

AAIL100: Introduction to Data Science

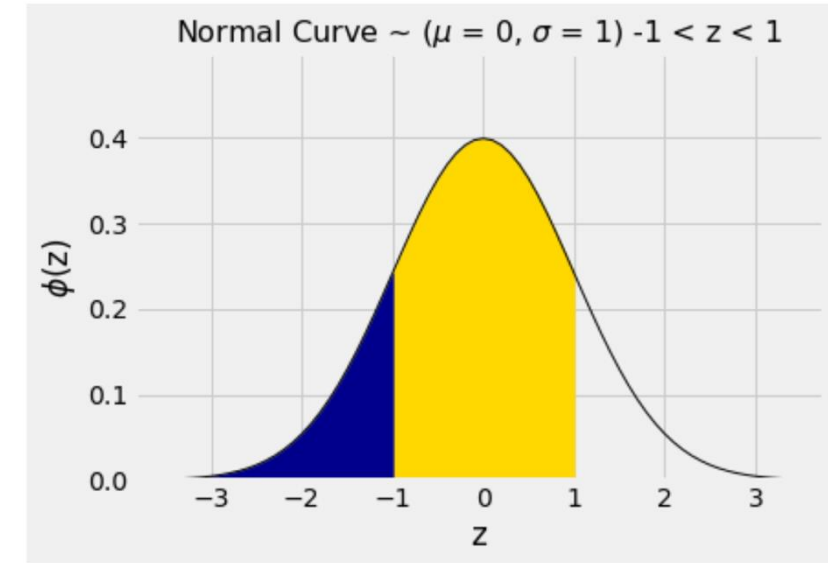
Ragesh Jaiswal, CSE, IIT Delhi

Mean, Normal, and the CLT

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$



Percent in Range	All Distributions: Bound	Normal Distribution: Approximation
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem**.

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*

Mean, normal, and the CLT

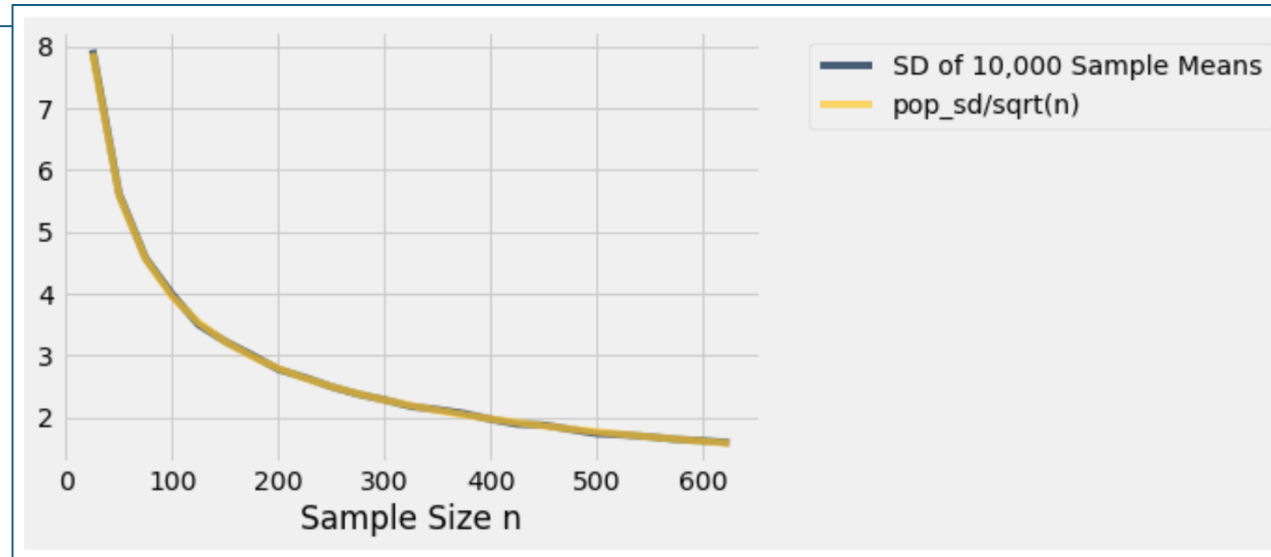
- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$
- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*



Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*
- **Central Limit Theorem for sample mean**: *If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population, the probability distribution of the sample mean is roughly normal, centered at the population mean, with an SD equal to the population SD divided by the square root of the sample size.*

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*
- **Central Limit Theorem for sample mean**: *If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population, the probability distribution of the sample mean is roughly normal, centered at the population mean, with an SD equal to the population SD divided by the square root of the sample size.*
- Question: How do you use the above theorem to obtain more and more accurate estimate of the population mean?

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*
- **Central Limit Theorem for sample mean**: *If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population, the probability distribution of the sample mean is roughly normal, centered at the population mean, with an SD equal to the population SD divided by the square root of the sample size.*
- Question: How do you use the above theorem to obtain more and more accurate estimate of the population mean? (Larger sample size)

Mean, normal, and the CLT

- One of the most commonly observed distributions is a **bell-shaped** distribution.
- In fact, if we plot the histogram of such bell-shaped distributions in **standard** units (so as to ignore the units), then they all resemble the same curve known as the **standard normal curve**.
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$
- Observation 1: Even if the data histogram is not bell-shaped, the empirical histogram of a statistic based on random samples from the data tends to be bell-shaped.
- The formal statement capturing the above observation is known as the **Central Limit Theorem (CLT)**.
 - **CLT**: *The probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.*
- Observation 2: How tightly is the distribution packed around the mean (i.e., SD) depends on the sample size (*larger the tighter*).
 - Question: *Is there a relationship between sample size and sample means?*
- **Central Limit Theorem for sample mean**: *If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population, the probability distribution of the sample mean is roughly normal, centered at the population mean, with an SD equal to the population SD divided by the square root of the sample size.*
- Question: How do you use the above theorem to obtain more and more accurate estimate of the population mean? (Larger sample size)

- Show that a sample size of 40,000 is sufficient to obtain a 95% confidence interval with 1% accuracy for voting data.

Prediction

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable based on the value of another*
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable based on the value of another*
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable based on the value of another*
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable based on the value of another*
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the the scatter plot is a perfect straight line sloping downwards.

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: *Based on a person's internet profile, which websites are likely to interest them?*
- We start with one-variable prediction:
 - *Predicting the value of one variable based on the value of another*
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: *Is there a measure for linear association?*
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: *r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.*

End