Distribution

& Hypothesis Testing

Statistics Tutorial

Day 5

Prabesh Dhakal

2020 April 30

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

DISTRIBUTION / . . .

We talk about data distribution. We also talk about hypothesis testing.

EXERCISE

Assignment!

Q&A and Recap

Please ask if you have any questions now.

Otherwise, we can move on to the recap.

R BASICS

- **Basic ideas in R**
  - Variables/objects, vectors / lists
  - Types of data (numbers, integers, Booleans, …)
  - Some good practices when writing code
  - Some useful functions

```
mean(x)       var(x)     sd(x)
boxplot(x)    hist(x)    quantile(x, ...)
sort(x)       seq(start, end, ...)
```

Data Distribution

Discuss different types of data distribution

Talk about normal distribution and why it is important

Box plots and Outliers

DISTRIBUTION OF THE DATA

1. **What?**
   - An arrangement of values of a variable showing their observed or theoretical frequency of occurrence
2. **Why?**
   - Shows how frequent each value is in a given data set
   - Enables us to get a better sense of the data than what just the numbers in the tables suggest
3. **How?**
   - *Bar charts / histograms / density plots / box plots*

PROBABILITY

Random Variable

A variable whose value is the outcome of a random event.

Probability

A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

$$P(E) = \frac{no.of\ favorable\ event\ E}{total\ no.of\ possible\ events} = \frac{n(E)}{N}$$

PROBABILITY - II

$$P(E) = \frac{n(E)}{N}$$

Where

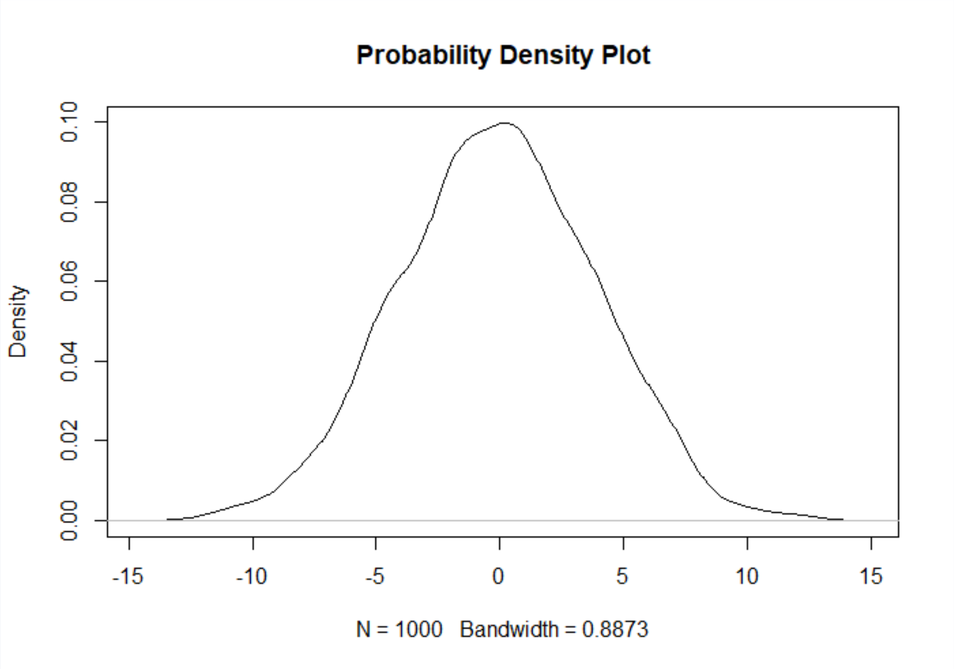
$P(E) =$ $probability\ of\ favorable\ event\ E\ occuring$

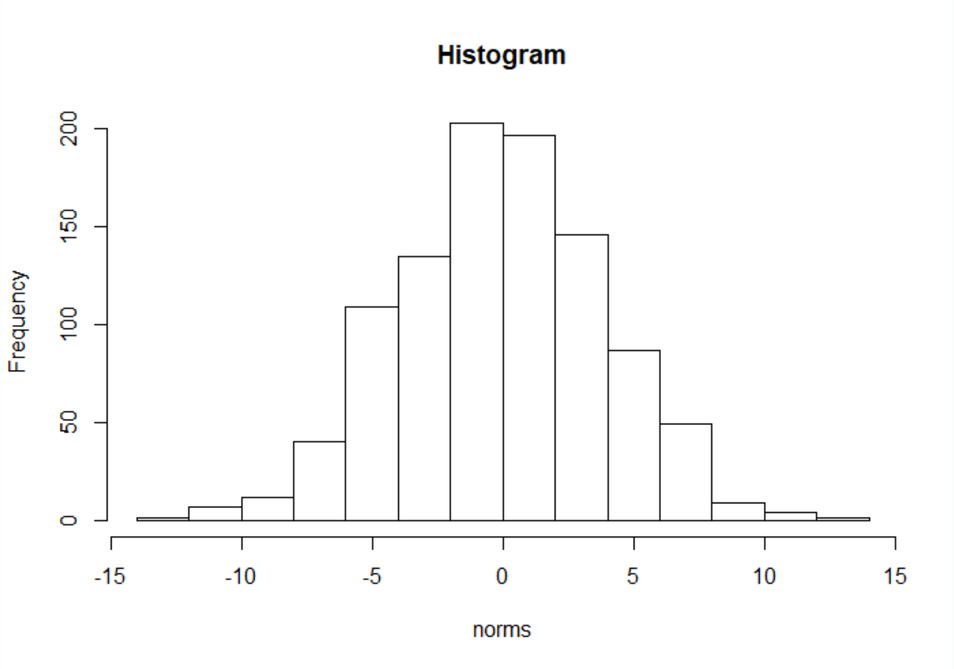
$n(E) =$ $the\ no.of\ favorable\ events\ E$

$N =$ $no.of\ possible\ events$

PROBABILITY DISTRIBUTION

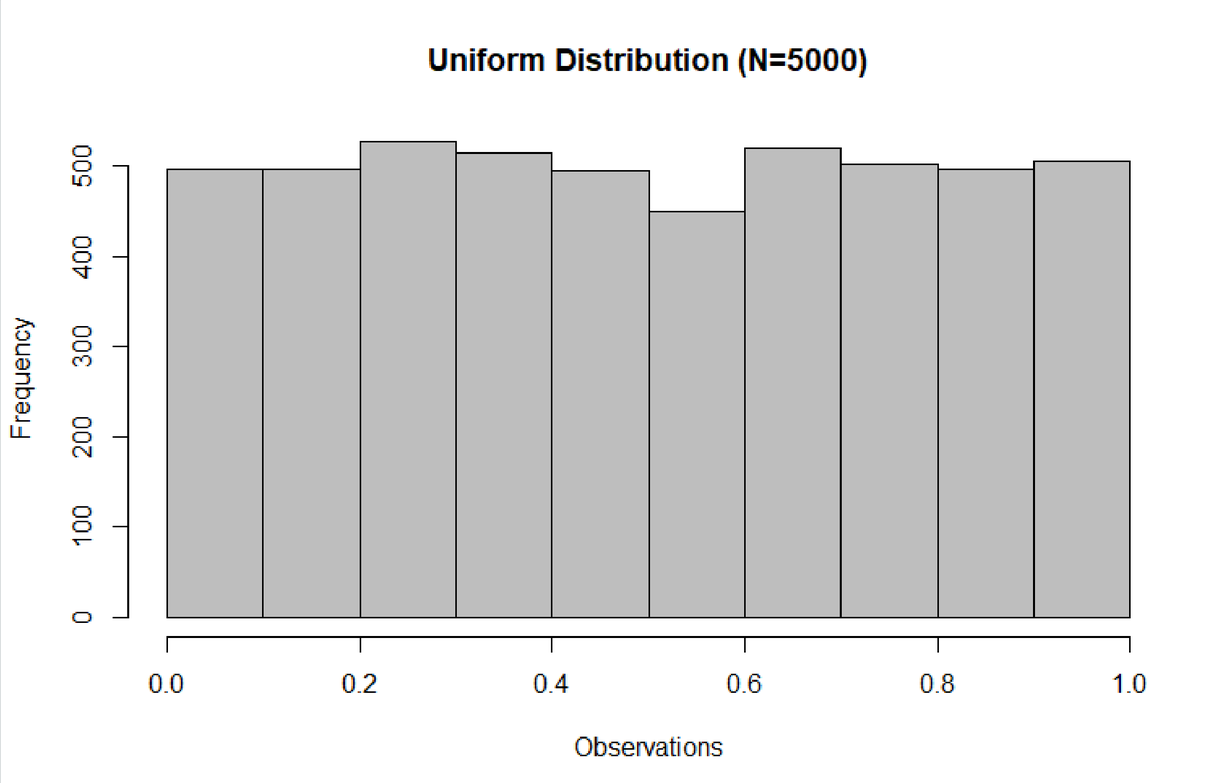
A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.





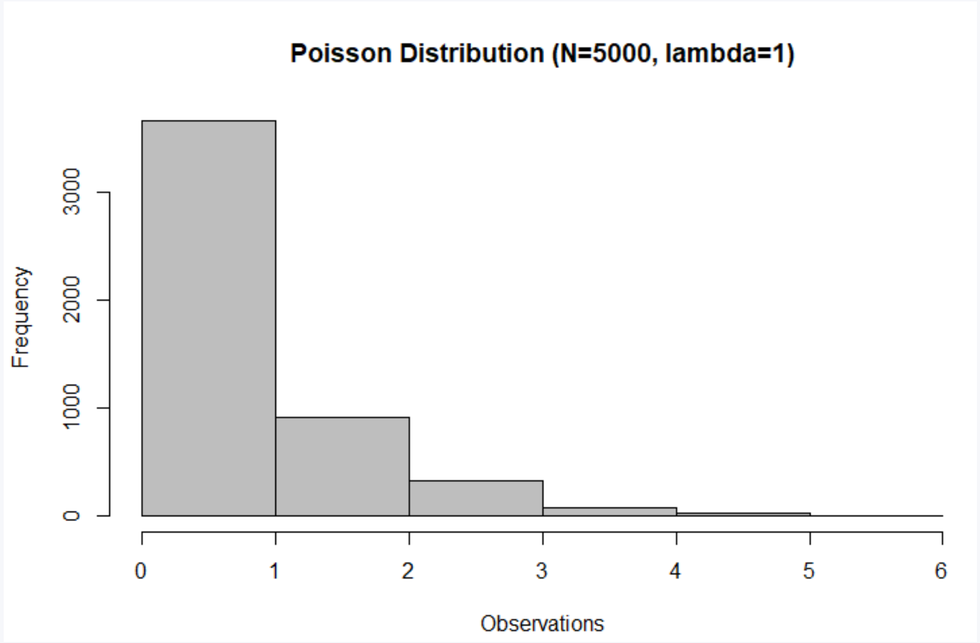
UNIFORM DISTRIBUTION

- Signify probability distribution with equally likely outcomes
- Looks (relatively) flat



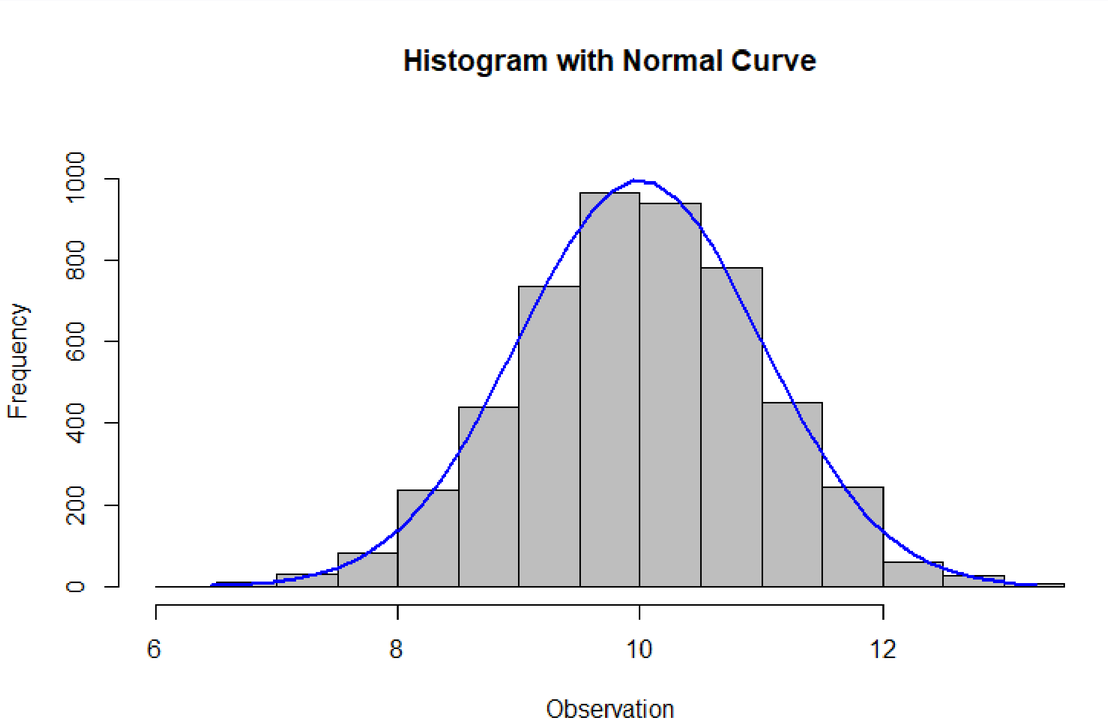
POISSON DISTRIBUTION

- expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time, or space since the last event



NORMAL DISTRIBUTION

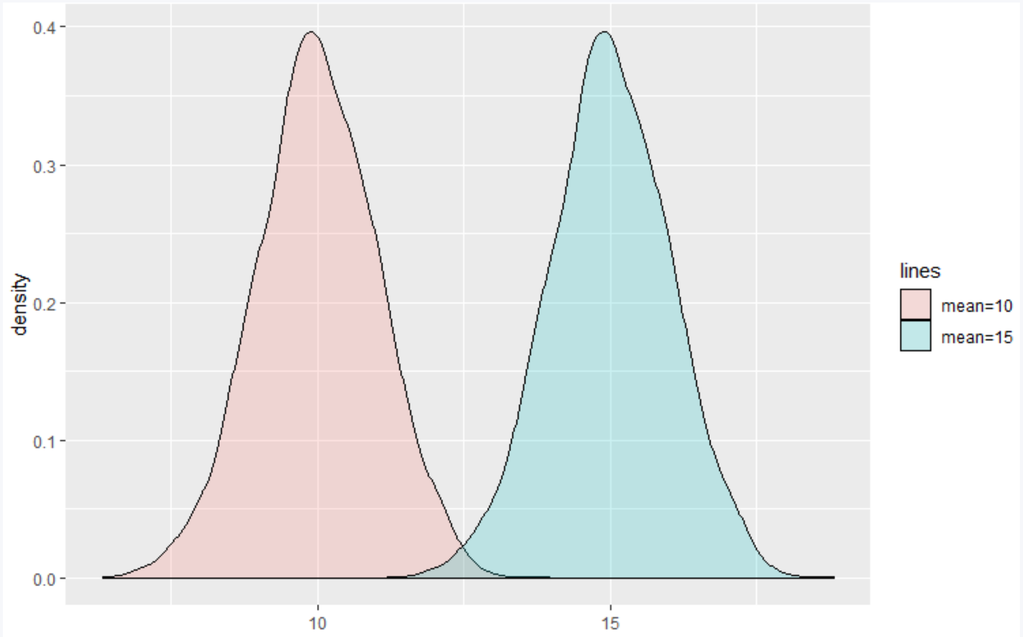
- Most values lies close to the mean
- Variance governs the spread of the values
- Symmetric, but can also be skewed



NORMAL DISTRIBUTION

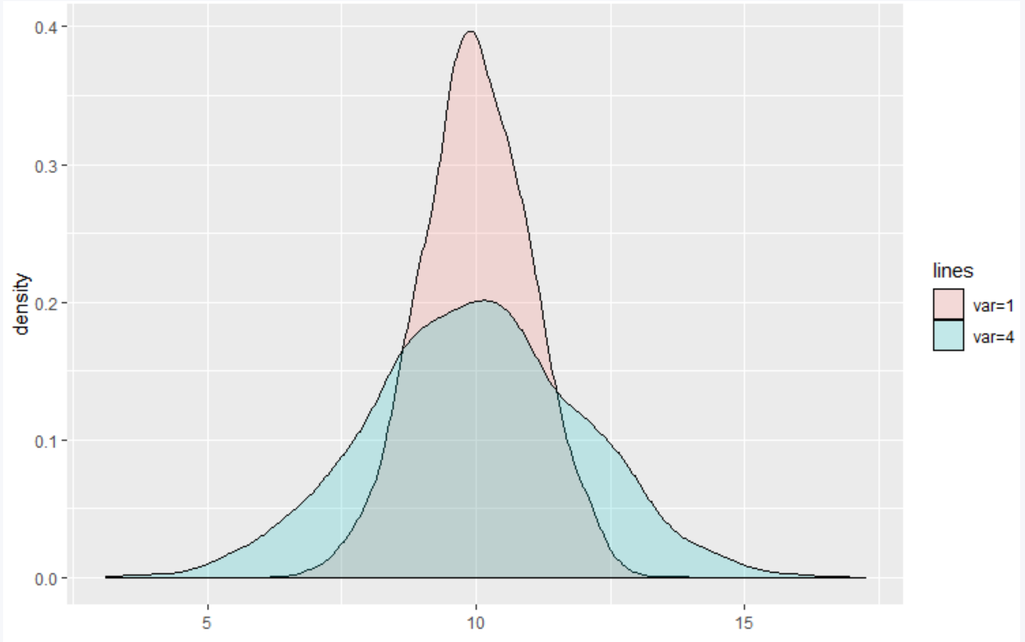
- Parameters: mean and variance

Difference in the mean



$$\bar{x} = \frac{\sum x}{N}$$

Difference in the variance



$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

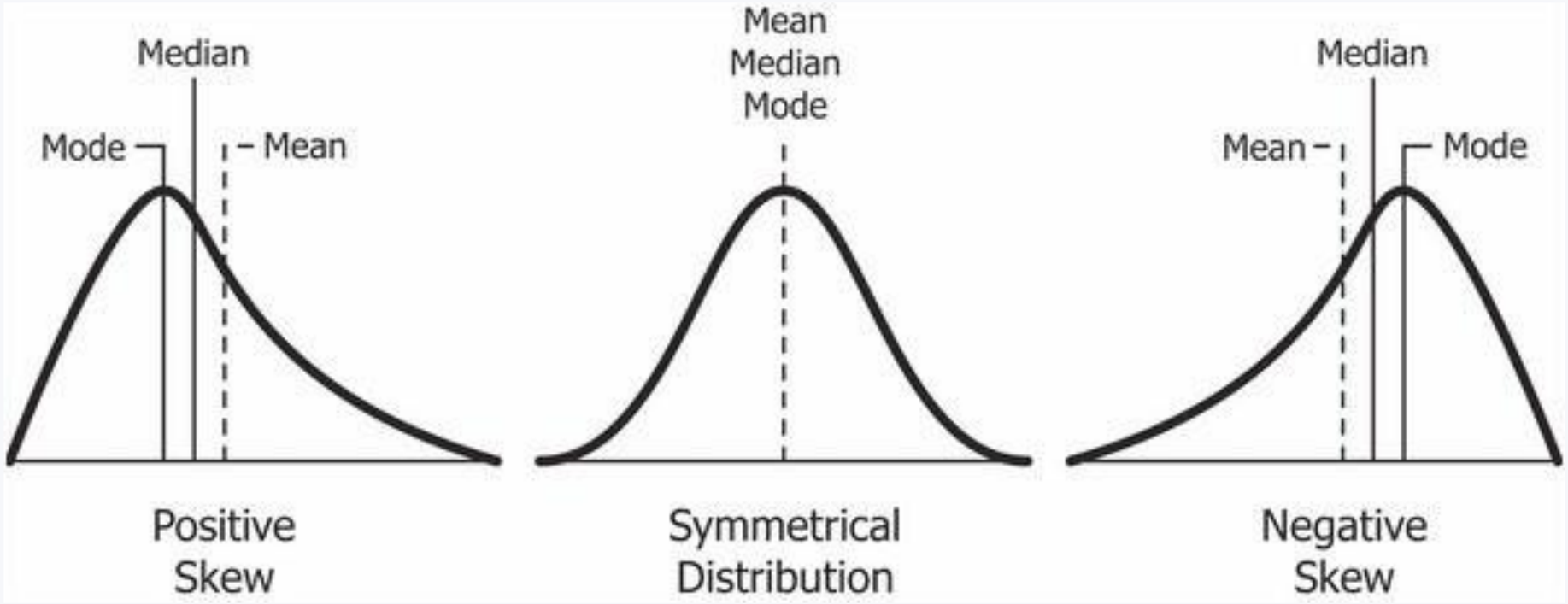
STANDARD NORMAL DISTRIBUTION

The Normal Distribution

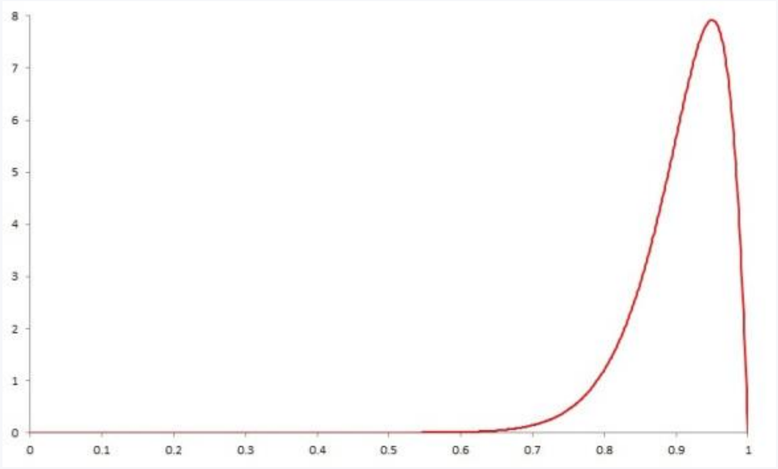
Z-Score, Standardization, Standard Normal Distribution

Reading:
Standard Normal Distribution

SKEWED DISTRIBUTIONS

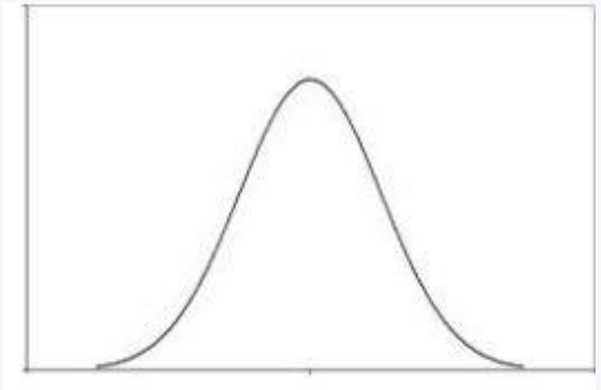


REVIEW: SKEW OF DISTRIBUTIONS



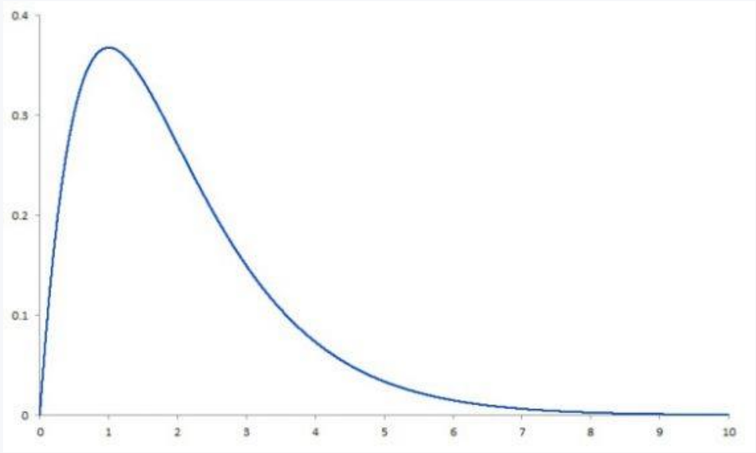
Left/Negative Skew

mean < median < mode



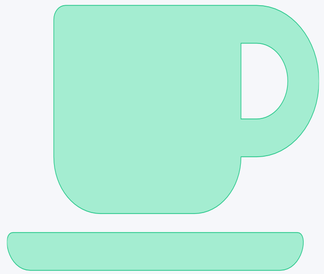
Symmetric

mean = median = mode



Right/Positive Skew

mode < median < mean



5 Minutes Break

Hypothesis Testing

Process of Hypothesis Testing

Significance Level (α)

$\mathbf{p-value}$

PROCESS FOR HYPOTHESIS TESTING

Step 1	Specify the hypotheses (H_0, H_1)
Step 2	Define a sample-based test statistic and the rejection region for the specified H_0
Step 3	Collect the sample data and calculate the test statistic
Step 4	Decide to either reject or fail to reject H_0
Step 5	Interpret the results/make recommendation for action

FORMULATING HYPOTHESES

- Null hypotheses are what you set out to prove wrong.
- Null and alternative hypotheses are mutually exclusive.
- You cannot *accept* a null hypothesis.
- You can either *reject a null hypothesis* or you can *fail to reject it.*
- Practice Exercise:
  Khan Academy Quiz

Alternative Hypothesis H_1:			**Null Hypothesis H_0:**
Symbol	Clue words	Type of test	**Symbol**
<	Less than, decreased, faster	Left tailed Test	≥
>	More than, increased, slower	Right tailed Test	≤
≠	Not equal to, has changed	Two Tailed Test	=

SIGNIFICANCE LEVEL (α)

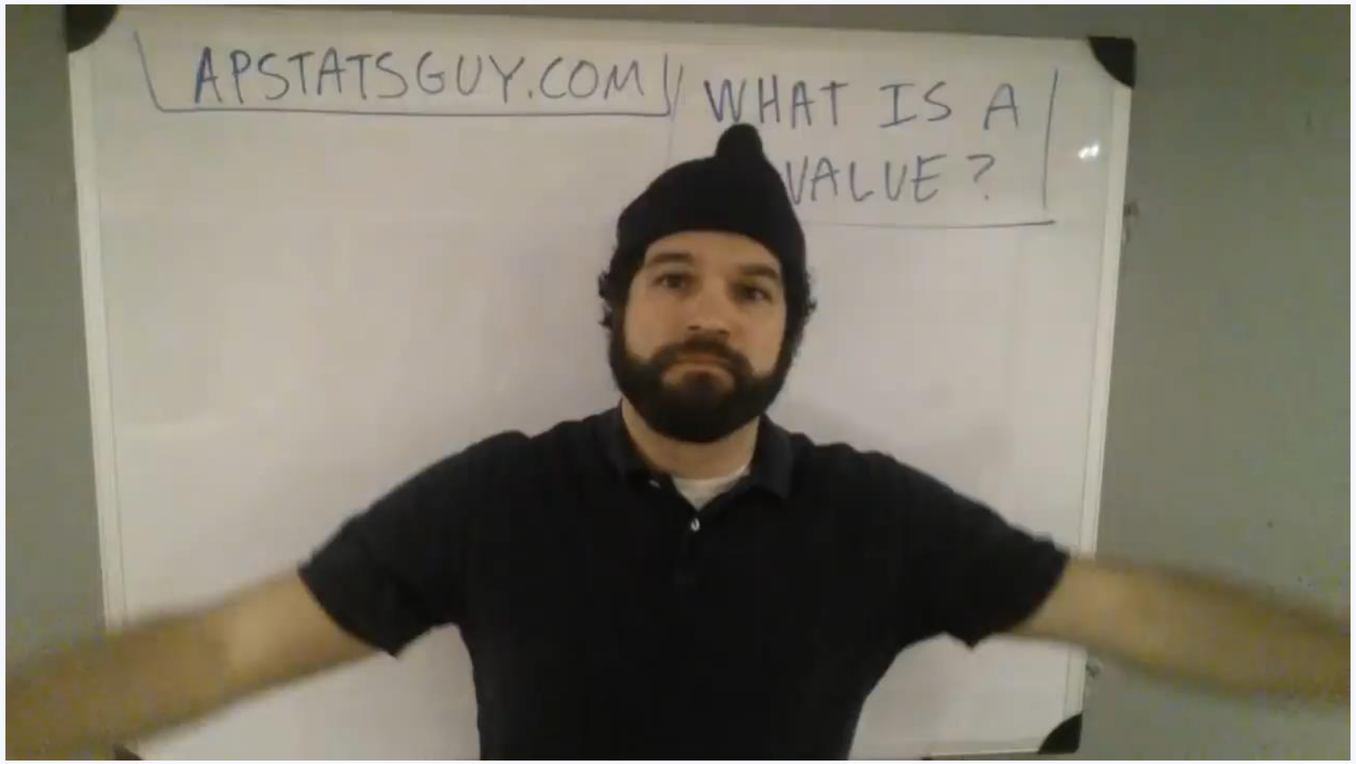
- α is used to set the rejection region.
- It denotes the probability of making a Type I Error (rejecting true H_0)
- The significance level (α) should be low so that the risk of incorrectly rejecting H_0 is minimized.
  (typically, $\alpha = 0.10\ or\ 0.05\ or\ 0.01$)

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

p-value

probability of you making the observations **if H_0 were true**

$$p - value = P(data \mid H_0 \; is \; true)$$



Video source: https://www.youtube.com/watch?v=-MKT3yLDkqk

COMPARING (α) AND $p - value$

When $p - value \leq \alpha$, we reject H_0

- The result is statistically significant
  - We are reasonably sure that there is something besides chance that gave us an observed sample

When $p - value > \alpha$, we fail to reject the H_0

- The result is not statistically significant.
  - We are reasonably sure that our observed data can be observed by chance alone

Exercise

Calculate variance and standard deviation.

Use R for simple data analysis.

ASSIGNMENT

Download the exercise from MyStudy.

You can work with your friends on the task.

PLAN FOR NEXT WEEK

That’s it for today! :-)

Next week, we are going to discuss:

- Chi-squared Test,
- Test of Normality
- t-Test

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`