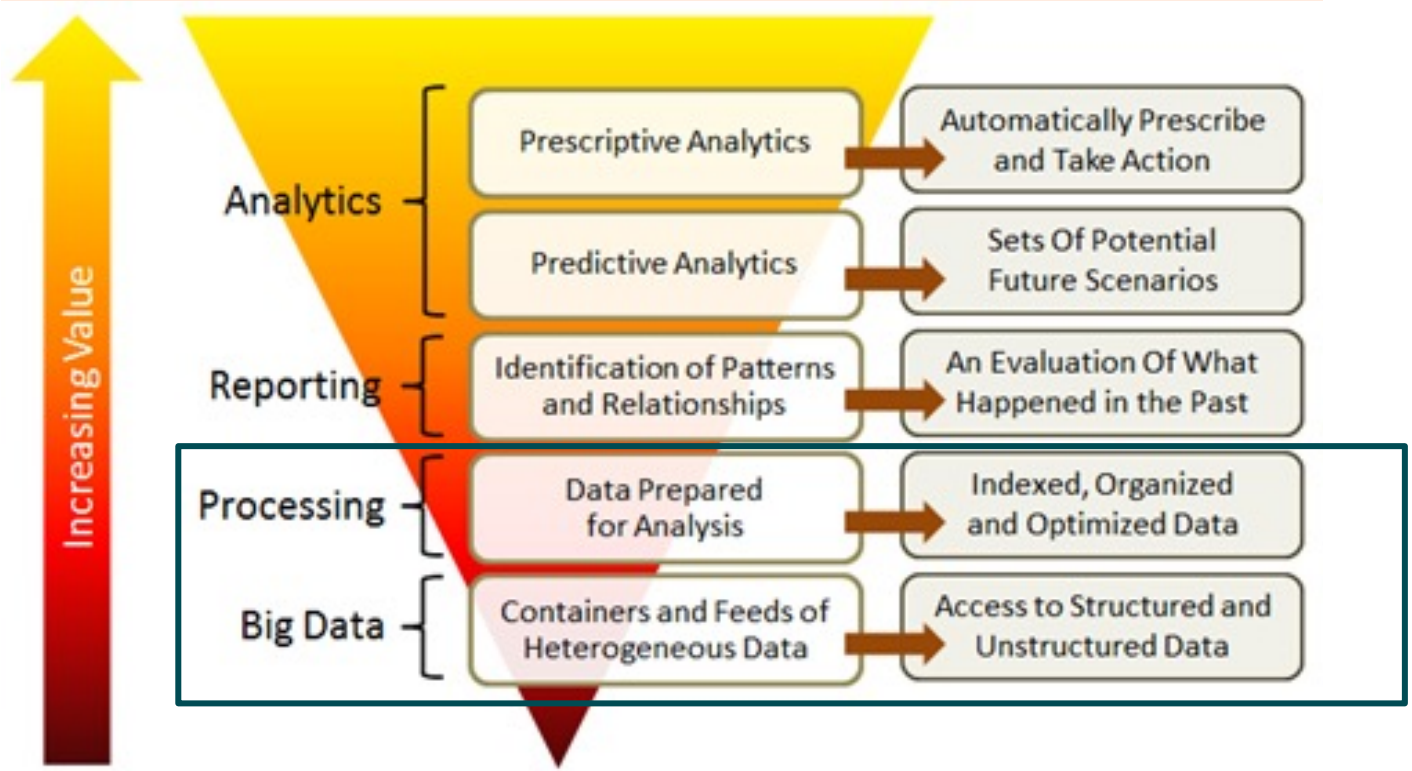
Group 1 Module 3, January 20th, 2026

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Stats review cont'd
- Exploring Datasets
  - Distributions
  - Visualizations
- Testing and evaluating the results (beginning)

Lower layers in the Analytics Stack



Stats review – cont'd

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency

Previous class

- Frequency distributions
- Probability
- Hypothesis (null and alternate)
- Significance tests
- P-value

Today's class

Measure of Central Tendency

• Mean: The most commonly used measure of central tendency, commonly referred to as “Average”, sensitive to extreme values (sensitive to outliers)

– Population Mean

– Sample Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Image Resource: https://www.onlinemathlearning.com/population-mean.html Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Standard Deviation

Population standard deviation of grades of eight students [edit]

Suppose that the entire population of interest is eight students in a particular class. For a finite set of numbers, the population standard deviation is found by taking the square root of the average of the squared deviations of the values subtracted from their average value. The marks of a class of eight students (that is, a statistical population) are the following eight values:

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = \frac{40}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{aligned} (2-5)^2 &= (-3)^2 = 9 & (5-5)^2 &= 0^2 = 0 \\ (4-5)^2 &= (-1)^2 = 1 & (5-5)^2 &= 0^2 = 0 \\ (4-5)^2 &= (-1)^2 = 1 & (7-5)^2 &= 2^2 = 4 \\ (4-5)^2 &= (-1)^2 = 1 & (9-5)^2 &= 4^2 = 16. \end{aligned}$$

The variance is the mean of these values:

$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = \frac{32}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{4} = 2.$$

https://en.wikipedia.org/wiki/Standard_deviation

Sample vs. Population

- Population
  - *All possible data points*
  - *May be of finite size (N) or infinite*
  - *Greek letters for parameters (μ, σ)*
  - *Parameters are estimated*

- *Sample*
  - *Finite subset of the population*
  - *Of finite size (n)*
  - *Latin letters for statistics (m, s)*
  - *Statistics are computed*

Sample vs. Population

- If sample *is* population, then $\mu = m$
- Realistically while they are not equal, m is a good estimator for μ

Law of large numbers

- *“the average of the results obtained from a large number of independent random samples converges to the true value, if it exists”*
- *“given a sample of independent and identically distributed values, the sample mean converges to the true mean."*

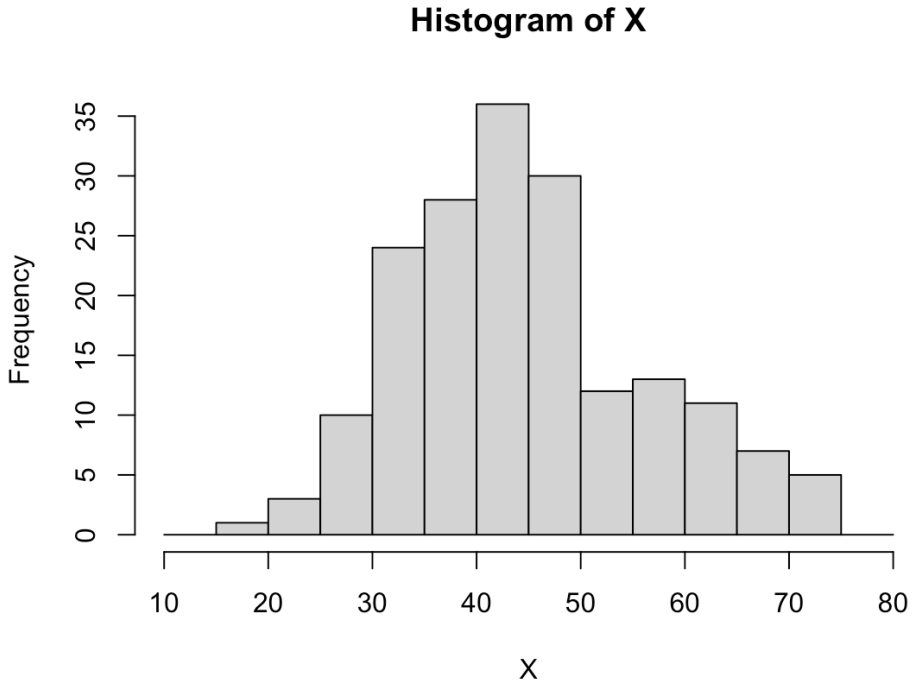
For this course

- Consider the observations in the given/acquired dataset the **entire** population.

Exploring distributions

- Histograms and binning
- Density curves
- Visual analysis
- Statistical tests

Grouped Frequency Distribution aka binning

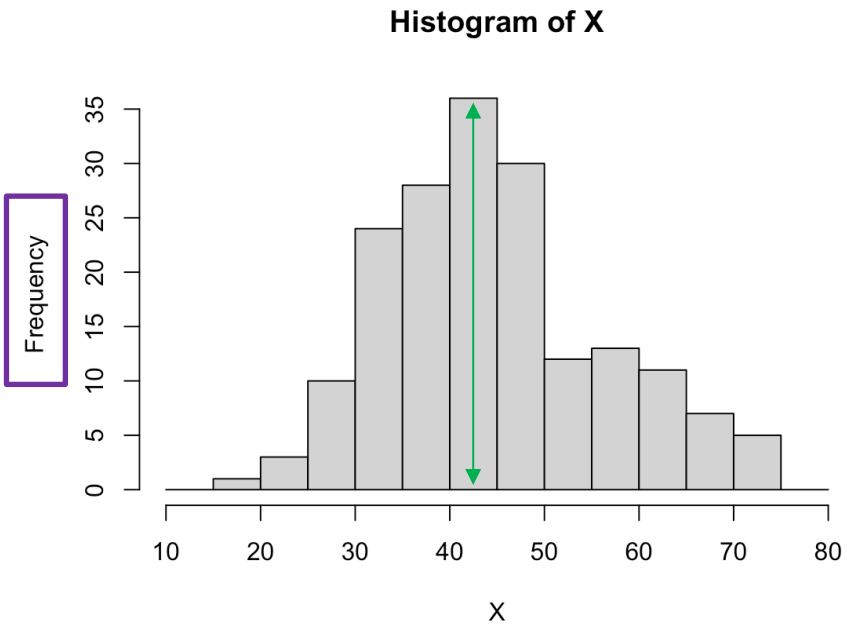


180 observations

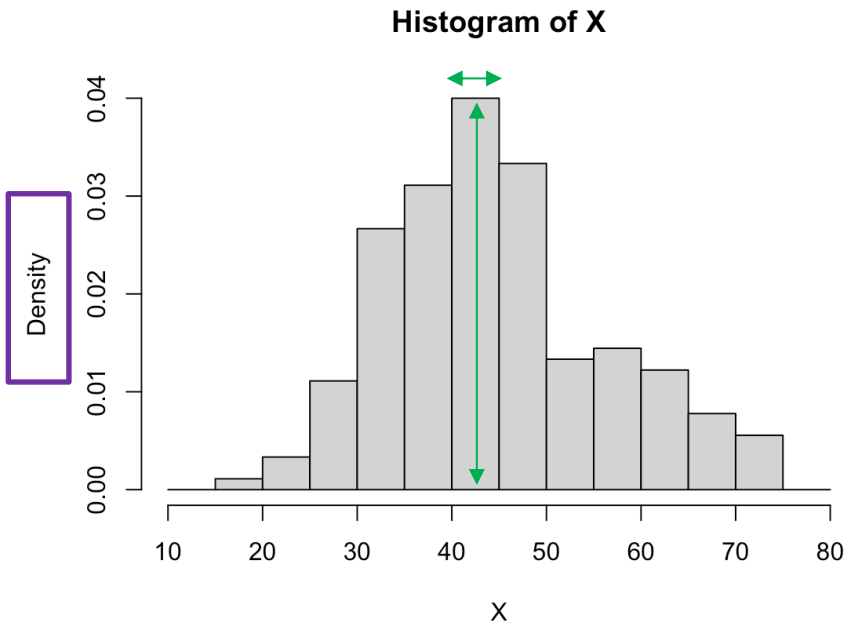
$15 \leq X \leq 75$

Histogram bin size = 5

Frequency vs. Density

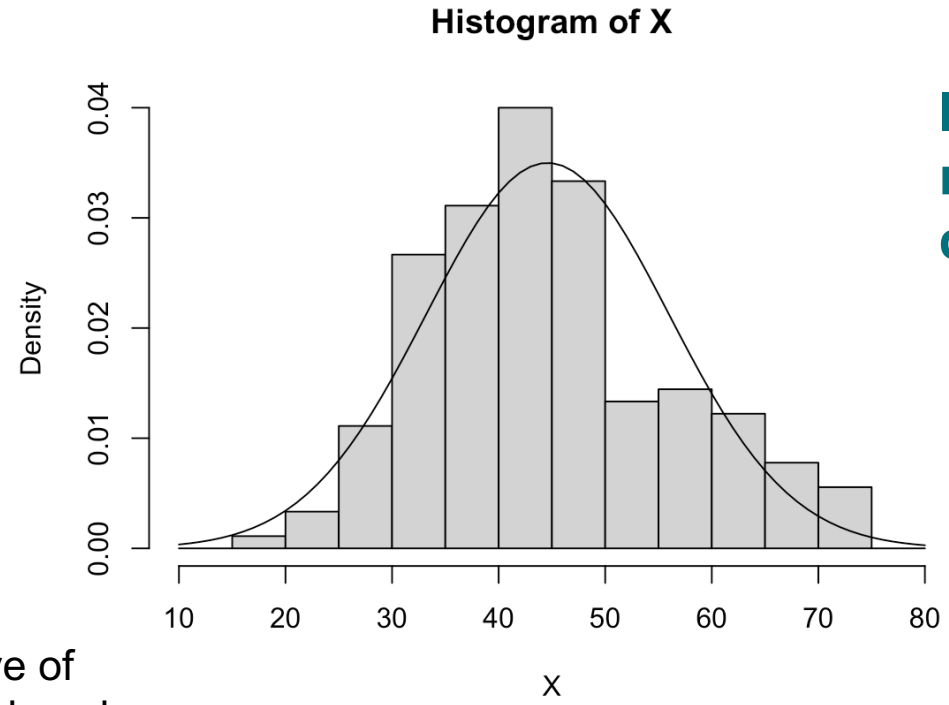


- 36 observations where 40 <= x < 45
- 36/180 (total) = 0.2 or 20%



- density = 0.04 where 40 < x < 45
- area of bar = 0.04 * 5 (width of bar) = 0.2 or 20%

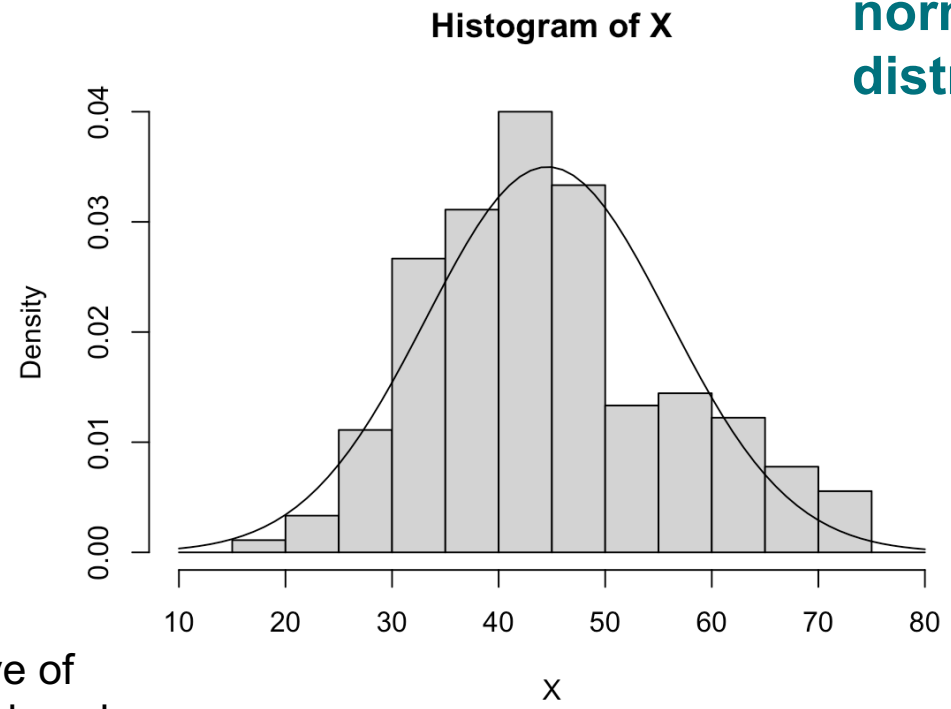
Empirical vs. Theoretical



Is variable X normally distributed??

- Probability density curve of normal distribution overlayed

Empirical vs. Theoretical



Is variable X normally distributed??

Stay tuned to find out!

- Probability density curve of normal distribution overlayed

Frequencies v. Probabilities

- Actual rate of occurrence in a sample or population – frequency
- Expected or estimate likelihood of a value or outcome – probability

e.g. Coin toss – two outcomes (binomial) p= 0.5 (of “heads”)

Probability Distributions

- Shape
- Parameter(s)

e.g. mean, standard deviation, rate, number of trials, etc.

- Probability distributions are functions!

- A probability density function describes the likelihood that a random variable will take on a specific value

- Which one fits?
  - Visually
  - Statistical tests..

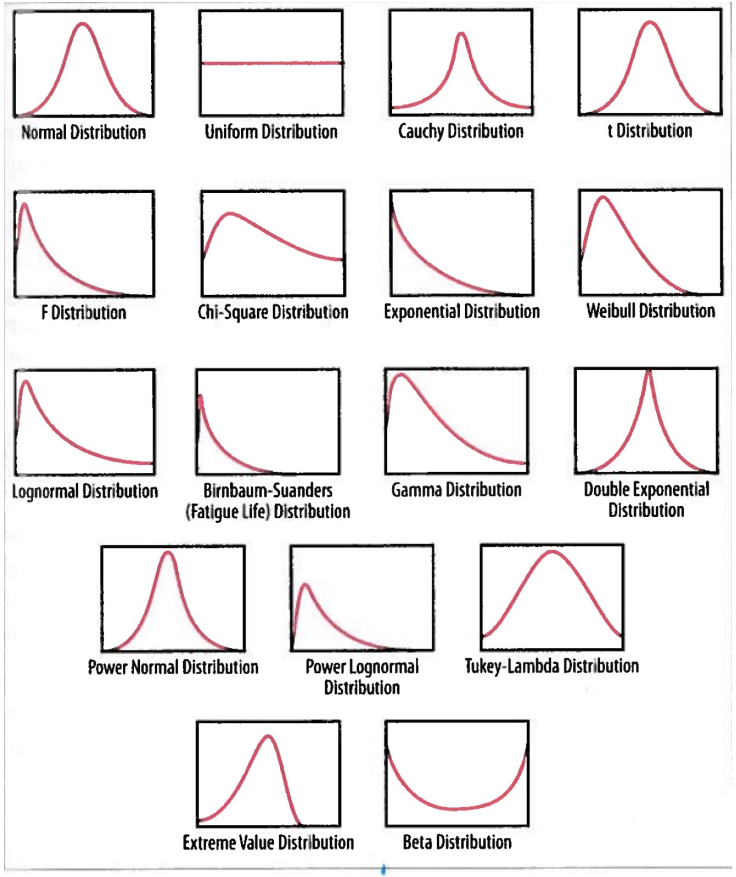


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Binomial Distribution

• Describes the outcome of coin toss experiments.

• Binomial distributions are discrete and are defined by 2 parameters: *p* (probability of success) and *n* (number of trials)

• Probability Mass Function

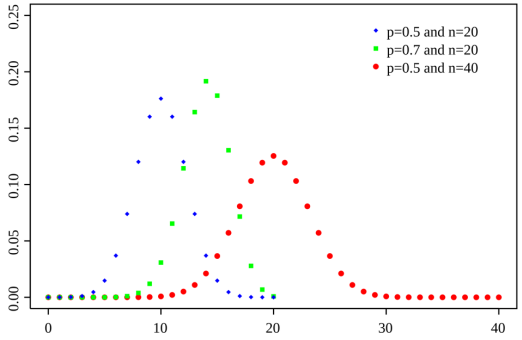
$$f(k, n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, ..., n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

e.g. Probability of obtaining exactly 1 head in 2 coin tosses:

$$C_1^2 * 0.5^1 * 0.5^{2-1} = 0.5$$



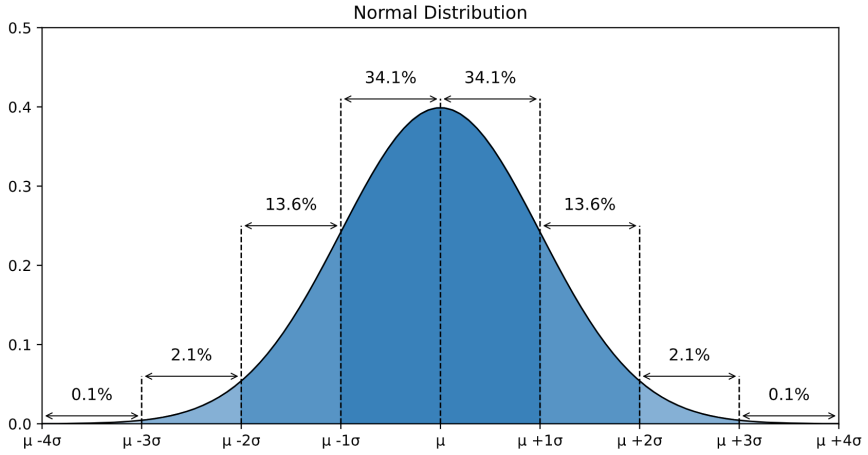
Probability Mass Function of Binomial Distribution

https://en.wikipedia.org/wiki/Binomial_distribution
Image credit: Tayste

Normal Distribution

- The normal distribution implies tight bounds on the probability of lying far from the mean.

- 68% of the values must lie within one sigma (standard deviation) of the mean, and 95% within two times the sigma (standard deviation) and 99.7% lie within the three the sigma (standard deviation)



- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

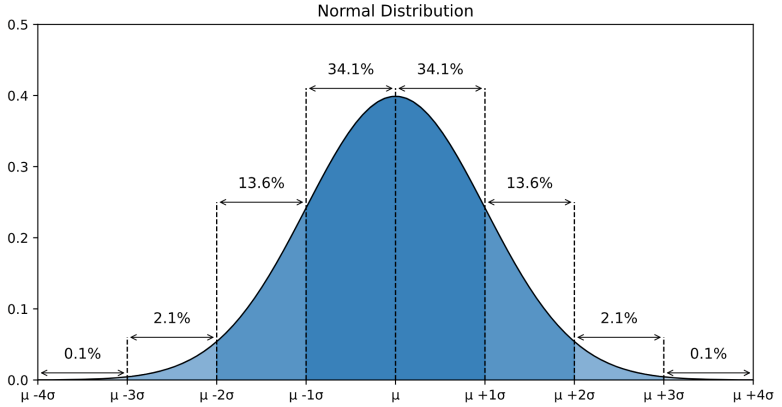
Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Normal Distribution

Probability Density Function:

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

μ: mean, σ^2: variance



- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Normal Distribution

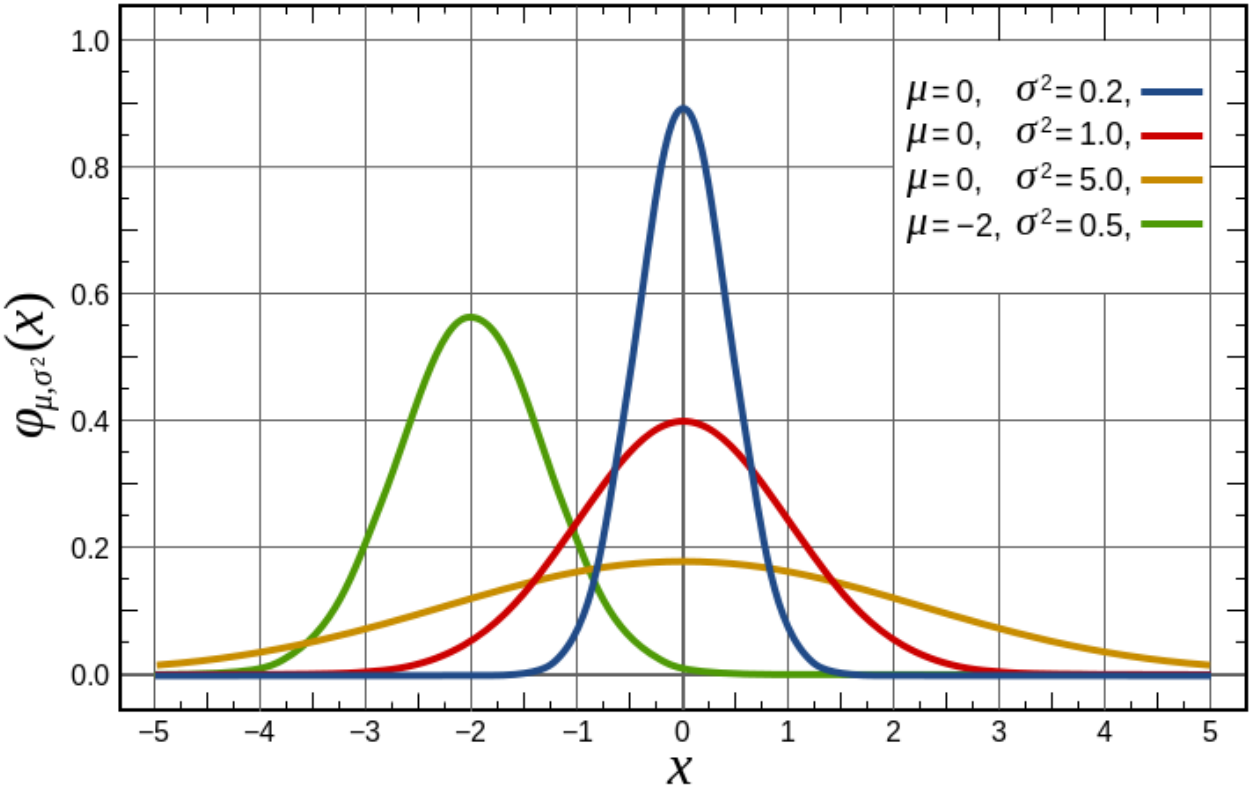
• The bell-shaped curve or Gaussian distribution describes a continuous random variable and is parameterized by the variable's mean and standard deviation.

• Many phenomena in the natural and social sciences are modeled with the normal distribution.

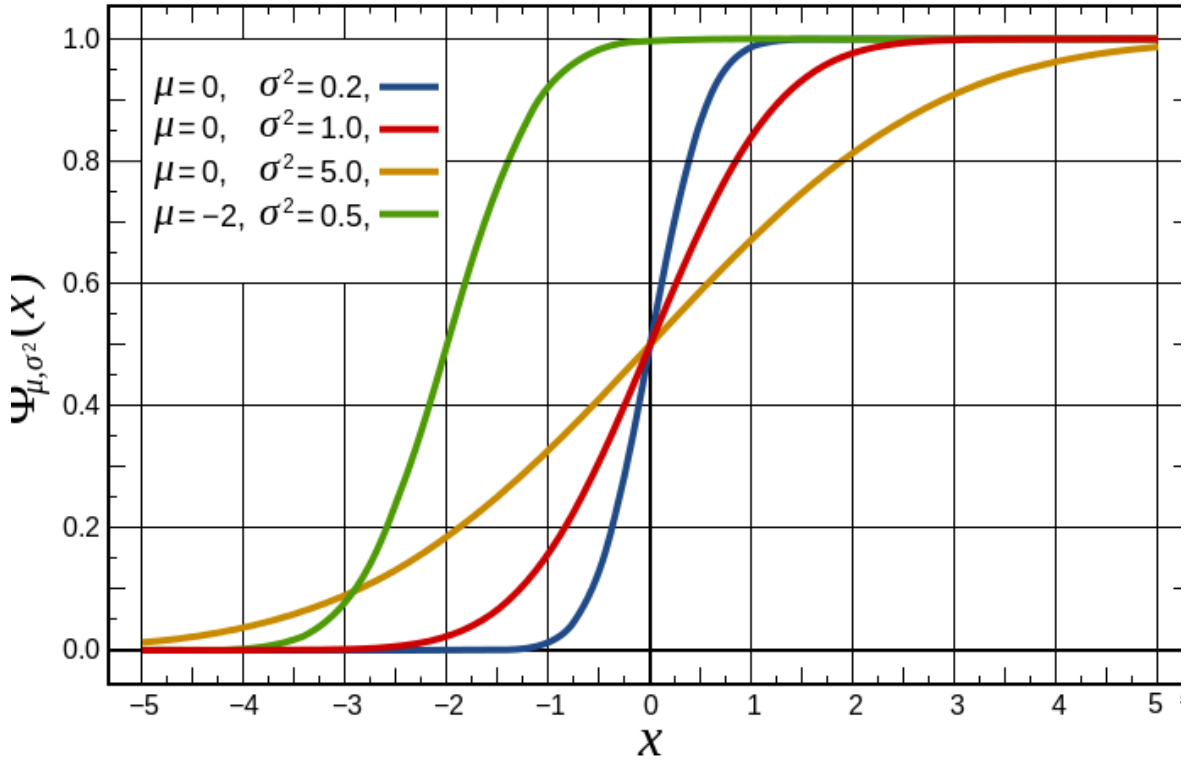
• Normal distributions can be used to approximate binomial distributions with large numbers of trials.

https://www.youtube.com/watch?v=4HpvBZnHOVI

Probability Density of Normal Distribution

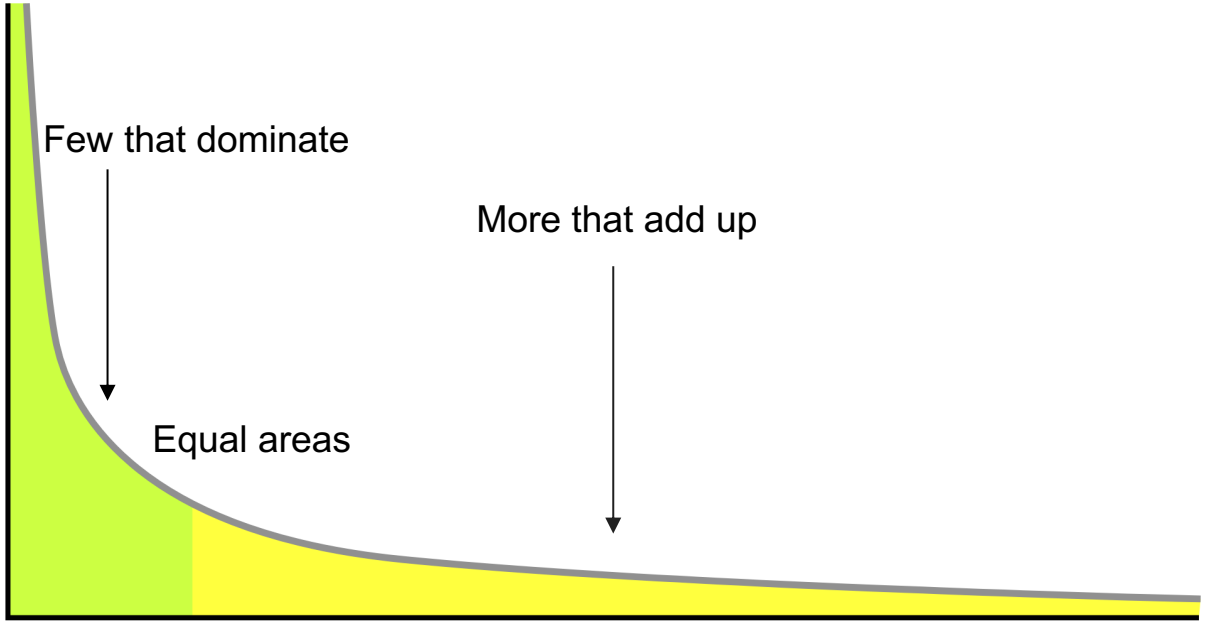


Cumulative Density Function



Heavy-tail distributions

- Probability distributions whose tails are not exponentially bounded
  e.g. long-tail distributions - common in business, marketing, social media mechanisms



http://en.wikipedia.org/wiki/Heavy-tailed_distribution

Distribution tests

Most distributions have tests:

- Wilcoxon-Mann-Whitney test

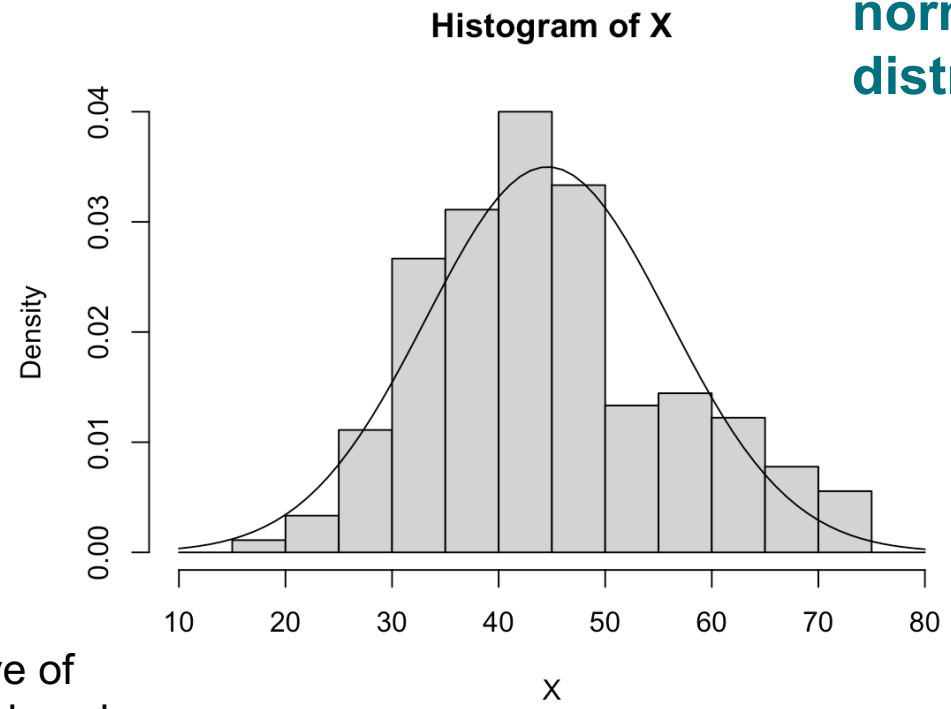
  – Comparing populations

  Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.
  http://www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test

- Kolmogorov-Smirnov
- Shapiro–Wilk
- Anderson–Darling

…

Empirical vs. Theoretical

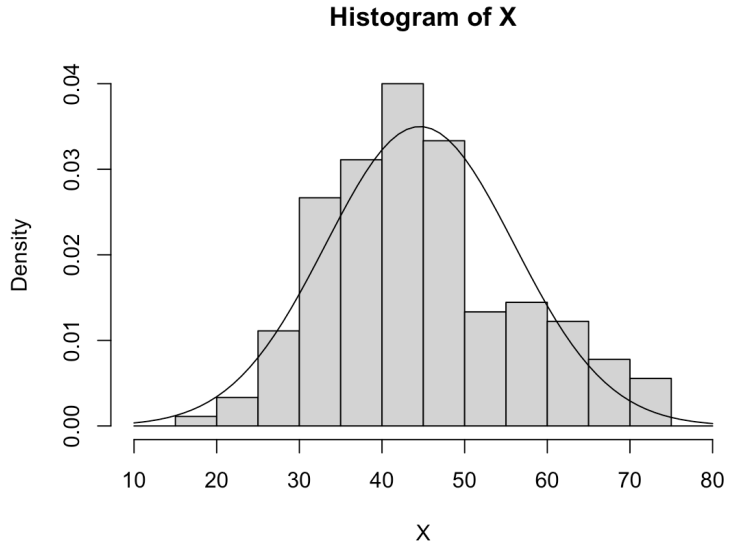


Is variable X normally distributed??

We can run a test!

- Probability density curve of normal distribution overlayed

Empirical vs. Theoretical



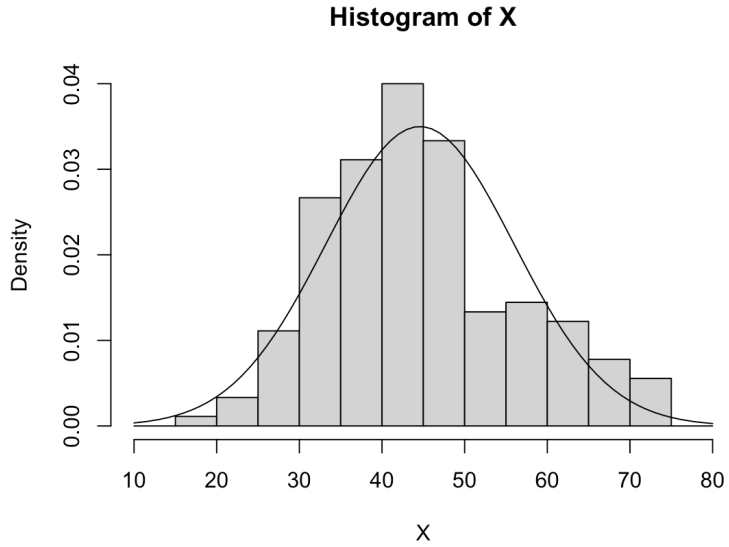
Is variable X normally distributed??

Shapiro-Wilk normality test

data: X
W = 0.96964, p-value = 0.0005824

How do we interpret the results??

Empirical vs. Theoretical



Is variable X normally distributed??

Shapiro-Wilk normality test

data: X
W = 0.96964, p-value = 0.0005824

How do we interpret the results??

Stay tuned!

Hypothesis-driven Research

1. Write the original claim and identify whether it is the null hypothesis or the alternative hypothesis.
2. Write the null and alternative hypotheses. Use the alternative hypothesis to identify the type of test.
3. Write down all information from the problem.
4. Compute the test statistic.
5. Make a decision to reject or fail to reject the null hypothesis. A figure showing the critical value and test statistic may be useful.
6. Write the conclusion.

Null and Alternate Hypotheses

- H0 – null
- H1 – alternate

- If a given claim contains equality, or a statement of no change from the given or accepted condition, then it is the null hypothesis, otherwise, if it represents change, it is the alternative hypothesis.

e.g.

N0: **There is no significant difference between sales of 2 stores.**

N0: **Test scores are normally distributed.**

N1: **Sales increase with more advertising.**

~~Accept or~~ Reject? **or fail to reject!**

• **Reject** the **null** hypothesis if the **p-value** is less than the level of significance (α).

• You will **fail to reject** the **null** hypothesis if the **p-value** is greater than or equal to the level of significance (α).

• **Typical significance $\alpha = 0.05$ (!)**

Shapiro-Wilk

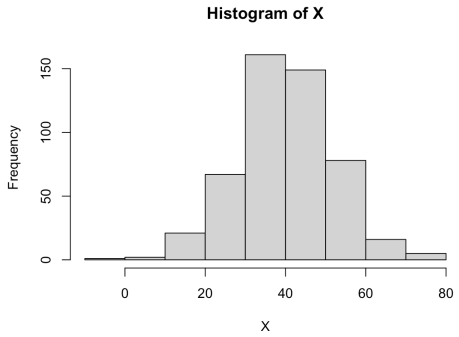
- “The Shapiro–Wilk test tests the null hypothesis that a sample *x1, ..., xn* came from a normally distributed population.”

- **H0: variable X is normally distributed**

shapiro.test(X)

Shapiro-Wilk normality test

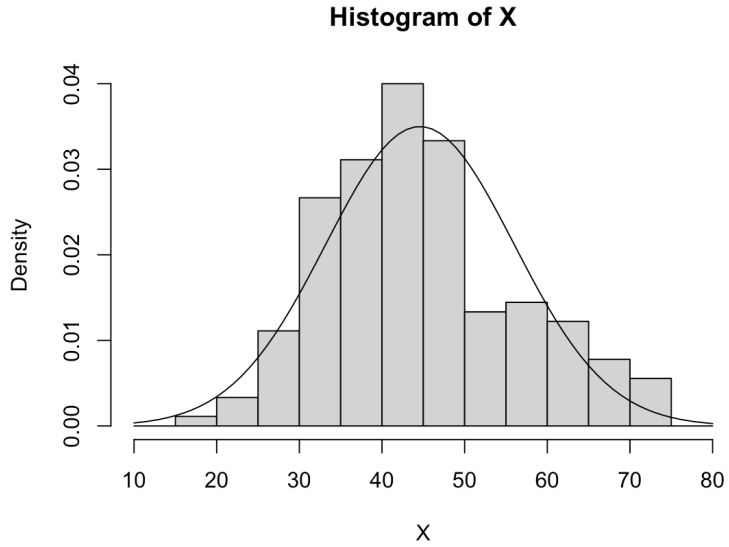
data: X
W = 0.99638, p-value = 0.3175



$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2},$$

https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Empirical vs. Theoretical



Is variable X normally distributed??

No!!

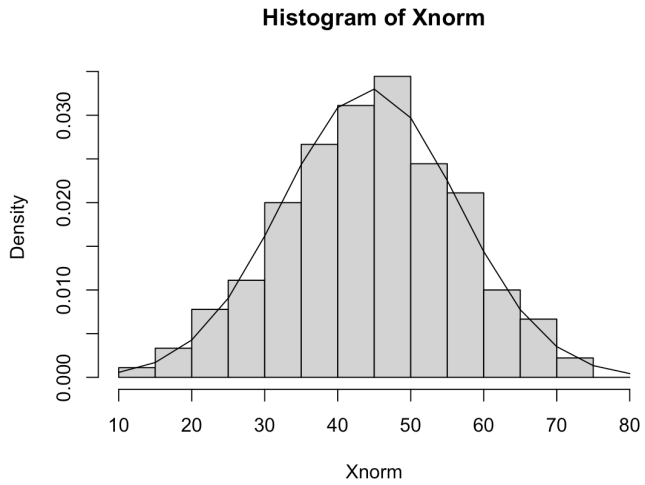
Shapiro-Wilk normality test

data: X
W = 0.96964, p-value = 0.0005824

Less than 0.05!

Empirical vs. Theoretical

Let's generate 180 numbers drawn at random from a normal distribution with the same mean and sd as X



Is variable X normally distributed??

Yes!!

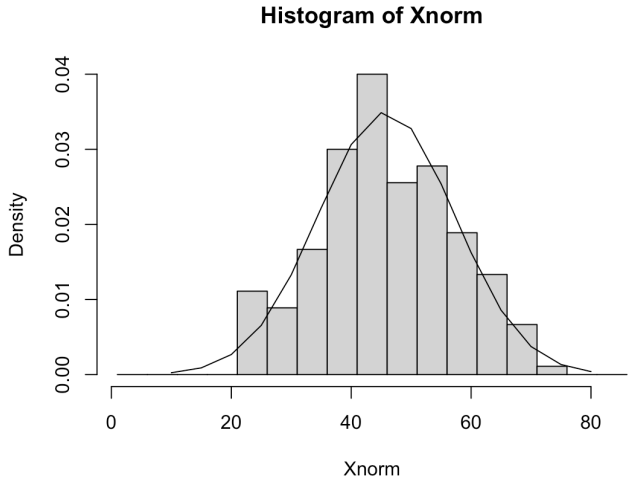
Shapiro-Wilk normality test

data: Xnorm
W = 0.99517, p-value = 0.8308

greater than 0.05!

Empirical vs. Theoretical

Let's generate 180 **NEW** numbers drawn at random from a normal distribution with the same mean and sd as X



Is variable X normally distributed??

Yes!!

Shapiro-Wilk normality test

data: Xnorm
W = 0.98952, p-value = 0.2079

greater than 0.05!

Anderson-Darling (Normality Test)

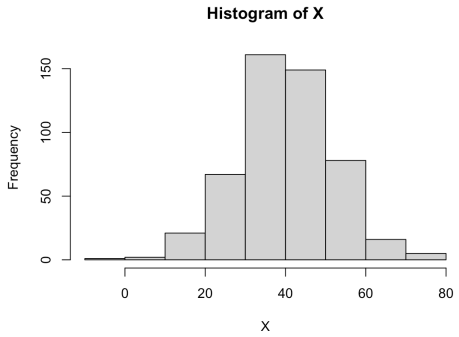
- “The Anderson–Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution”

- H0: variable X is normally distributed

ad.test(X)

Anderson-Darling normality test

data: X
A = 0.26158, p-value = 0.7048



$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) \, dF(x),$$

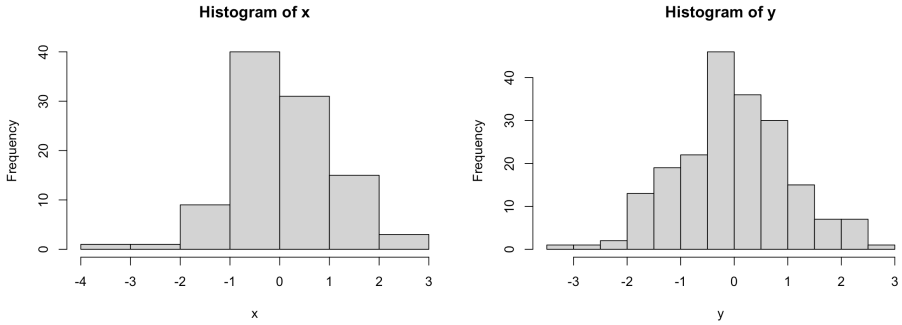
https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

Kolmogorov-Smirnov

- “to test whether two samples came from the same distribution (two-sample K–S test)”
- H0: x and y are from the same distribution

ks.test(x,y)

$$D_n = \sup_x |F_n(x) - F(x)|$$



Asymptotic two-sample Kolmogorov-Smirnov test

data: x and y
D = 0.08, p-value = 0.787
alternative hypothesis: two-sided

https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Wilcoxon-Mann-Whitney

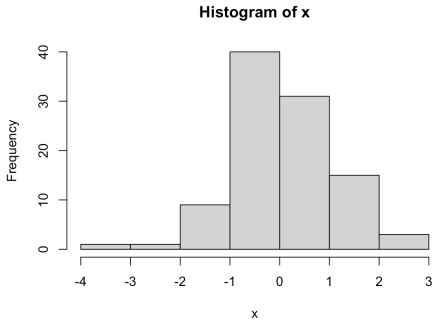
- “a nonparametric statistical test of the null hypothesis that randomly selected values X and Y from two populations have the same distribution”
- H0: x and y have the same distribution

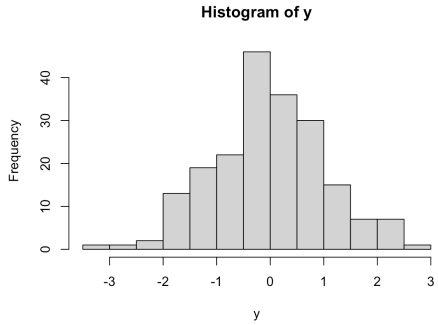
$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

wilcox.test(x,y)

Wilcoxon rank sum test with continuity correction

data: x and y
W = 10220, p-value = 0.7566
alternative hypothesis: true location shift is not equal to 0





https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

T-test

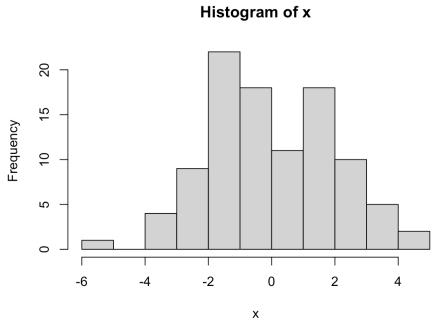
- “to test whether the difference between the response of two groups is statistically significant or not.”
- Assumes variables are normally distributed and have equal variance
- H0: x and y are not statistically significantly different

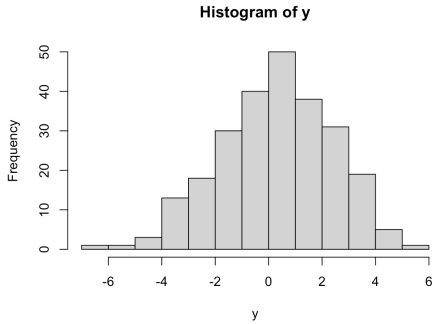
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

t.test(x,y)

Welch Two Sample t-test

data: x and y
t = -1.2348, df = 194.41, p-value = 0.2184
alternative hypothesis: true difference in means is not equal to 0





https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

F-test

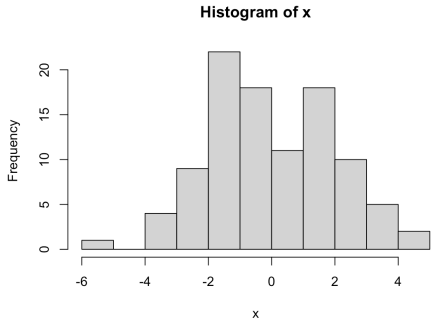
- “It is used to determine if the variances of two samples, or if the ratios of variances among multiple samples, are significantly different.”
- H0: x and y have equal variance

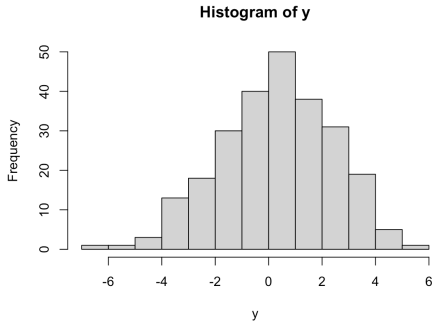
$$F = \frac{S_A^2}{S_B^2}$$

var.test(x,y)

F test to compare two variances

data: x and y
F = 0.87183, num df = 99, denom df = 249, p-value = 0.4338
alternative hypothesis: true ratio of variances is not equal to 1





https://en.wikipedia.org/wiki/F-test

Random Numbers

- Can a computer generate a random number?
- Can you?
- Why? – to reduce selection bias!
- In R – many ways – see help on Random {base} and get familiar with set.seed()

Preliminary (Exploratory) Analysis

• Determining if there is one or more common distributions involved – i.e. parametric statistics (assumes or asserts a probability distribution)

• Fitting that distribution -> provides a model!

Or NOT:

– Non-parametric (statistics) – more on this to come

Considerations

- Quality, uncertainty and bias – you will often spend a lot of time with the data
- Distributions – the common and not-so common ones and how digital vs. natural data can have distinct distributions
- How simple statistical distributions can mislead us
- Populations and samples and how inferential statistics will lead us to model choices
- Preparing data for exploratory analysis

Reminder:

- Make sure you have R & RStudio installed for Lab 1

- Experiment with R

* Create the Github repository for this class if you have not created it yet and email the repo URL to me (eleisa2@rpi.edu)

Thanks!
(See you Friday)