

[Image Source](#)

Welcome to Week 15, Lecture 01!

Intro to Distributions &
Hypothesis Testing



06/06/22

Agenda

- Announcements
- Assignments & Belt Exam Eligibility
- Intro to Distributions
- Intro to Hypothesis Testing
- CodeAlong: From Distributions to Hypotheses



Announcements

- **Optional Class - Reviewing Mock Belt Exam Solution**
 - Thursday, 06/09/22 @ 5 pm PST.
 - This will replace the code review time slots for Thursday
 - It is completely optional.
 - It will be recorded.

Assignments & Belt Exam Eligibility

This week's assignments Due Sunday by 11:59 PM PST.

- Project 3 Part 3(Core)
- Describing Distributions (Core)
- Hypothesis Testing with Insurance Data(Core)

Remember that Belt Exam eligibility is this Friday at 9AM PST

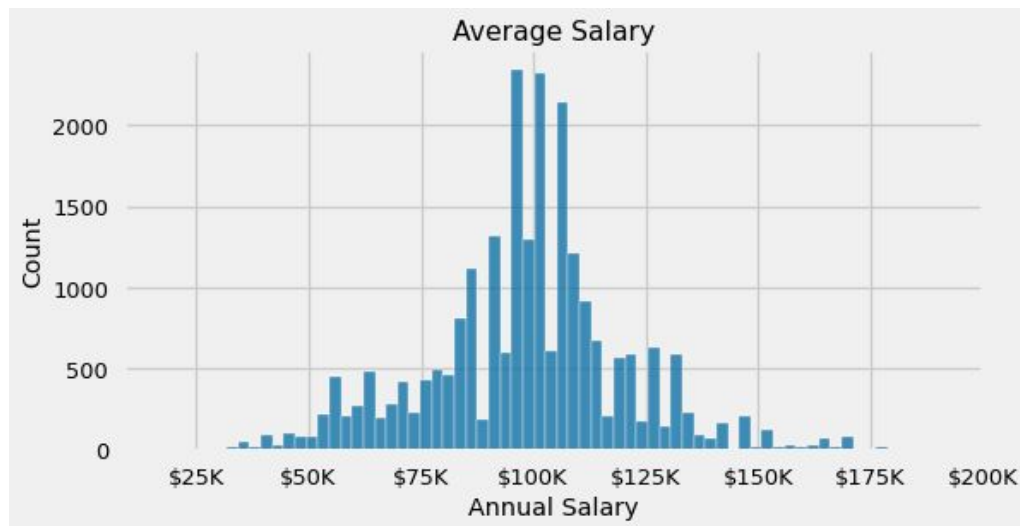
- Make sure you have **all of week 1 and 2 assignments submitted** and **any resubmits from week 1 addressed.**
- Note that content from this week **(week 3 of the stack) is assessed on the exam!**

Intro to Statistical Distributions

What is a “distribution”?

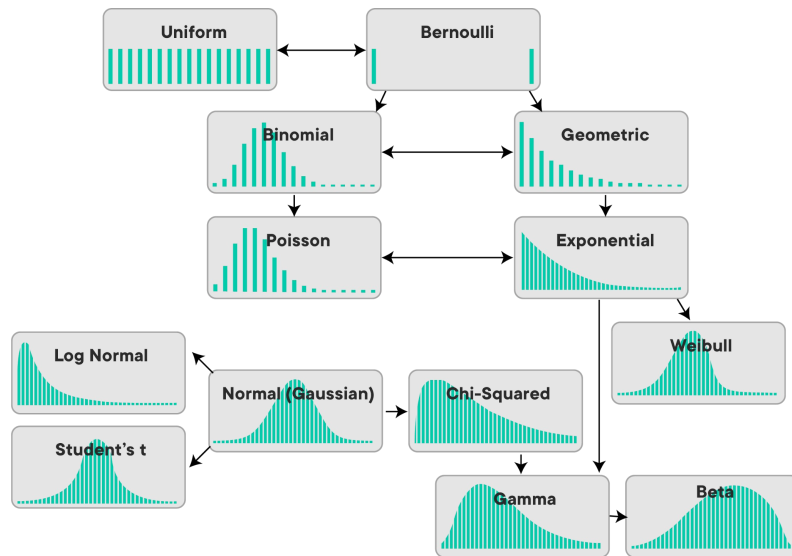
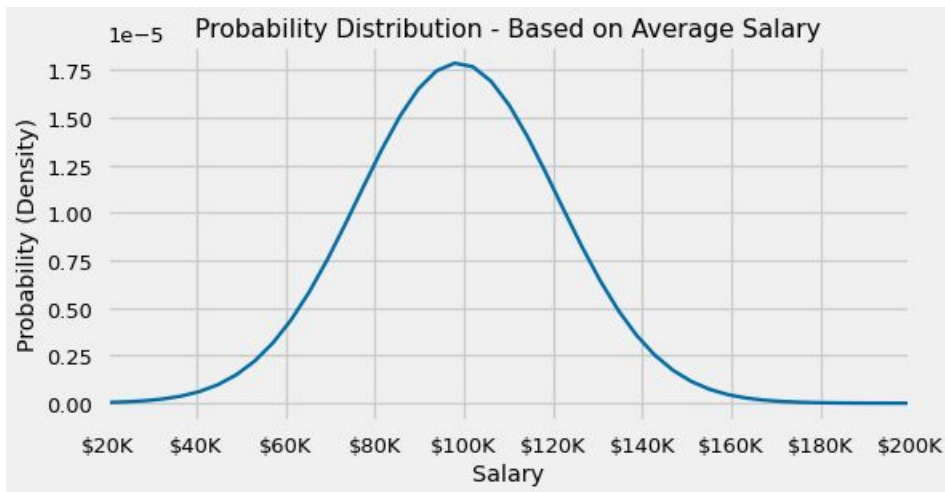
“A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.”

— Page 6, [Statistics in Plain English](#), Third Edition, 2010.



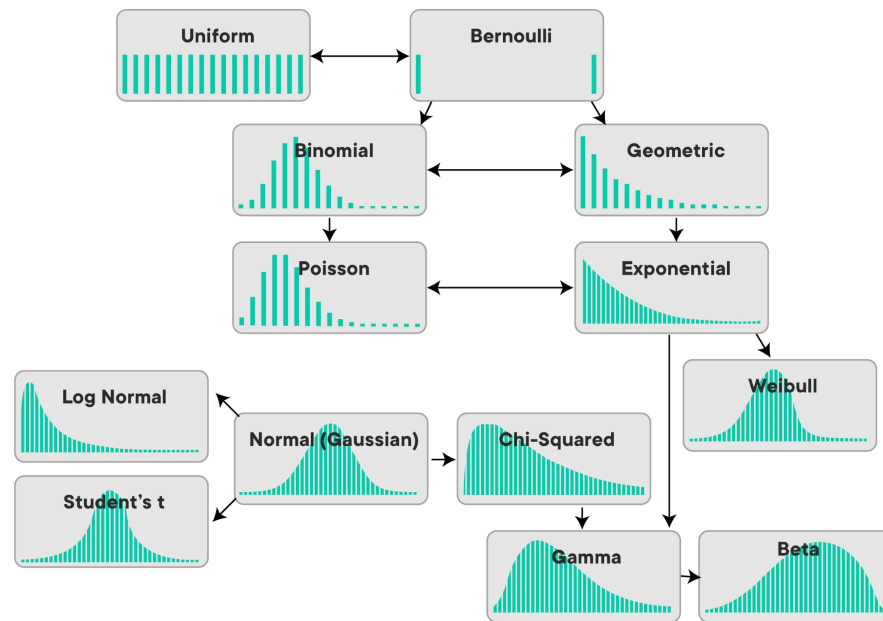
What is a "probability distribution"?

- A mathematically-defined archetype of a data distribution with a specific shape.
- Prob Distributions are plotted as "density" - which is basically probability.



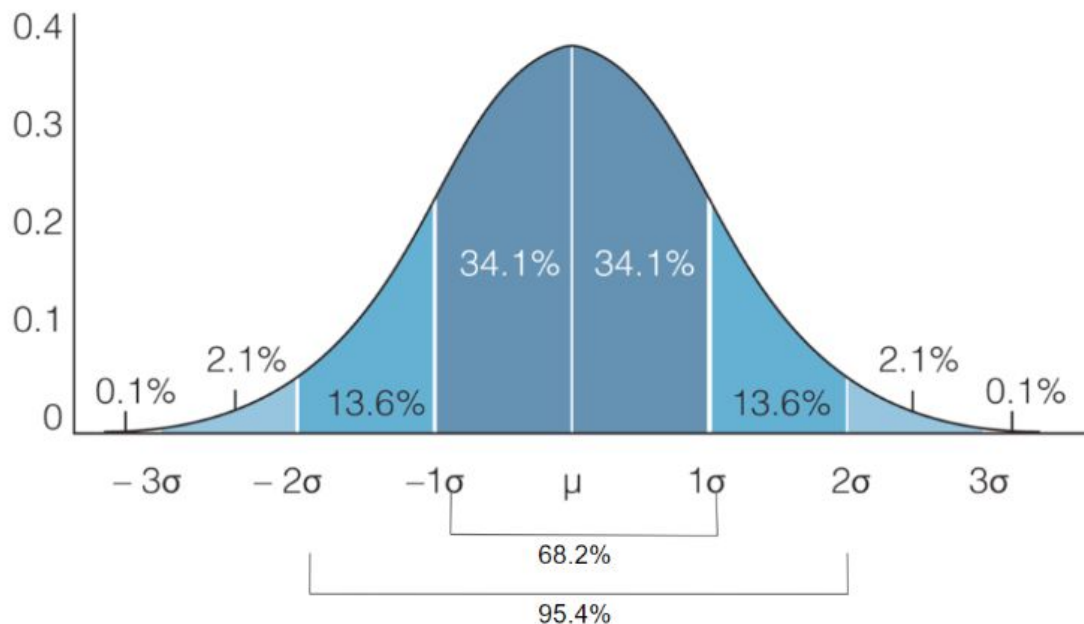
Probability Distributions - Why do we care?

- **Once we know if our data matches / “follows” a certain distribution:**
 - We can calculate probability of specific values occurring.
 - Can use just 1-2 parameters to define the distribution (e.g. mean and standard deviation)
- **Because of a natural phenomenon called the “Central Limit Theorem”,** we can use a normal distribution to calculate the probabilities of observing specific means.







The (Standard) Normal Distribution

- **Normal distributions can be “standardized”** by:
subtracting the mean from the feature and dividing by the standard deviation.
- The data is now in “Z-scores”/“standard deviation units”, not the original units.
- We know what % of the distribution should occur between any 2 points on the x-axis by taking Area Under the Curve



Additional Resources

-  **A Gentle Introduction to Statistical Data Distributions:**
 - Article Link: <https://machinelearningmastery.com/statistical-data-distributions/>
-  **What is a Distribution in Statistics?**
 - Article Link: <https://365datascience.com/tutorials/statistics-tutorials/distribution-in-statistics/>
-  **StatQuest: Statistics Fundamentals**
 - Playlist Link: <https://youtube.com/playlist?list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9>
 -  **If You Only Watched 30 Minutes Worth:**
 - The Main Ideas Behind Probability Distributions (5 mins): <https://youtu.be/oI3hZJqXJuc>
 - The Normal Distribution (5 mins): <https://youtu.be/rzFX5NWoip0>
 - Sampling From a Distribution - Clearly Explained! (3 mins): <https://youtu.be/XLCWeSVzHUU>
 - Hypothesis Testing and the Null Hypothesis - Clearly Explained (<15 mins): <https://youtu.be/0oc49DyA3hU>
 - **Honorable Mentions:**
 - Central Limit Theorem - Clearly Explained (7 mins): <https://youtu.be/YAIJCEDH2uY>
 - p-Values: What they are and how to interpret them (11 mins): <https://youtu.be/vemZtEM63GY>

Intro to Hypothesis Testing

Hypothesis Testing: Getting Started

- Is there actually a *significant* difference between two groups or are the differences just due to randomness?
- Example:
 - A survey goes out and asks students to rate Jupyter notebooks and Google Colab each on a scale of 1-5.
 - On average, Jupyter scores higher, but can we conclude that students really do prefer Jupyter notebooks or is it just due to randomness?

State the null and alternate hypothesis

Null Hypothesis:

- The null hypothesis is the one that seems like it doesn't need to be stated! It is that there is nothing special going on. In our case:
- **Null Hypothesis (H_0):** There is NO difference between ratings for Jupyter notebooks and Google Colab

Alternate Hypothesis:

- The alternate hypothesis is usually what you would just think of as “the” hypothesis! It states that something significant is going on.
- **Alternate Hypothesis: (H_a)** There is a significant difference between ratings for Jupyter notebooks and Google Colab

Establish the significance Level (alpha)

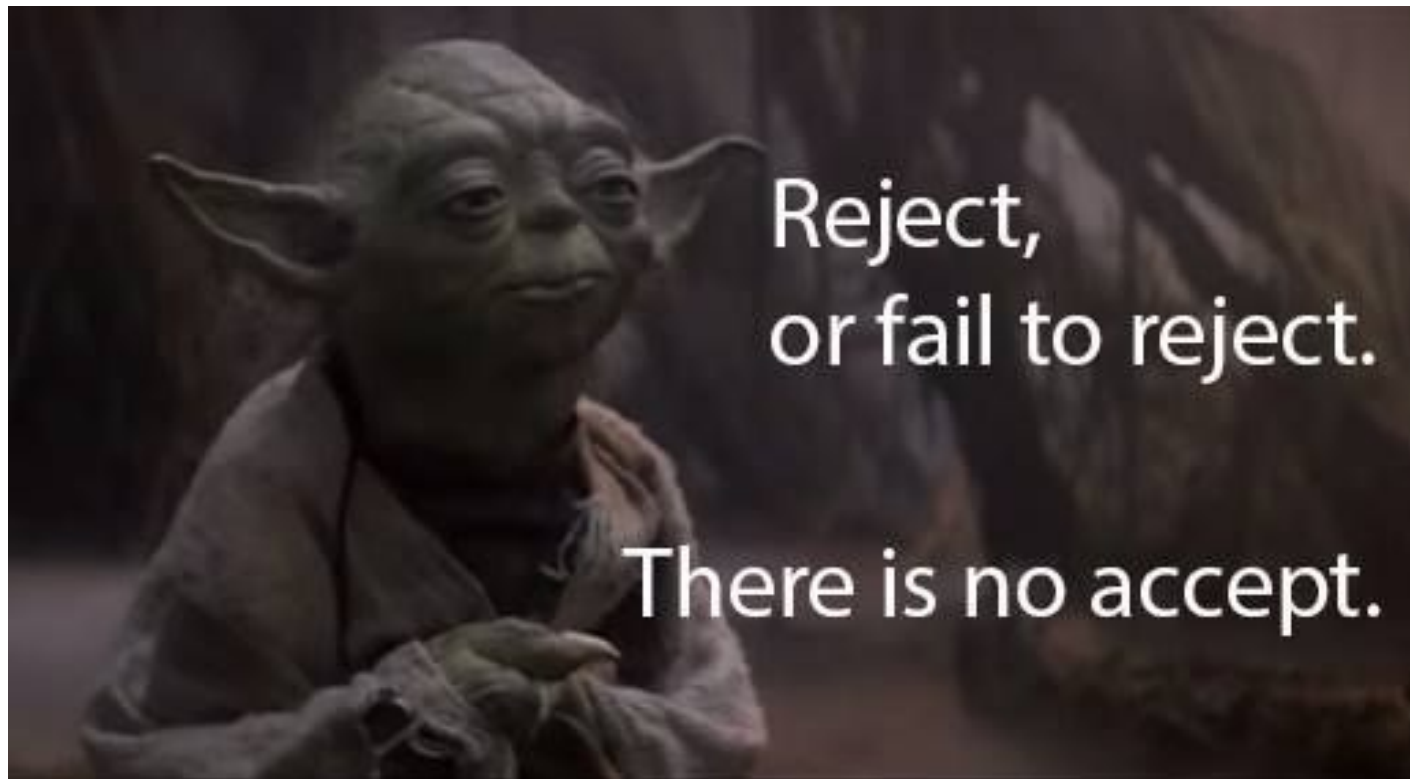
- The most common significance value is $\alpha = 0.05$.
- This means that if the likelihood of the results due to random chance (p) is less than 5%, we consider the results significant (and not just random).
- If we use a significance value of 0.01, it means that the likelihood of the results due to randomness (p) must be less than 1% in order for us to claim it is significant.
- Setting an alpha value establishes your willingness to accept Type 1 or Type 2 errors, and really it depends on your data and situation. When in doubt, consult a SME!

[Examples of Type 1 and 2 Errors with Hypothesis Testing](#)

Perform the test to get a p value


There are many different statistical tests, but the interpretation is generally the same!

- Whichever statistical test you use, you will be given a p-value in your results.
- The p-value is the probability of your situation (like the difference in ratings for Jupyter and Colab) occurring at random.
- If your p is $< \alpha$, you have a significant result and reject the null hypothesis.
 - You conclude there is a significant difference in the ratings.
- If your p is $> \alpha$, you do not have a significant results and “fail to reject” the null hypothesis.
 - You conclude there is NOT a significant difference in the ratings.



CodeAlong: From Distributions to Hypotheses

CodeAlong Repository

- <https://github.com/coding-dojo-data-science/data-enrichment-wk15-lect01-codealong>
 -  The “PreClass - From Distributions to Hypothesis.ipynb” notebook.
 - Post Class Notebook [06/06/22]:
<https://github.com/coding-dojo-data-science/data-enrichment-wk15-lect01-codealong/blob/06-06-22-class/Class%20-%20From%20Distributions%20to%20Hypothesis.ipynb>
 -

Next Class: Selecting the Right Test

Test Selection Overview

- **STEP 1: Stating our Hypothesis**
- **STEP 2: Determine the category/type of test based on your data.**
 - Answer 2 questions about what you are comparing to determine which test.
 - Once you know which test, check the assumptions of the test.
- **STEP 3: Does the data meet the assumptions of the selected test?**
 - If you meet the assumptions, run the desired test.
 - If you don't meet the assumptions, run the non-parametric equivalent test.
- **STEP 4: Perform Test & Interpret Result**
 - Run the selected statistical test to get your p-value and interpret it.
 - Do you reject your null hypothesis or did you fail to reject it?
- **STEP 5: Post-hoc multiple comparison tests (if needed)**