Data Analysis for Business Analytics

CIS621

Week 5 - Statistical Distributions

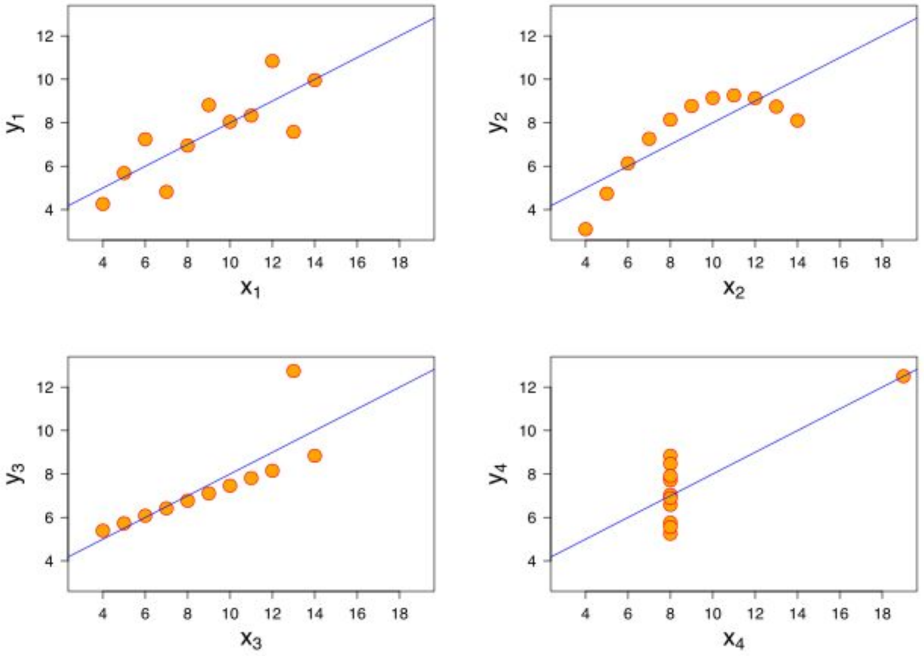
Lecture Goals

1. Understand why we look at distributions.
2. Understand the normal distribution, its basic properties, and its importance.
3. Understand other common types of distributions and when they might occur.
4. Understand common statistical tests to determine if a sample of data follows a certain type of distribution (e.g., is this data normally distributed?).

This lecture can be viewed as part one of a two-part series. The lecture next week will extend these ideas to talk about statistical significance - which examines the relationship between multiple data distributions.

Anscombe’s Quartet

Image Credit - Wikipedia



How Summary Statistics Can Lead Us Astray

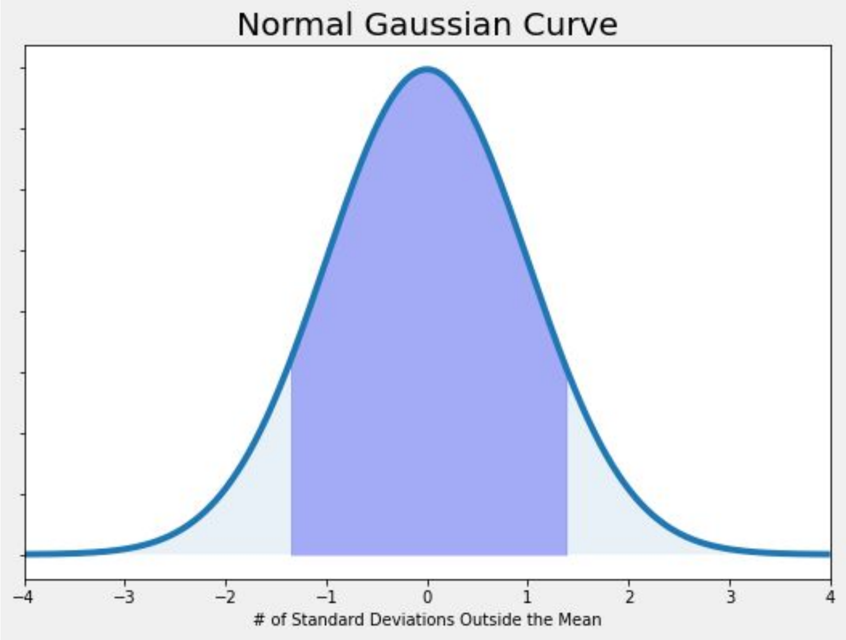
- *Summary* statistics are just that - a *summary*.
- They are meant to remove complexity and give us a quick, high-level understanding of our data.
- Therefore, they hide nuance and patterns.

Why We Look at Data Distributions

- Data distributions tell us a bigger story.
- How does our data *look*?
- Do we have lots of spread? If so, in what direction(s)?
- How many “rare” events do we have?
- What’s the behavior around the mean?
- Do we perhaps have multiple modes?
- Do we have interesting peaks and valleys?

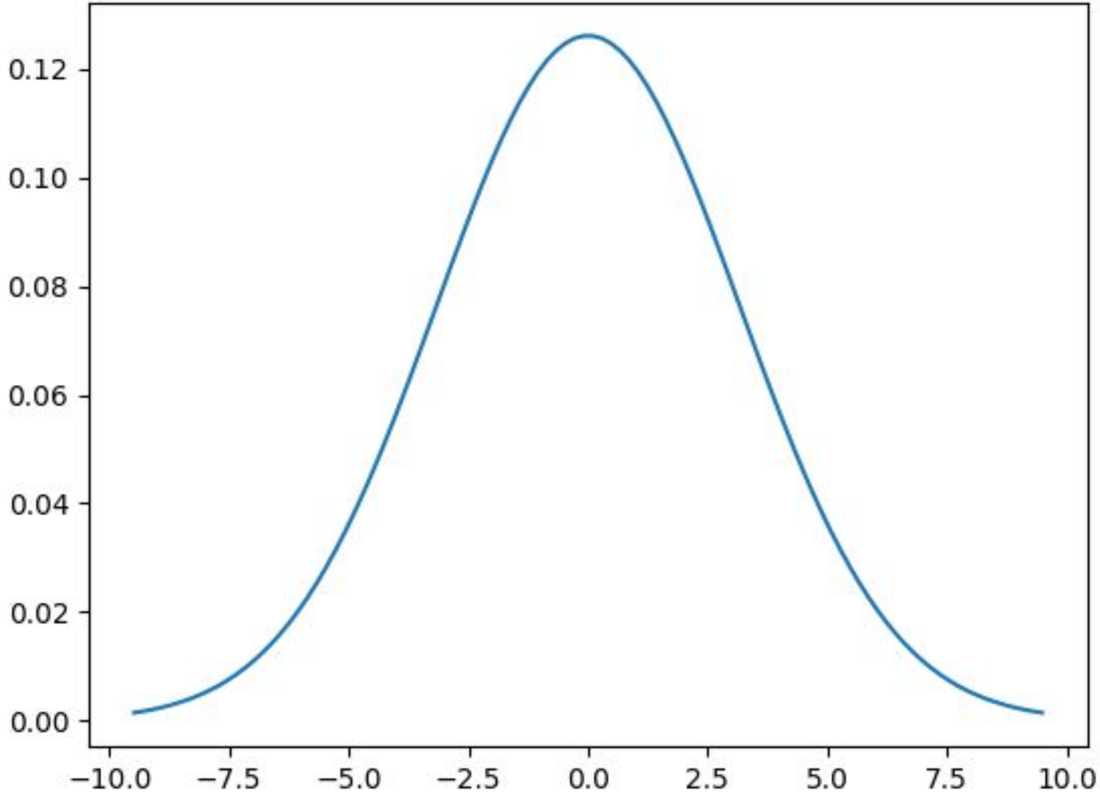
Normal Distribution

- The normal distribution is the most important distribution to know. Many statistical methodologies assume a normal distribution. In sum, this distribution appears everywhere.

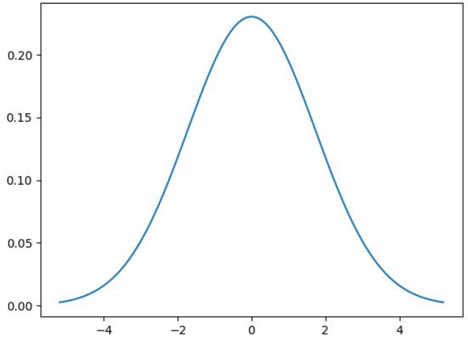


Normal Distributions with Different Means and Standard Deviations

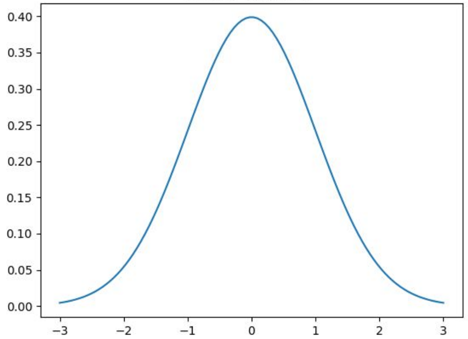
- To construct a normal distribution, we need to know the mean and the standard deviation of our sample.



Mean 0 Sigma 3.16



Mean 0 Sigma 1.73



Mean 0 Sigma 1.0

Empirical Rule

- The following nice rule can be applied to a normal distribution.
  - 68% of observations will fall within one standard deviation of the mean
  - 95% of observations will fall within two standard deviations of the mean
  - 99.7% of observations will fall within three standard deviations of the mean
- If you have normally-distributed data, this provides powerful insight!

Central Limit Theorem

- This theorem can be a bit confusing (even mesmerizing), but it is very powerful to understand. Here is how it works.
- We have a population of data.
- We capture a series of samples from the population.
- We take the mean of each sample.
- As we take more samples, our distributions of means approaches a normal distribution, regardless of distribution of the population.
- Generally, samples sizes of 30-40 are considered “large enough” for the CLT to kick in.

CLT Real-World Application

- You run a test to see if sending prospects a coupon via email entices them to buy.
- You randomly assign the prospects to treatment (received coupon) or control (did not receive coupon) groups.
- You run the test for a week.
- You end up with the following results:
  - Treatment conversion rate: 5.1%
  - Control conversion rate: 4.5%
- Should we celebrate? Not yet.

CLT Real-World Application

- If this test were to happen again in a parallel universe, we would very likely get different results! In fact, the new results would each lie somewhere on a normal distribution! Meaning - the outcome we experienced was not the only outcome but rather *one of the outcomes*.
- **Don’t view results as a single point but rather as a point on a distribution.**
- More concretely, let’s say you re-ran the same trial for 30 weeks in a row. Assuming temporal and exogenous factors are not confounders, the conversion rates of the treatment group would be normally distributed, and the conversion rates of the control group would also be normally distributed. In some cases, then, you might find the control was better than the treatment!

Statistical Significance Preview

- We can use the concept of statistical significance to determine if our results are *different enough* to be called *actually different,* controlling for the fact we will receive normally distributed outcomes. This will be covered more next week - hold your excitement!

Normal Distributions in Everyday Life

- In what may seem a bit mind-boggling, normal distributions appear all the time in everyday life. Here are some items that are normally distributed.
  - Height
  - Birth weight
  - Shoe size
  - Blood pressure
  - ACT scores and IQ

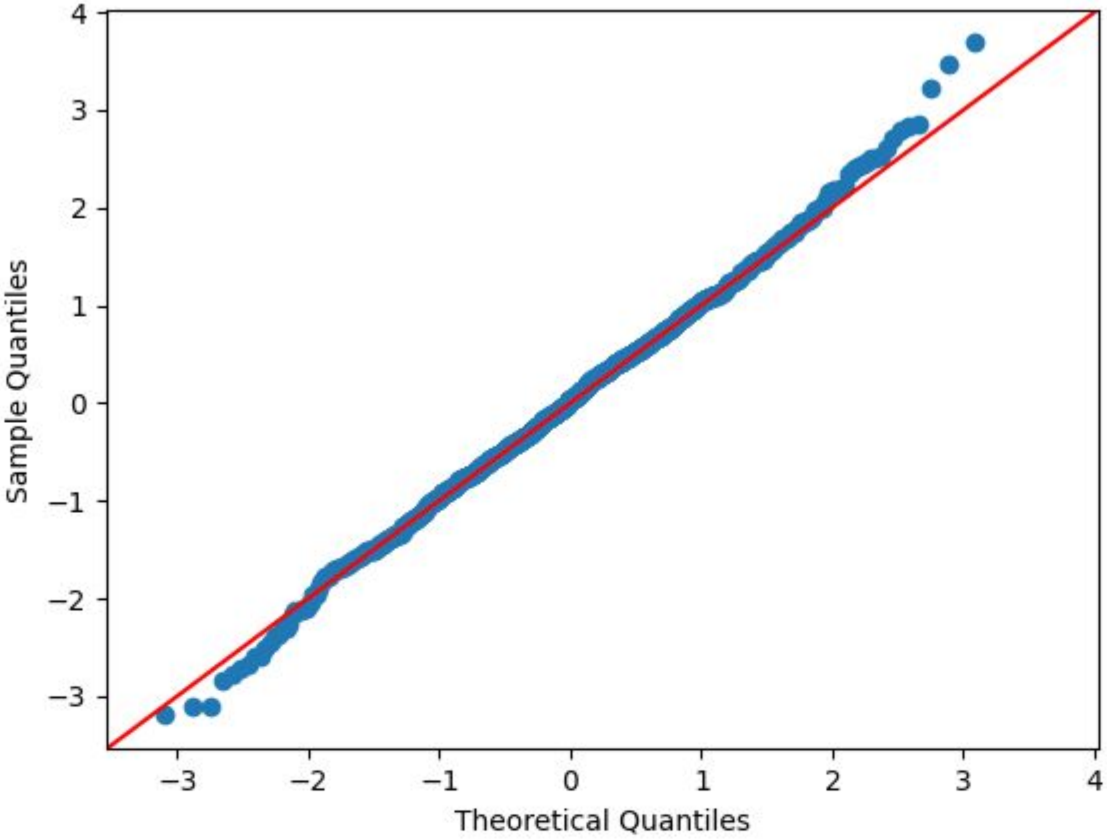
Citation: https://www.statology.org/example-of-normal-distribution/

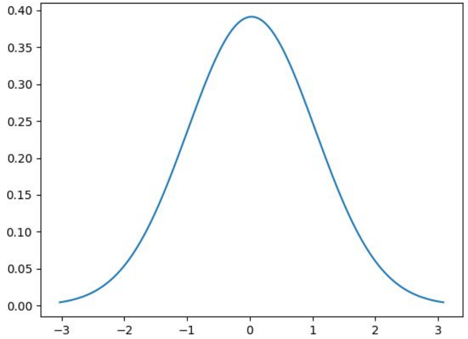
Normal Distributions in Statistics

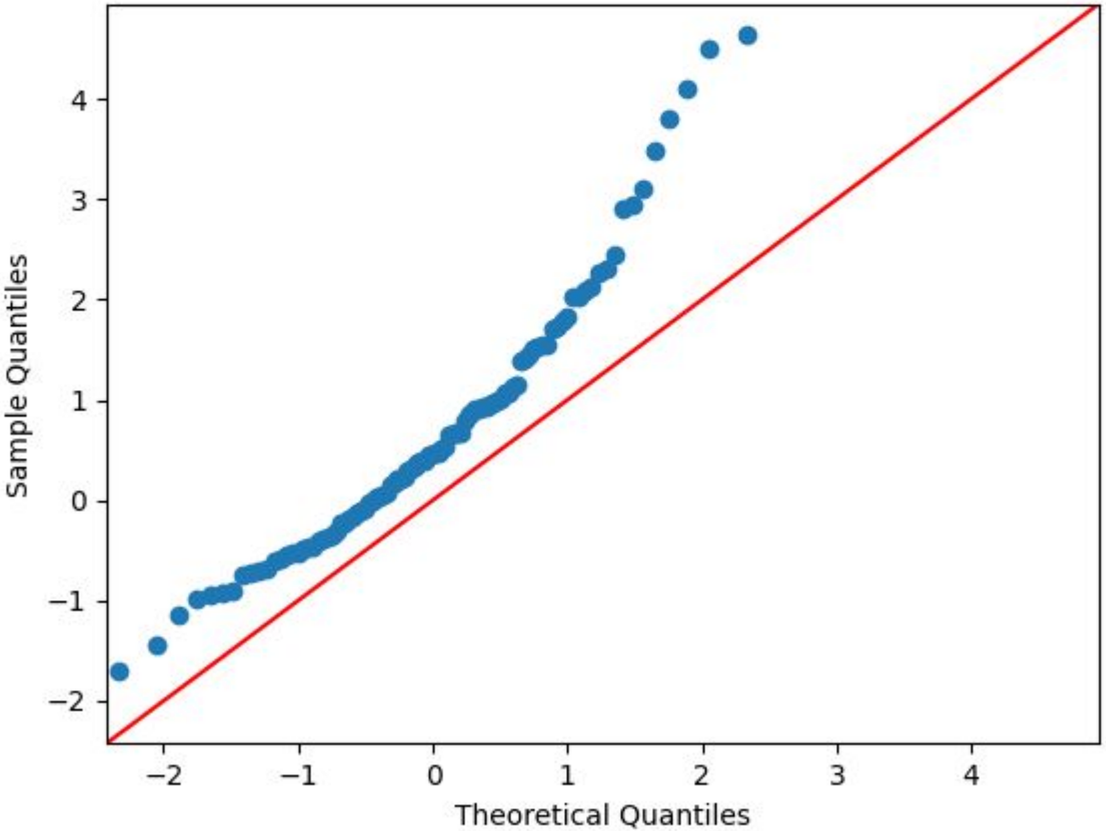
- Linear regression assumes errors follows a normal distribution.
- Many statistical tests, including the t-test for statistical significance, assume normal distributions. These are broadly known as parametric tests since they make assumptions about your data.

Testing for Normality

- Visually, we can use a Q-Q plot to inspect normality. A histogram is a cruder, but potentially quick way to assess normality.
- We could also use a formal statistical test for normality (we will cover this topic more soon):
  - Shapiro–Wilk Test
  - Kolmogorov–Smirnov Test
  - Anderson-Darling Test

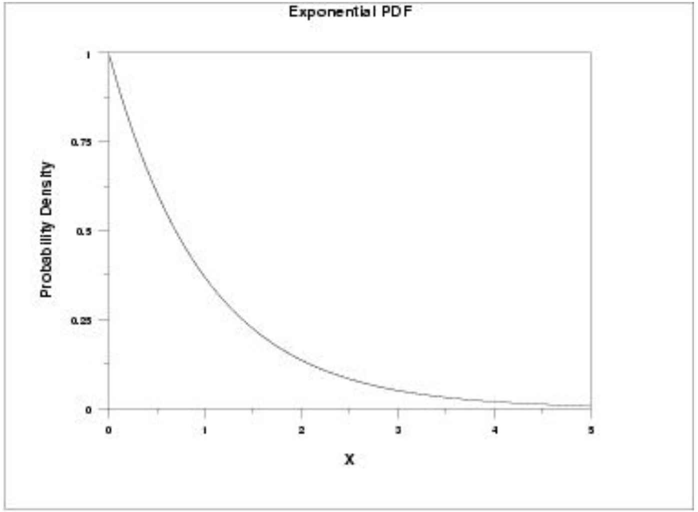

4
3
2
1
0
−1
−2
−3
Sample Quantiles
−3
−2
−1
0
1
2
3
4
Theoretical Quantiles


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3


Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4

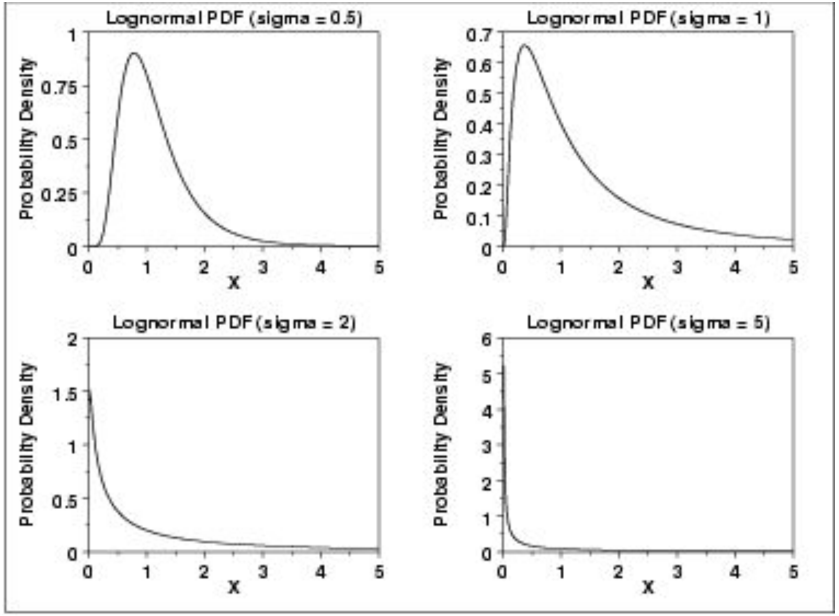
Exponential Distribution

- Phenomena with rapid decay may follow an exponential distribution.
- Class Question: What metrics “in the real world” might follow such a distribution?
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda3667.htm



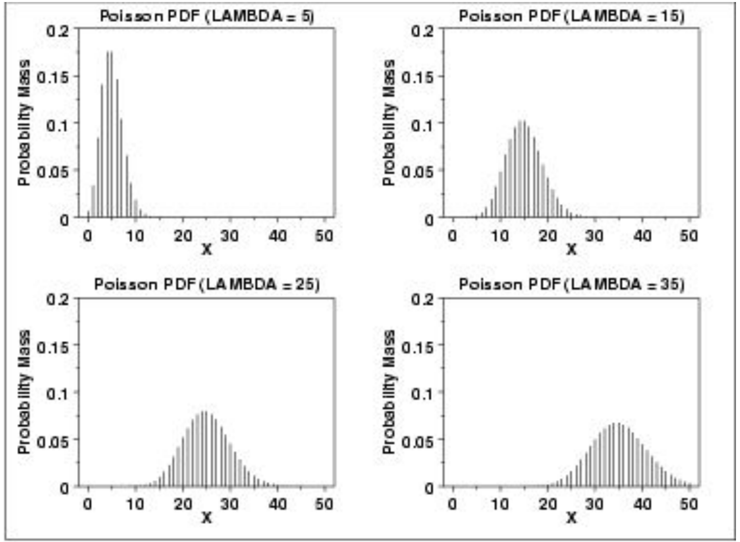
Lognormal Distribution

- Broadly, a lognormal distribution represents phenomena with long tails.
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm



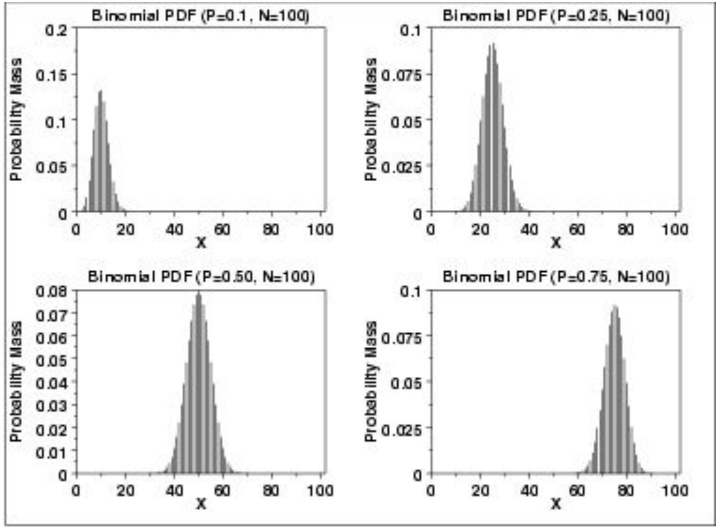
Poisson Distribution

- Displays the probability of a discrete event occurring over a given period of time.
- Class Question: What types of cases might following a poisson distribution?
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda366j.htm



Binomial Distribution

- A binomial distribution shows the probability of “success” in a given time period.
- Class Question: What types of cases might following a binomial distribution?
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda366i.htm



Kolmogorov–Smirnov Test

- The Kolmogorov–Smirnov Test assess if two samples are drawn from the same distribution.
- We can use this test to determine if a sample is, say, normally distributed.
  - We generate a synthetic normal distribution and compare it to our real data.

Anderson-Darling Test

- The Anderson-Darling Test also assess if two samples are drawn from the same distribution. It is an alternative to the Kolmogorov–Smirnov Test. The former is more sensitive to the tails than the latter.

Proportions Z-Test

- How do we compare if two samples of categorical data have the same distribution (count)?
- We can use a Proportions Z-Test!

Plotting Distributions

- Familiar friends:
  - Histogram
  - Density Plot (also known as a PDF plot)
- Cumulative density plot (known as a CDF plot)
- Boxplots
- Stripplots
- Violin plots