

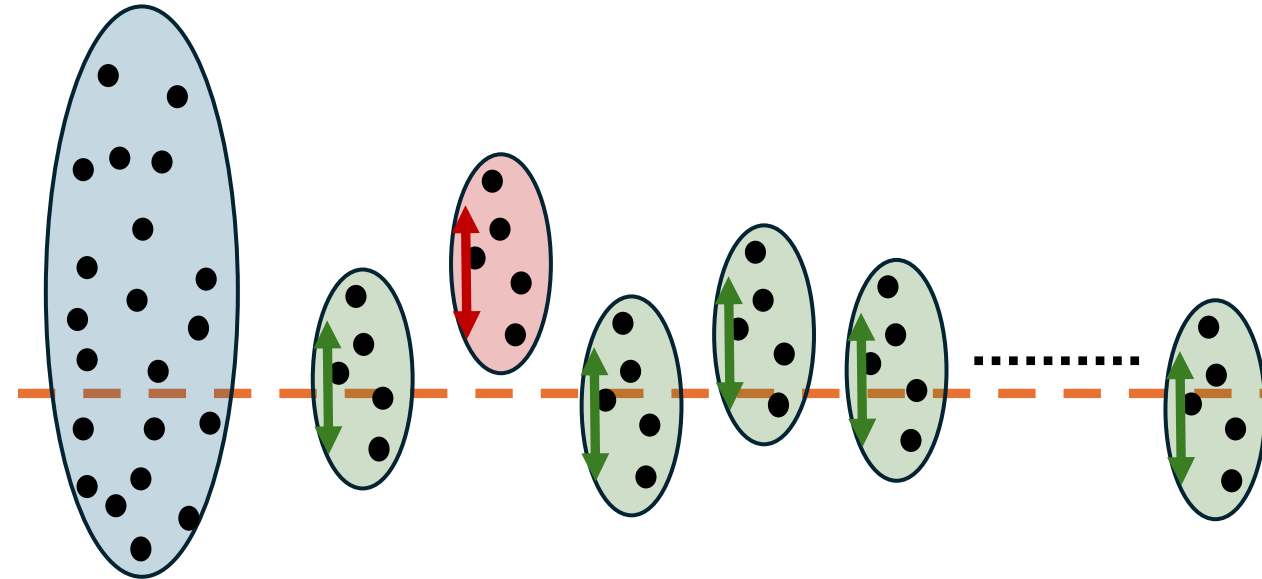
AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue:** Only data from a random sample is available.
 - **Solution:** *Estimate* the parameter from the *random sample*.
 - **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?
 - Typically, no.
 - **Solution:** Obtain multiple random samples.
 - **Issue:** It is typically very difficult to return to the population and obtain another sample.
 - **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
 - Let us see the method known as “*the Bootstrap,*” which provides an interval estimate.
 - Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.



- Interesting duality between p-value and CI:
 - *If you are testing whether or not the population mean is a particular value x , and you use the 5% cutoff for the p-value, then you will reject the null hypothesis if x is not in your 95% confidence interval for the mean.*

Mean, Normal, and the CLT

Mean, normal, and the CLT

- In multiple previous discussions, we have used the *sample mean*.
- *Why is the **mean** the property of choice when analyzing data?*
 - ... because the mean of random samples satisfies a strong property **regardless of the population**.
- In this discussion, we'll try to address the following questions related to sample means:
 - *What exactly does the mean measure?*
 - *How close to the mean are most of the data?*
 - *How is the sample size related to the variability of the sample mean?*
 - *Why do empirical distributions of random sample means come out bell-shaped?*
 - *How can we use sample means effectively for inference?*

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- Answer the following questions related to mean:
 - Is it necessarily an element of the collection?
 - Is it necessarily an integer if all the elements of the collection are integers?
 - Is it necessarily somewhere between the smallest and largest values in the collection?
 - Is it necessarily halfway between the two extremes; is it true in general that half the elements in a collection are above the mean?
 - If the collection consists of values of a variable measured in specified units, then does the mean have the same units too?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- Answer the following questions related to mean:
 - Is it necessarily an element of the collection? (**No**)
 - Is it necessarily an integer if all the elements of the collection are integers? (**No**)
 - Is it necessarily somewhere between the smallest and largest values in the collection? (**Yes**)
 - Is it necessarily halfway between the two extremes; is it true in general that half the elements in a collection are above the mean? (**No**)
 - If the collection consists of values of a variable measured in specified units, then does the mean have the same units too? (**Yes**)

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- Median = *the 50th percentile number*
- Mean Vs Median:
 - For symmetric distributions, the mean and the median are equal
 - If the histogram has a tail on one side (*the formal term is “skewed”*), then the mean is pulled away from the median in the direction of the tail

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$? ($\leq 1/2$)
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq \alpha \cdot \mu\}|}{|C|}$ for $\alpha > 1$?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$? ($\leq 1/2$)
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq \alpha \cdot \mu\}|}{|C|}$ for $\alpha > 1$? ($\leq 1/\alpha$)

Mean, normal, and the CLT

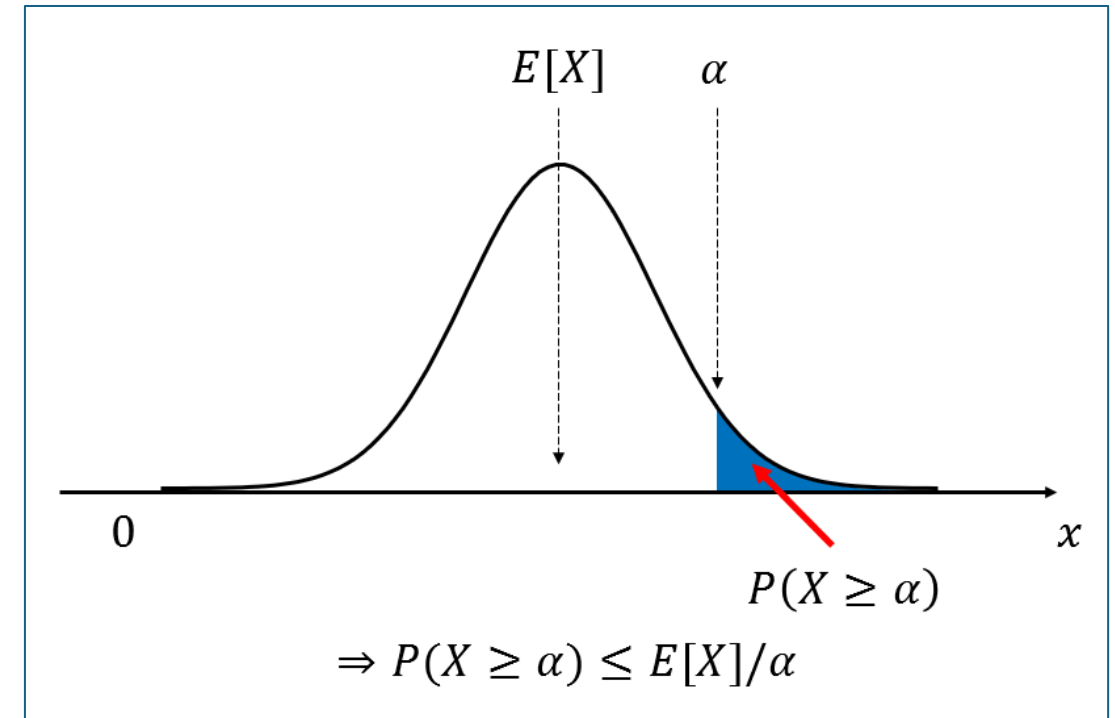
- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$? ($\leq 1/2$)
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq \alpha \cdot \mu\}|}{|C|}$ for $\alpha > 1$? ($\leq 1/\alpha$)
 - Suppose X denotes a random variable that takes values in C with uniform probability. What is $\Pr[X \geq \alpha \cdot \mu]$?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$? ($\leq 1/2$)
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq \alpha \cdot \mu\}|}{|C|}$ for $\alpha > 1$? ($\leq 1/\alpha$)
 - Suppose X denotes a random variable that takes values in C with uniform probability. What is $\Pr[X \geq \alpha \cdot \mu]$? ($\leq 1/\alpha$)
 - This is known as *Markov's inequality*.

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- An interesting property of Mean:
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$, where $\forall i \ c_i \geq 0$
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq 2 \cdot \mu\}|}{|C|}$? ($\leq 1/2$)
 - What can you say about the quantity $\frac{|\{i \mid c_i \geq \alpha \cdot \mu\}|}{|C|}$ for $\alpha > 1$? ($\leq 1/\alpha$)
 - Suppose X denotes a random variable that takes values in C with uniform probability. What is $\Pr[X \geq \alpha \cdot \mu]$? ($\leq 1/\alpha$)
 - This is known as **Markov's inequality**.
 - One-sided tail.



https://angeloyeo.github.io/2022/09/12/Markov_Chebyshev_Inequality_en.html

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data *deviates* from the mean value.
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
 - Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
 - What can you say about the average value of the deviations (i.e., $\frac{\sum_i d_i}{n}$)?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data *deviates* from the mean value.
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
 - Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
 - What can you say about the average value of the deviations (i.e., $\frac{\sum_i d_i}{n}$)? (Ans:0)
 - The average deviation does not give any additional information. What could be a better measure?

Mean, normal, and the CLT

- Mean = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data *deviates* from the mean value.
 - Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
 - Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
 - What can you say about the average value of the deviations (i.e., $\frac{\sum_i d_i}{n}$)? (Ans:0)
 - The average deviation does not give any additional information. How about losing the signs?
 - Consider absolute values
 - **Consider the squares of the deviations**

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data *deviates* from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*
- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- Can you upper bound ($\alpha > 1$):
$$\frac{|\{i : d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq ???$$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*
- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- You can upper bound ($\alpha > 1$):
$$\frac{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2} \quad (\text{from Markov's})$$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- You can upper bound ($\alpha > 1$):
$$\frac{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2} \quad (\text{from Markov's})$$
- So,
$$\frac{|\{i: (c_i - \mu)^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2}$$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

- Let μ denote the mean of a collection $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = \left(\sum_i d_i^2\right)^{1/2}$.
- You can upper bound ($\alpha > 1$):
$$\frac{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2} \quad (\text{from Markov's})$$
- So, $\frac{|\{i: (c_i - \mu)^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2}$
- So, $\frac{|\{i: |c_i - \mu| \geq \alpha \cdot \sigma\}|}{n} \leq \frac{1}{\alpha^2}$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- You can upper bound ($\alpha > 1$):
$$\frac{n}{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|} \leq \frac{1}{\alpha^2} \quad (\text{from Markov's})$$
- So, $\frac{n}{|\{i: (c_i - \mu)^2 \geq \alpha^2 \cdot \sigma^2\}|} \leq \frac{1}{\alpha^2}$
- So, $\frac{n}{|\{i: |c_i - \mu| \geq \alpha \cdot \sigma\}|} \leq \frac{1}{\alpha^2}$
- So, if X denotes the random variable that takes values in C with uniform probability. What is
$$\Pr[|X - \mu| \leq \alpha \cdot \sigma] \geq ???$$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- You can upper bound ($\alpha > 1$):
$$\frac{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2} \quad (\text{from Markov's})$$
- So, $\frac{|\{i: (c_i - \mu)^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2}$
- So, $\frac{|\{i: |c_i - \mu| \geq \alpha \cdot \sigma\}|}{n} \leq \frac{1}{\alpha^2}$
- So, if X denotes the random variable that takes values in C with uniform probability. Then,

$$\Pr[|X - \mu| \leq \alpha \cdot \sigma] \geq 1 - \frac{1}{\alpha^2}$$

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - **Informal reason:** *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*

- Let μ denote the mean of a collection $C = \{c_1, c_2, \dots, c_n\}$.
- Suppose d_i denotes the deviation from the mean. So, $d_i = c_i - \mu$.
- Variance = σ^2 = mean of squared deviations = $\sum_i d_i^2 / n$
- SD = $\sigma = (\sum_i d_i^2)^{1/2}$.
- You can upper bound ($\alpha > 1$):
 $\frac{|\{i: d_i^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2}$ (from Markov's)
- So, $\frac{|\{i: (c_i - \mu)^2 \geq \alpha^2 \cdot \sigma^2\}|}{n} \leq \frac{1}{\alpha^2}$
- So, $\frac{|\{i: |c_i - \mu| \geq \alpha \cdot \sigma\}|}{n} \leq \frac{1}{\alpha^2}$
- So, if X denotes the random variable that takes values in C with uniform probability. Then,
$$\Pr[|X - \mu| \leq \alpha \cdot \sigma] \geq 1 - \frac{1}{\alpha^2}$$
- This is known as **Chebychev's inequality**.

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*
- We can use the above result to express any collection of numerical values ($c_1, \dots, c_n$) in terms of the number of standard deviations above/below average:
$$Z_i = \frac{c_i - \mu}{\sigma}$$
- This is known as conversion to **standard units**.

Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*
- We can use the above result to express any collection of numerical values ($c_1, \dots, c_n$) in terms of the number of standard deviations above/below average:
$$z_i = \frac{c_i - \mu}{\sigma}$$
- This is known as conversion to **standard units**.
- What can you say about the range of values in standard units?

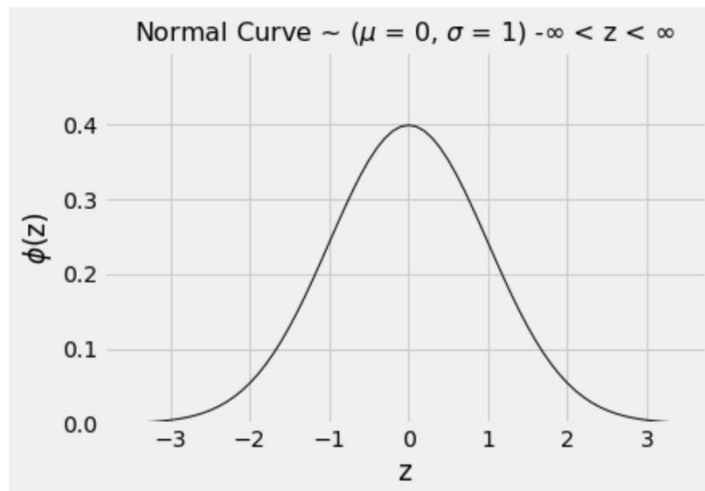
Mean, normal, and the CLT

- **Mean(μ)** = $\frac{\text{sum of numbers in the collection}}{\text{size of the collection}}$
- It is useful to know how much the data **deviates** from the mean value.
- **Variance(σ^2)** = mean of squared deviations
- **Standard deviation(σ)** = square root of variance
 - SD: root mean squared deviations from average
- Why does “standard deviation” occur so frequently in data analysis conversations?
 - Informal reason: *In all numerical data sets, the bulk of the entries are within the range “average $\pm$ a few SDs”.*
- We can use the above result to express any collection of numerical values $(c_1, \dots, c_n)$ in terms of the number of standard deviations above/below average:
$$Z_i = \frac{c_i - \mu}{\sigma}$$
- This is known as conversion to **standard units**.
- What can you say about the range of values in standard units?
 - *Mostly in the range $(-5, 5)$.*

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**.

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

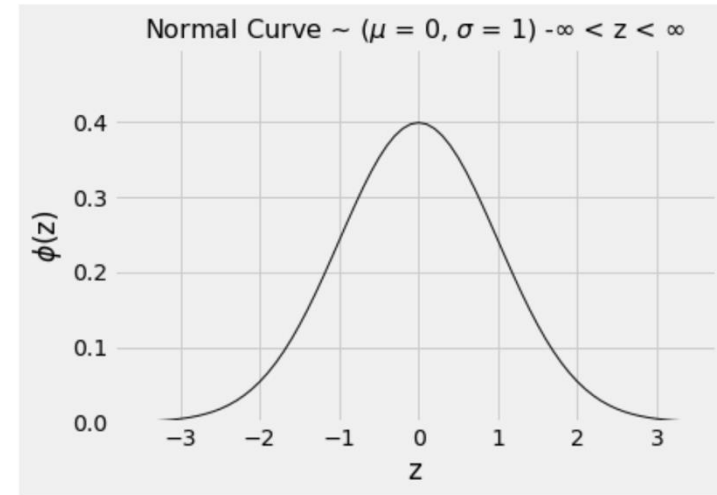


The values are standard units on the Horizontal axis of a standard normal curve.

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**.

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

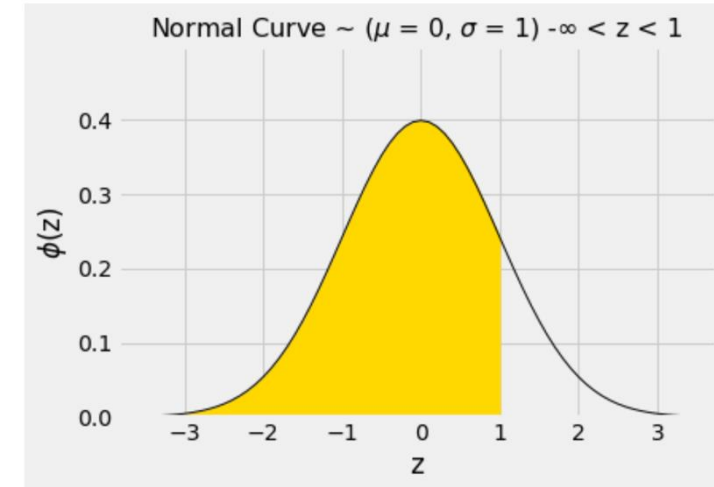


- Properties:
 - *The total area under the curve is 1. So, you can think of it as a histogram drawn to the density scale.*
 - *The curve is symmetric about 0. So, if a variable has this distribution, its mean and median are both 0.*
 - *The points of inflection of the curve are at -1 and +1.*
 - *If a variable has this distribution, its SD is 1. The normal curve is one of the very few distributions that has an SD so clearly identifiable on the histogram.*

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**.

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

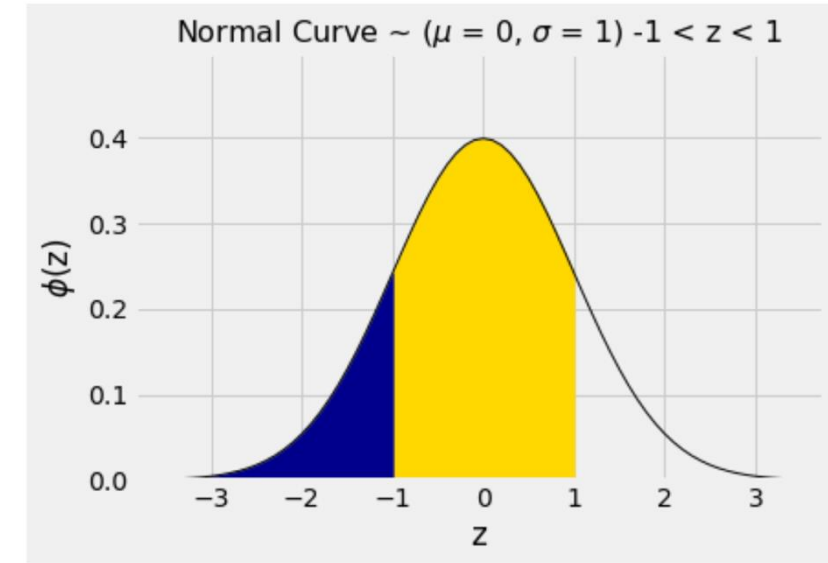


- The proportion of data can be calculated as the area under the curve.
- **Cumulative Distribution Function (cdf)**: The area under the curve to the left of the given data point.

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**.

End