Basic Statistical Concepts

Part 2.2 Descriptive Statistics, Random Variables,

Measures of Location / Dispersion

Descriptive Statistics

Q. What does the data look like using numbers?

Two ways to describe data:

- Data visualization / statistical graphics (Part 1)
- Summary measures

We may want to describe data with a few numbers.

- Q. What is the 'middle' height in the class?
  - Measures of Location: Mean, Median, Mode
- Q. How spread out are the heights in the class?
  - Measures of Dispersion / Spread: Variance, Standard Deviation, Range

Measures of Location

Q. Where is most of the data?

Mean: add all the values and divide the total by the number of points:

$$mean\ of\ x_1, x_2, \ldots, x_N: \qquad \bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Median: the value separating the higher half of a set of data values, from the lower half.

- If there are an odd number of values, choose the middle-ranked value
- If there are an even number of values, take the mean of the middle-ranked values

Mode: the value that appears most often.

Measures of Location: Example

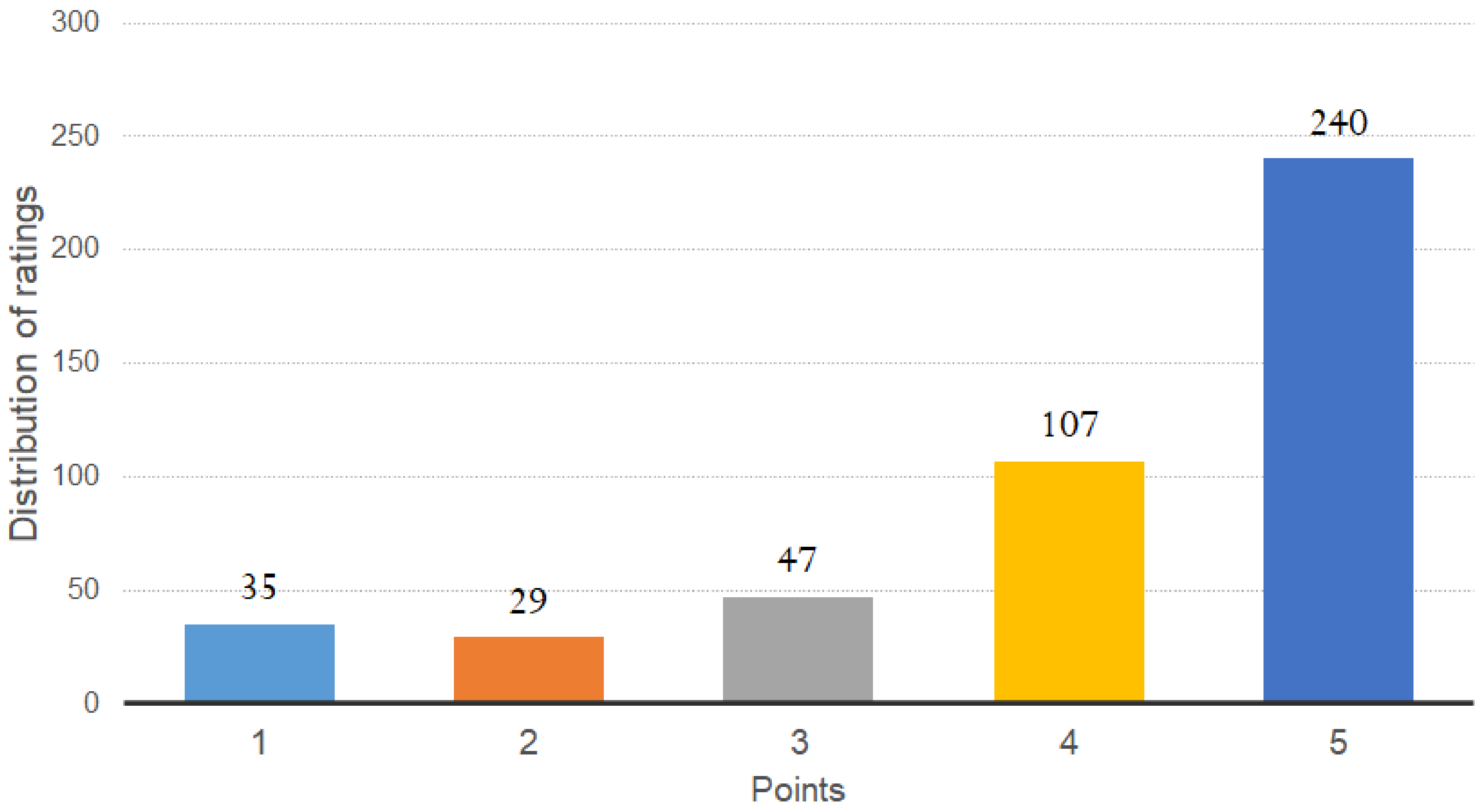
Q. Where is most of the data?

Data on customer ratings for Pittsburgh public transit in December 2017.

Mode: **?**

Median: **?**

Mean: **?**



Measures of Location: Example

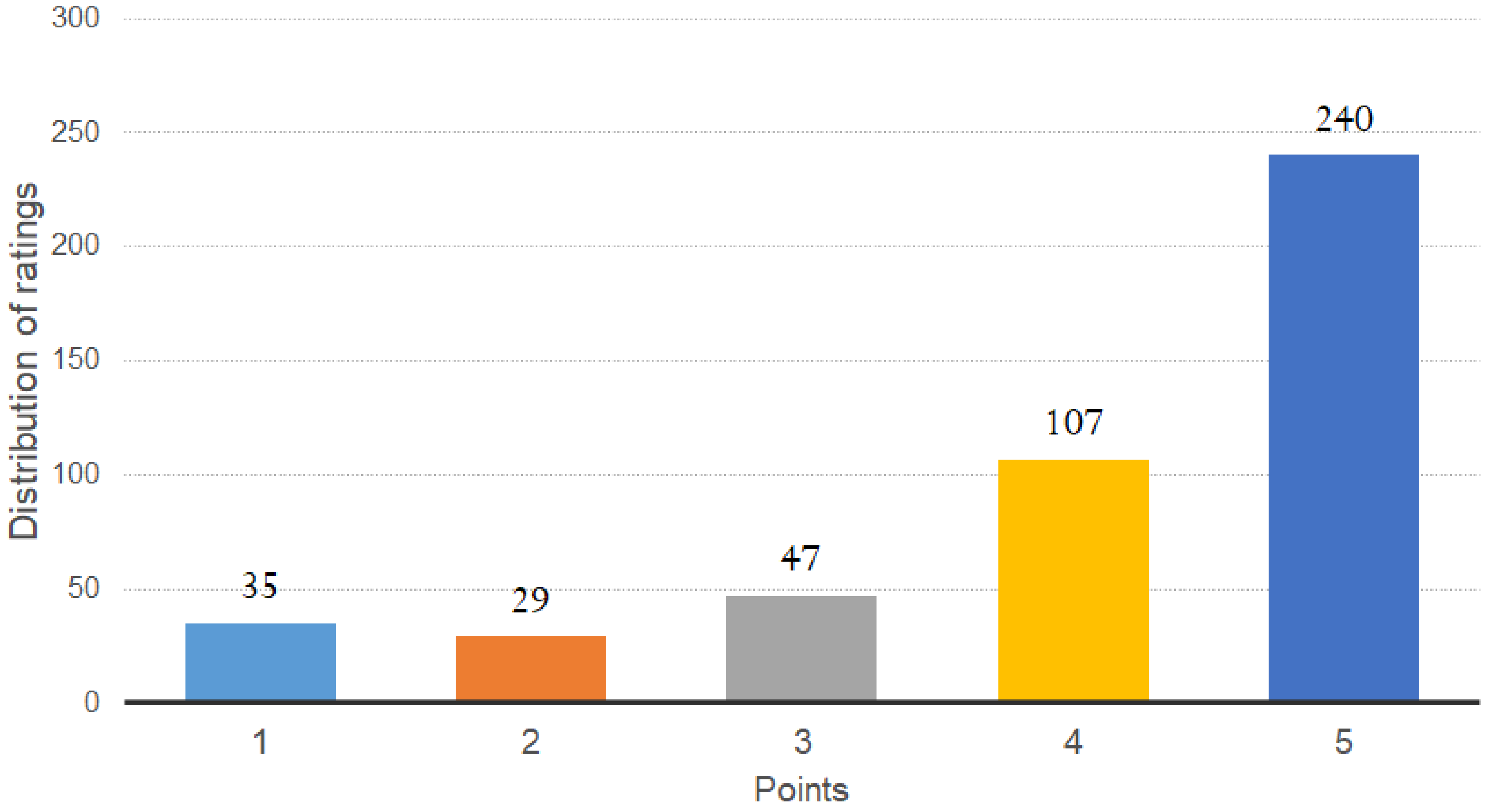
Q. Where is most of the data?

Data on customer ratings for Pittsburgh public transit in December 2017.

Mode: 5

Median: ?

Mean: ?



Measures of Location: Example

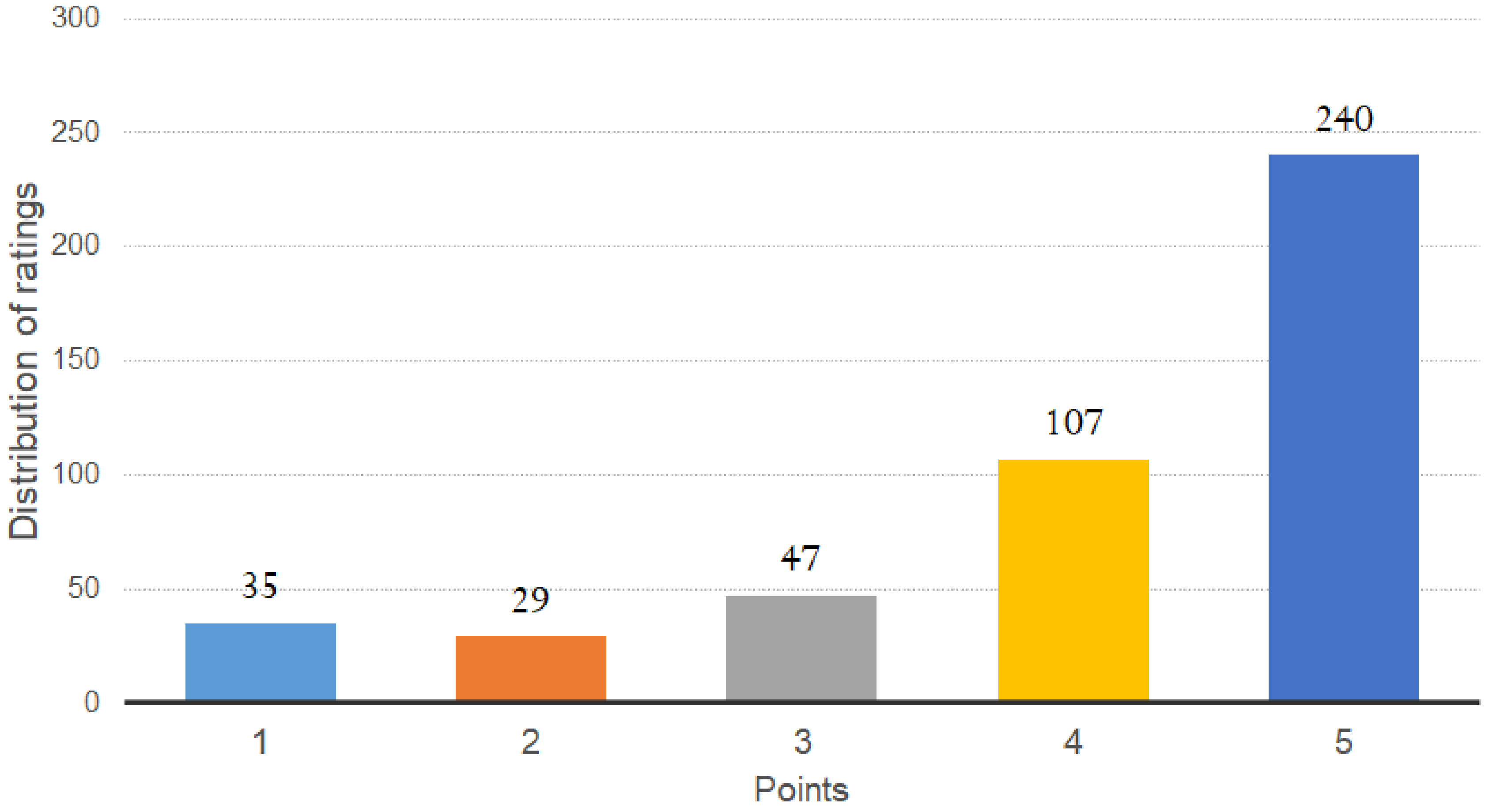
Q. Where is most of the data?

Data on customer ratings for Pittsburgh public transit in December 2017.

Mode: 5

Median: 5

Mean: ?



Measures of Location: Example

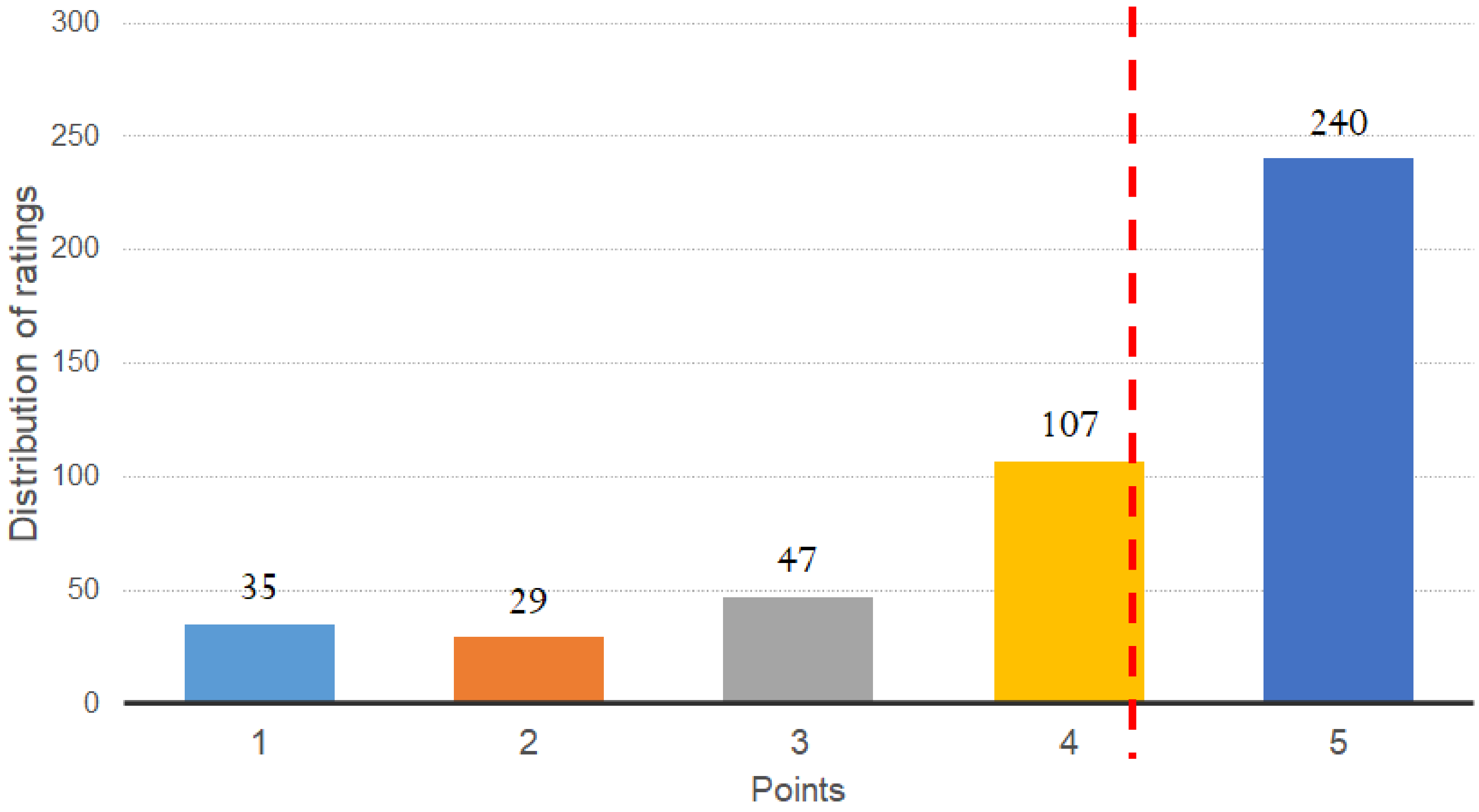
Q. Where is most of the data?

Data on customer ratings for Pittsburgh public transit in December 2017.

Mode: 5

Median: 5

Mean: ?



Measures of Location: Example

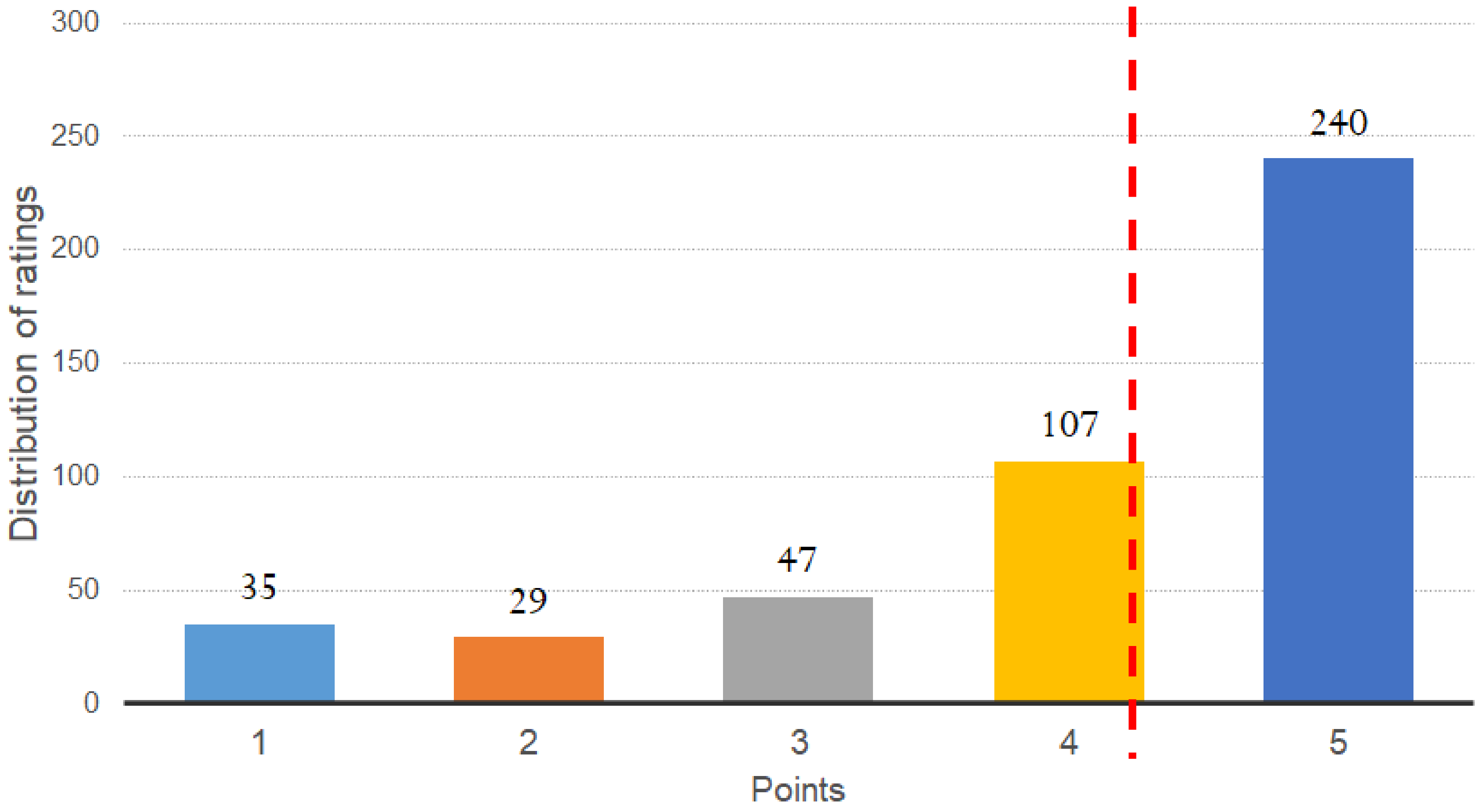
Q. Where is most of the data?

Data on customer ratings for Pittsburgh public transit in December 2017.

Mode: 5

Median: 5

Mean: 4.1



Measures of Location

Q. Where is most of the data?

Which one is better: Mean vs. Median vs. Mode

- Make a guess:
  - Mean number of twitter followers = ?
  - Median number of twitter followers = ?
- Depends on:
  - What we are trying to measure
  - Shape of the distributions of values
- Median is a better measure of central value when a small number of outliers could drastically skew the mean

Measures of Dispersion

Q. How spread out is the data?

What do we mean by dispersion?

- Deviation from the mean
- How common is each deviation

Measures of Dispersion

Q. How spread out is the data?

What do we mean by dispersion?

- Deviation from the mean
- How common is each deviation

Types of measures of dispersion:

- Range
- Variance
- Standard Deviation

Measures of Dispersion

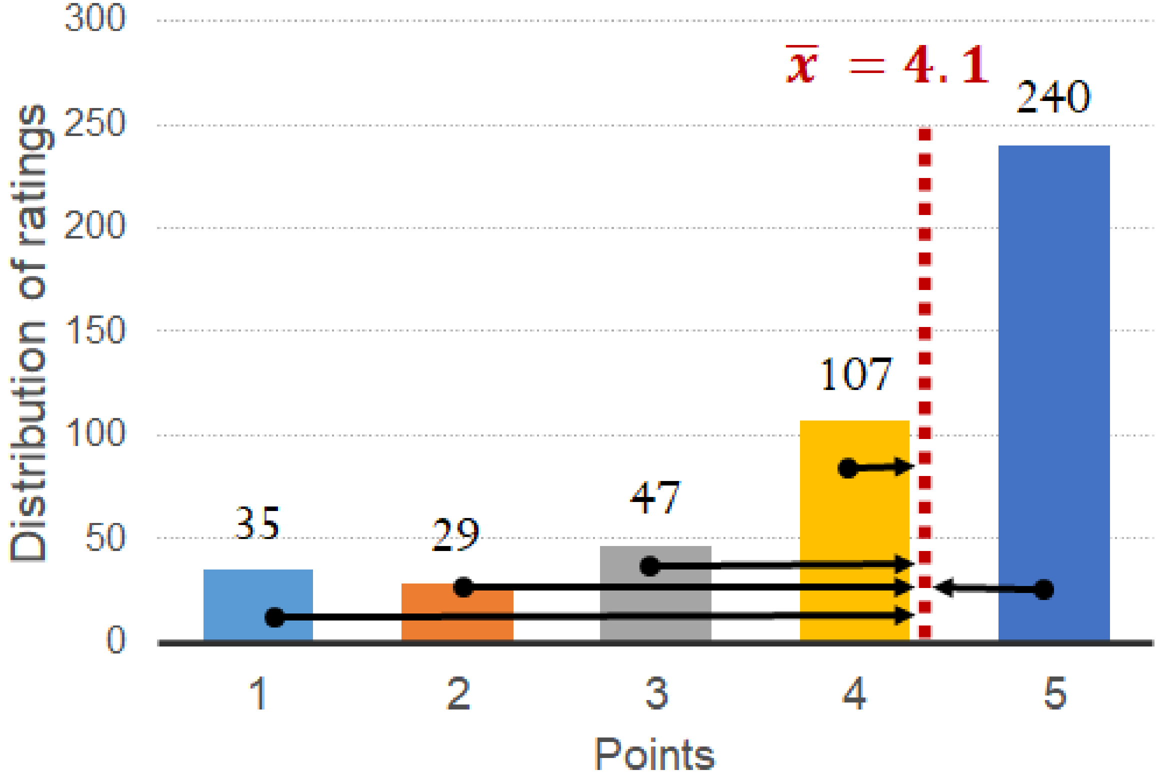
Q. How spread out is the data?

What do we mean by dispersion?

- Deviation from the mean
- How common is each deviation

Types of measures of dispersion:

- Range
- Variance
- Standard Deviation



Measures of Dispersion

Q. How spread out is the data?

Range: difference between the largest and smallest value in the data.

Variance: average squared difference from the mean

$$VAR(x_1, x_2, \dots, x_N) = \sigma^2$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N}$$

Variance is nice but the units are squared: Unit of variance = $(\text{original unit})^2$

Eg. $2(\text{ft})^2$

Measures of Dispersion

Q. How spread out is the data?

Standard Deviation: square root of variance.

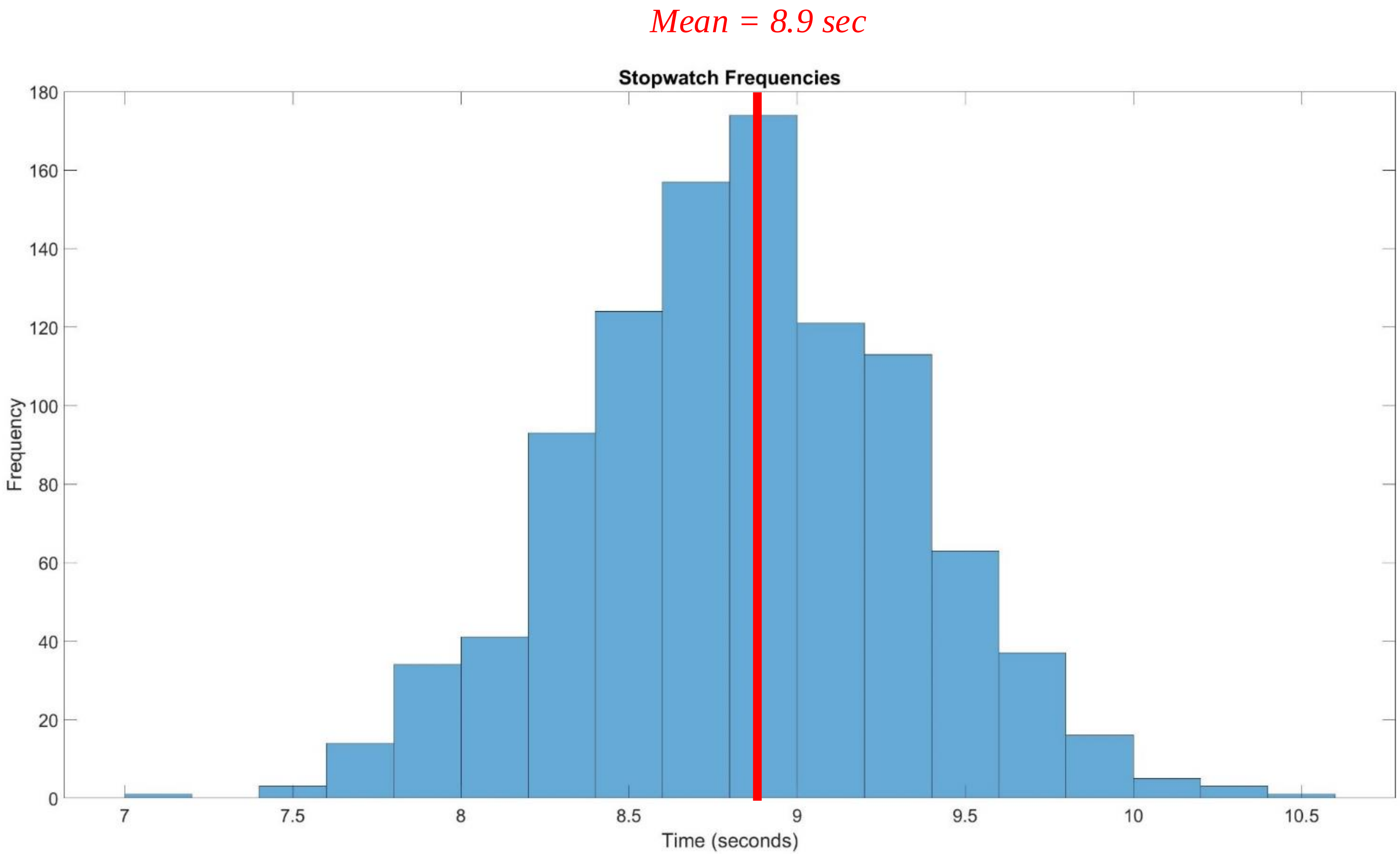
$$StDev(x_1, x_2, \dots, x_N) = \sigma$$

$$= \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}}$$

Standard deviation is nicer (in some ways) because it's in the original units.

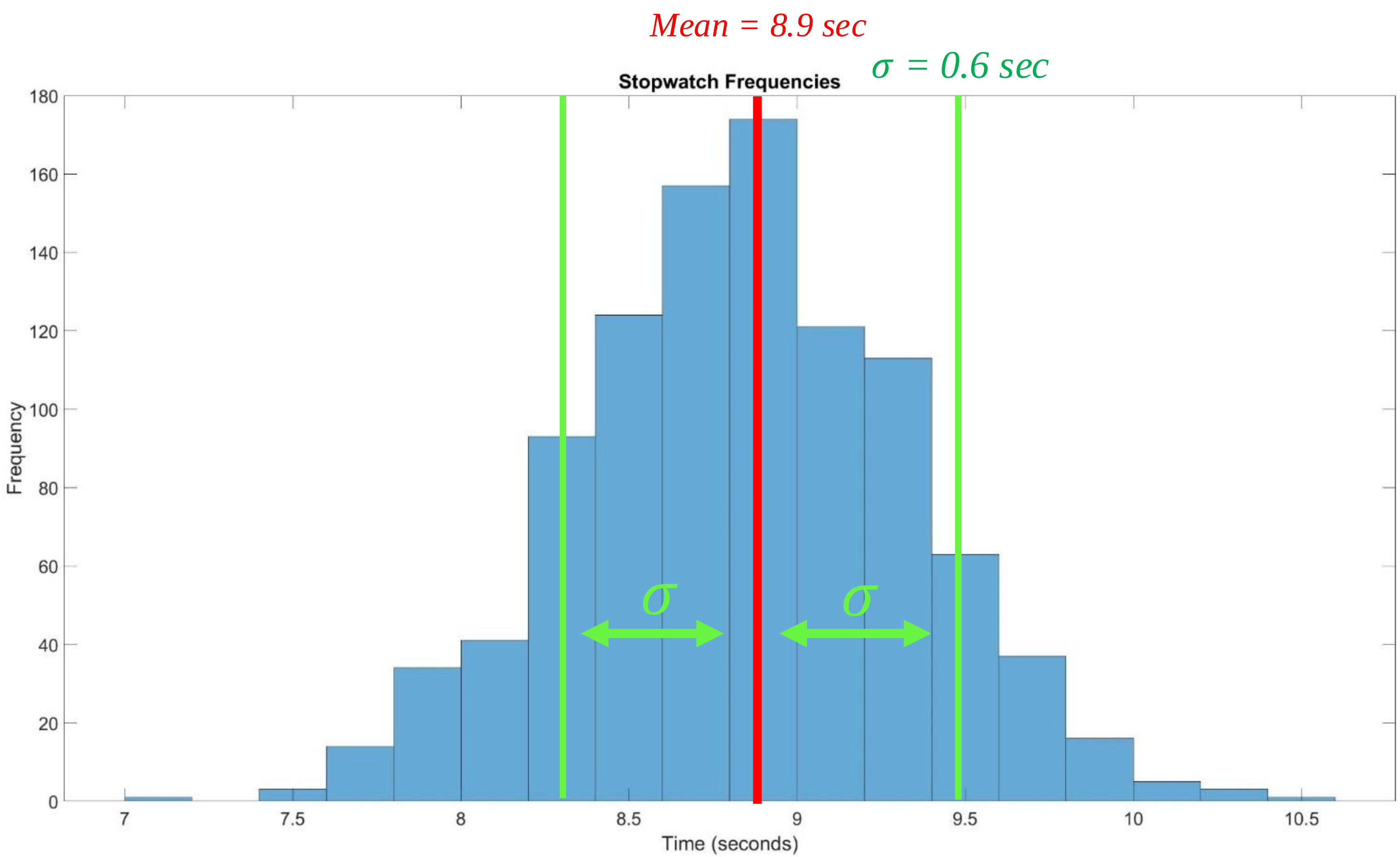
Summary Statistics: Example

Q. Summarize the ball drop times from the Empire State Building.



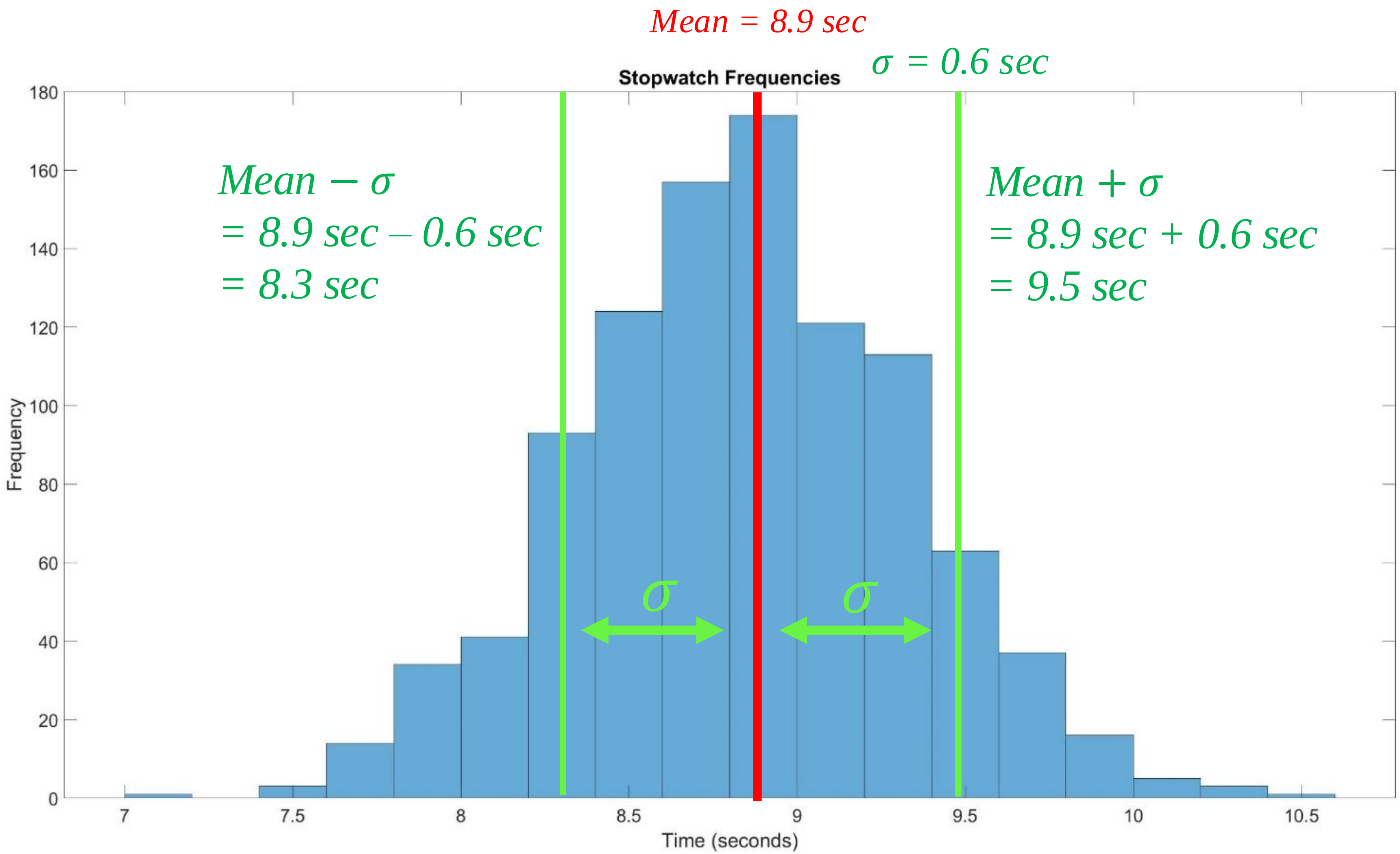
Summary Statistics: Example

Q. Summarize the ball drop times from the Empire State Building.



Summary Statistics: Example

Q. Summarize the ball drop times from the Empire State Building.



Random Variables

Q. Is there a way to summarize random variables like we did with data?

- We treat data as a realization of a random variable.
- A random variable is a model of the data.
- We can also describe the random variable.
    - Measures of Location
    - Measures of Dispersion / Spread
- Summary statistics describe the data, which are realizations, not the underlying random variable itself.
    - Now we use information about the likelihood of each outcome of X
    - This information is contained in the probability distribution of X

Discrete Random Variables

A discrete random variable takes a *finite* number of values

- Can list *all* the possible values
- Number of the die, number of customers, etc.

The probability mass function of a discrete random variable lists the probabilities associated with each of its possible values

- Can list all possible probabilities: $P(X = x)$ for each value x
- The probabilities must be positive and less than one: $0 \leq P(X = x) \leq 1$
- The probabilities must sum to one:

$$\sum_x P(X = x) = 1$$

Discrete Random Variables: Example

Q. Let X take 5 different values.

Values of X	Probability of Occurrence
$X = 1$	$P(X = 1) = 0.12$
$X = 2$	$P(X = 2) = 0.4$
$X = 3$	$P(X = 3) = 0.35$
$X = 4$	$P(X = 4) = 0.03$
$X = 5$	$P(X = 5) = 0.1$

Discrete Random Variables: Example

Q. Let X take 5 different values.

Values of X	Probability of Occurrence
$X = 1$	$P(X = 1) = 0.12$
$X = 2$	$P(X = 2) = 0.4$
$X = 3$	$P(X = 3) = 0.35$
$X = 4$	$P(X = 4) = 0.03$
$X = 5$	$P(X = 5) = 0.1$

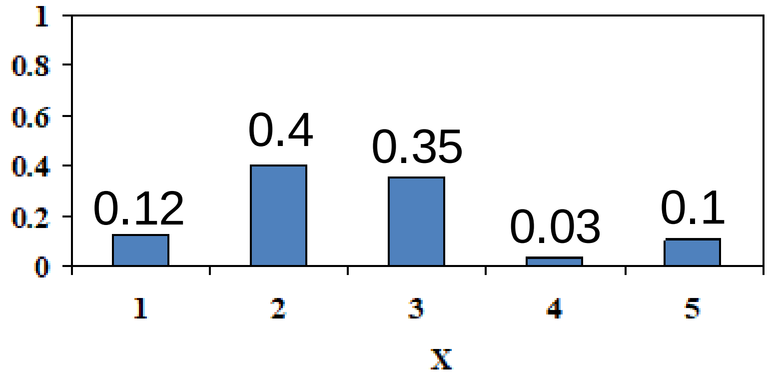
Sum = 1

Discrete Random Variables: Example

Probability Distribution:

- Probability that value of *X* is *equal to* *x*

$$\mathrm{P}(X = x)$$

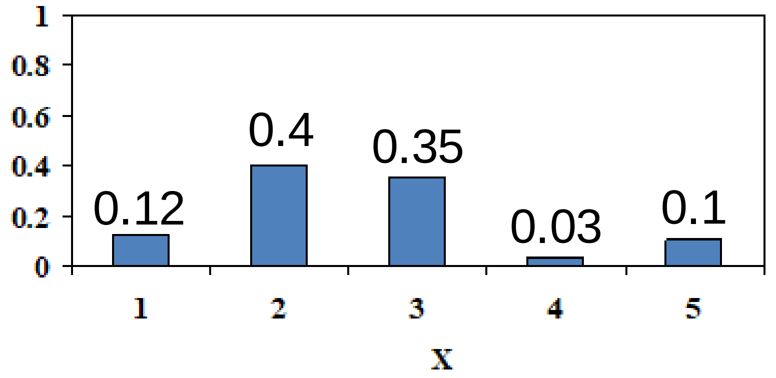


Discrete Random Variables: Example

Probability Distribution:

- Probability that value of *X* is *equal to* *x*

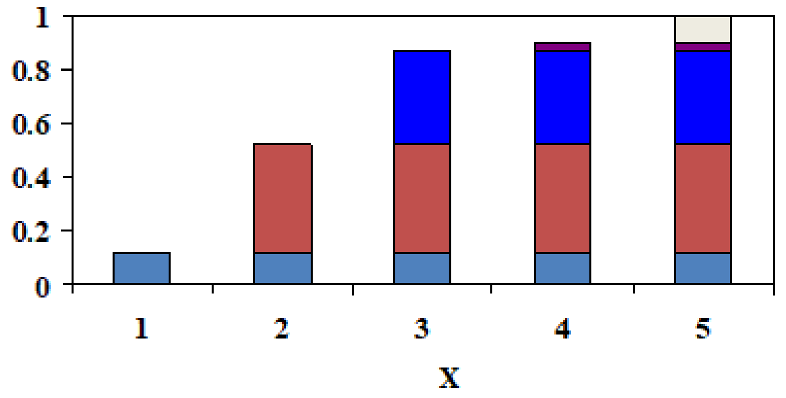
$$\mathrm{P}(X = x)$$



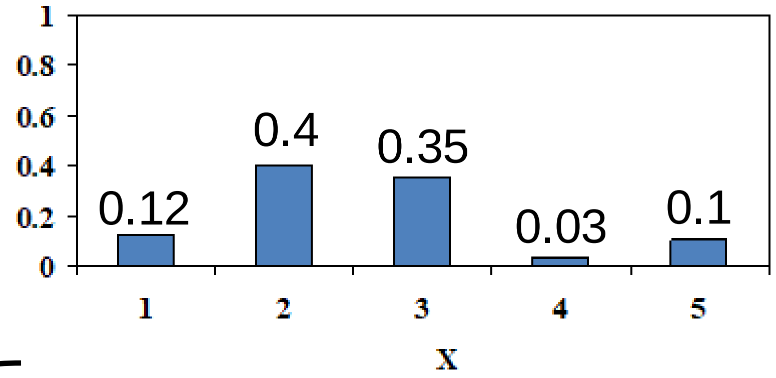
Cumulative Probability Distribution

- Probability that value of *X* is smaller than or equal to *x*

$$\mathrm{P}(X \leq x)$$



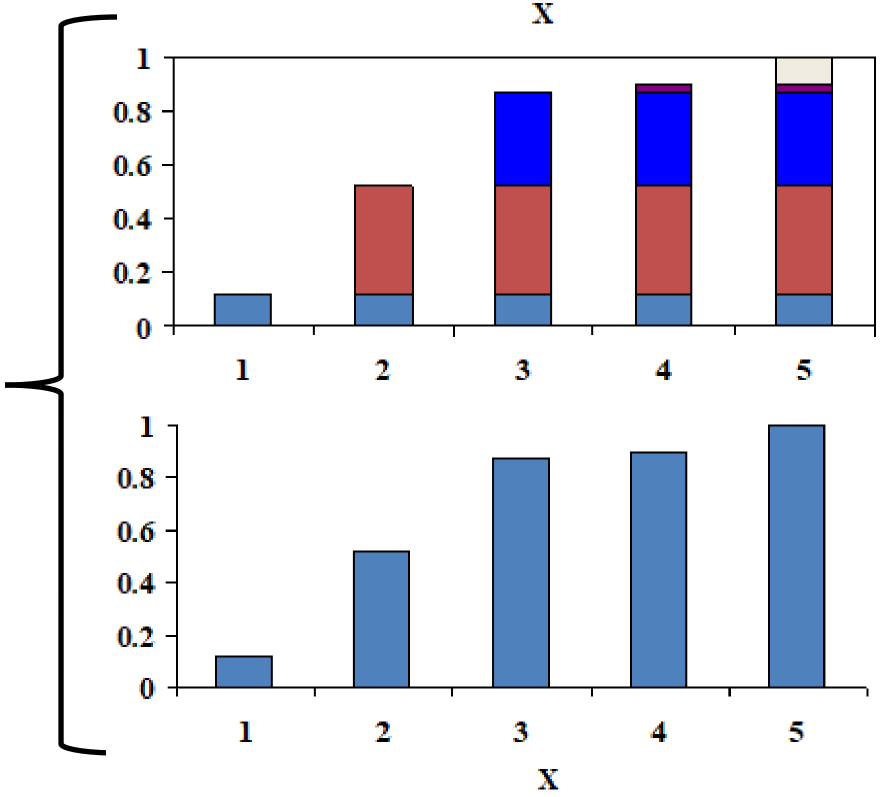
Discrete Random Variables: Example



Probability Distribution:

- Probability that value of X is *equal to* x

$$P(X = x)$$



Cumulative Probability Distribution

- Probability that value of X is smaller than or equal to x

$$P(X \leq x)$$

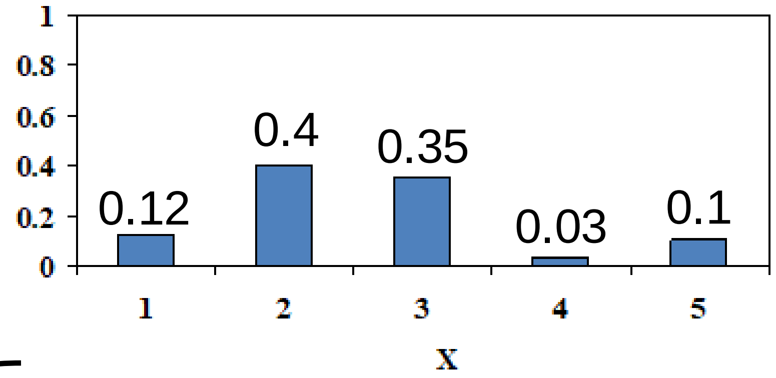
$$P(X \leq 1) = P(X = 1)$$

Discrete Random Variables: Example

Probability Distribution:

- Probability that value of *X* is *equal to x*

$$P(X = x)$$



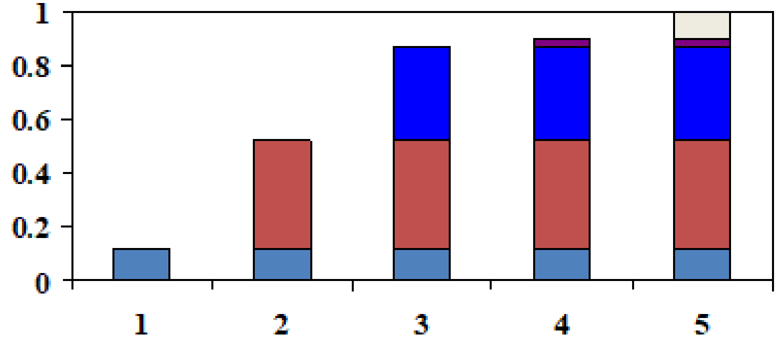
Cumulative Probability Distribution

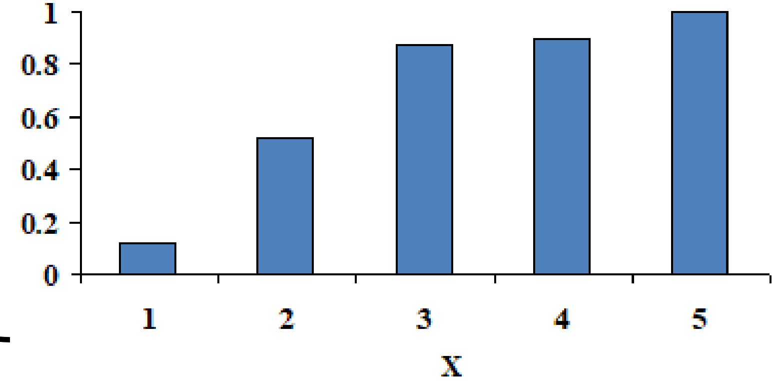
- Probability that value of *X* is smaller than or equal to *x*

$$P(X \leq x)$$

$$P(X \leq 1) = P(X = 1)$$

$$P(X \leq 2) = P(X = 1) + P(X = 2)$$



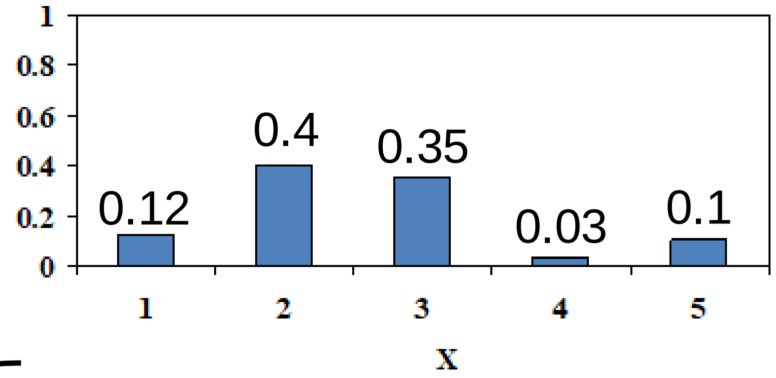


Discrete Random Variables: Example

Probability Distribution:

- Probability that value of X is *equal to* x

$$P(X = x)$$



Cumulative Probability Distribution

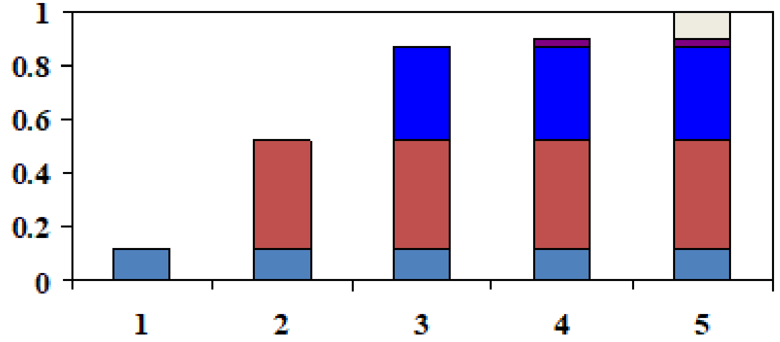
- Probability that value of X is smaller than or equal to x

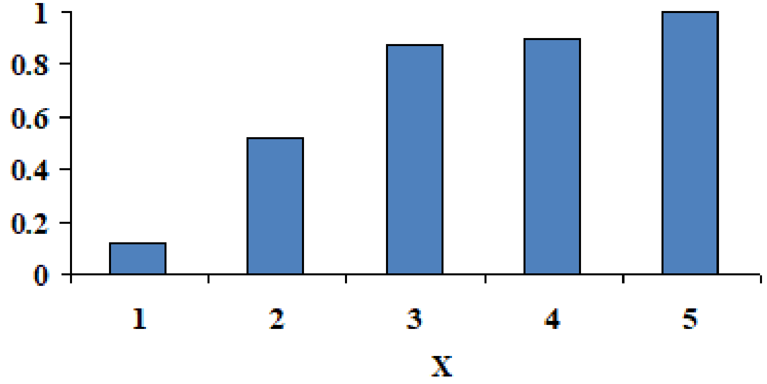
$$P(X \leq x)$$

$$P(X \leq 1) = P(X = 1)$$

$$P(X \leq 2) = P(X = 1) + P(X = 2)$$

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$$





Discrete Random Variables: Expected Value

A random variable X can take a number of different values: $x_1, x_2, \ldots$

with corresponding probabilities: $P(X = x_1),\ \ P(X = x_2),\ \ \ldots$

- Expected Value of a random variable X: $\left(\text{denoted as } \mu_X \text{ or } E(X)\right)$

$$\mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

Discrete Random Variables: Expected Value

A random variable X can take a number of different values: $x_1, x_2, \ldots$

with corresponding probabilities: $P(X = x_1),\ \ P(X = x_2),\ \ \ldots$

- Expected Value of a random variable X: $\big($denoted as μ_X or $E(X)\big)$

$$\mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

- Mean of a series of values $\{x_1, x_2, \ldots x_N\}$ is a very similar idea:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = x_1 \cdot \frac{1}{N} + x_2 \cdot \frac{1}{N} + \ldots + x_N \cdot \frac{1}{N}$$

Discrete Random Variables: Expected Value

A random variable X can take a number of different values: $x_1, x_2, \ldots$

with corresponding probabilities: $P(X = x_1),\ P(X = x_2),\ \ldots$

- Expected Value of a random variable X: $\left(\text{denoted as } \mu_X \text{ or } E(X)\right)$

$$\mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

- μ_X is a measure of the "center" of the distribution
- Weighted average of the outcomes, weighted by the probabilities

Discrete Random Variables: Variance

Variance of random variable X: (denoted as σ_X^2 or $Var(X)$)

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2) + \cdots$$

Find the value's squared deviation from the "mean" (expected value μ_X).

Then take the weighted average.

Discrete Random Variables: Variance

Variance of random variable X: (denoted as σ_X^2 or $Var(X)$)

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2) + \cdots$$

Find the value's squared deviation from the "mean" (expected value μ_X).

Then take the weighted average.

Variance of a series of numbers $\{x_1, x_2, \ldots x_N\}$

$$\sigma^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}$$

Each number's squared deviation from the mean $\overline{x}$

Then take average.

Discrete Random Variables: Standard Deviation

Standard Deviation of random variable X: (denoted as σ_X)

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)}$$

… a measure of the dispersion of the distribution in original units.

Discrete Random Variables: Summary

- Discrete Random Variable: numerical valued outcomes, can list all possible values it may take
- Mean (*expected value*) of X:

$$E(X) = \mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

Discrete Random Variables: Summary

- Discrete Random Variable: numerical valued outcomes, can list all possible values it may take
- Mean (*expected value*) of X:

$$E(X) = \mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

- Variance of X:

$$Var(X) = \sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2) + \cdots$$

Discrete Random Variables: Summary

- Discrete Random Variable: numerical valued outcomes, can list all possible values it may take
- Mean (*expected value*) of X:

$$E(X) = \mu_X = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots$$

- Variance of X:

$$Var(X) = \sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2) + \cdots$$

- Standard deviation of X:

$$Std.Dev.(X) = \sigma_X = \sqrt{Var(X)} = \sqrt{\sigma_X^2}$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get $10 million.*
> - *If it is tails, you'll have to pay me $1 million"*

What are *expected value*, *variance*, and *standard deviation* of the change in your wealth after this coin toss?

- X = change in your wealth (in millions of dollars)

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

What are *expected value*, *variance*, and *standard deviation* of the change in your wealth after this coin toss?

- X = change in your wealth (in millions of dollars)
- Expected value of X:

$$\mu_X = \underline{x_1} \cdot P(X = \underline{x_1}) + \underline{x_2} \cdot P(X = \underline{x_2})$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

What are *expected value*, *variance*, and *standard deviation* of the change in your wealth after this coin toss?

- X = change in your wealth (in millions of dollars)
- Expected value of X:

$$\mu_X = \underline{x_1} \cdot P(X = \underline{x_1}) + \underline{x_2} \cdot P(X = \underline{x_2})$$

$$= \underline{(10)} \cdot P(X = \underline{10}) + \underline{(-1)} \cdot P(X = \underline{-1})$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

What are *expected value*, *variance*, and *standard deviation* of the change in your wealth after this coin toss?

- X = change in your wealth (in millions of dollars)
- Expected value of X:

$$\mu_X = \underline{x_1} \cdot P(X = \underline{x_1}) + \underline{x_2} \cdot P(X = \underline{x_2})$$

$$= \underline{(10)} \cdot P(X = \underline{10}) + \underline{(-1)} \cdot P(X = \underline{-1})$$

$$= (10) \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} \quad = 4.5 \text{ (\$ million)}$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

Variance of X:

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2)$$

Example: A Bet With A Rich Person

"We'll toss a coin once:

- *If it is heads, you get \$10 million.*
- *If it is tails, you'll have to pay me \$1 million"*

Variance of X:

$$\sigma_X^2 = (\underline{x_1} - \mu_X)^2 \cdot P(X = \underline{x_1}) + (\underline{x_2} - \mu_X)^2 \cdot P(X = \underline{x_2})$$

$$= (\underline{10} - \underline{\mu_X})^2 \cdot P(X = \underline{10}) + (\underline{-1} - \underline{\mu_X})^2 \cdot P(X = \underline{-1})$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

Variance of X:

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2)$$

$$= (10 - \mu_X)^2 \cdot P(X = 10) + (-1 - \mu_X)^2 \cdot P(X = -1)$$

$$= (10 - 4.5)^2 \cdot P(X = 10) + (-1 - 4.5)^2 \cdot P(X = -1)$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
>
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

Variance of X:

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2)$$

$$= (10 - \mu_X)^2 \cdot P(X = 10) + (-1 - \mu_X)^2 \cdot P(X = -1)$$

$$= (10 - 4.5)^2 \cdot P(X = 10) + (-1 - 4.5)^2 \cdot P(X = -1)$$

$$= (10 - 4.5)^2 \cdot \frac{1}{2} + (-1 - 4.5)^2 \cdot \frac{1}{2}$$

Example: A Bet With A Rich Person

> *"We'll toss a coin once:*
> - *If it is heads, you get \$10 million.*
> - *If it is tails, you'll have to pay me \$1 million"*

Variance of X:

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot P(X = x_1) + (x_2 - \mu_X)^2 \cdot P(X = x_2)$$

$$= (10 - \mu_X)^2 \cdot P(X = 10) + (-1 - \mu_X)^2 \cdot P(X = -1)$$

$$= (10 - 4.5)^2 \cdot P(X = 10) + (-1 - 4.5)^2 \cdot P(X = -1)$$

$$= (10 - 4.5)^2 \cdot \frac{1}{2} + (-1 - 4.5)^2 \cdot \frac{1}{2}$$

$$= (5.5)^2 \cdot 0.5 + (-5.5)^2 \cdot 0.5 \qquad = 30.25 \text{ (\$ million}^2\text{)}$$

Example: A Bet With A Rich Person

"We'll toss a coin once:

- *If it is heads, you get $10 million.*
- *If it is tails, you'll have to pay me $1 million"*

Standard Deviation of X:

$$\sigma_X = \sqrt{\sigma_X^2}$$

$$= \sqrt{30.25}$$

$$= 5.5 \text{ (\$ million)}$$

Example: Sales Calls

- A salesperson for a national clothing company makes five calls to potential customers every day.
- The following probability distribution describes the number of successful calls each day:

Number of Successful Calls	Probability
0	0.15
1	0.40
2	0.20
3	0.10
4	0.10
5	0.05

- How many successful calls does this salesperson expect to make each day?

Continuous Random Variables

Q. What if the data is continuous?

A continuous random variable takes an *infinite* number of values

- Cannot list *all* the possible values
- Weight, height, time, etc.

Continuous Random Variables

Q. What if the data is continuous?

A continuous random variable takes an *infinite* number of values

- Cannot list *all* the possible values
- Weight, height, time, etc.

The probability density function *f(x)* describes the distribution of a continuous random variable

- For any number a, the area under the curve of *f(x)* between the negative infinity to a gives the probability that $X \leq a$:

$$\int_{-\infty}^{a} f(x)dx = P(X \leq a)$$

Continuous Random Variables

Q. What if the data is continuous?

A continuous random variable takes an *infinite* number of values

- Cannot list *all* the possible values
- Weight, height, time, etc.

The probability density function *f(x)* describes the distribution of a continuous random variable

- For any number a, the area under the curve of *f(x)* between the negative infinity to a gives the probability that $X \leq a$: $$\int_{-\infty}^{a} f(x)dx = P(X \leq a)$$
- The density at any point x must not be less than zero: $f(x) \geq 0$

Continuous Random Variables

Q. What if the data is continuous?

A continuous random variable takes an *infinite* number of values

- Cannot list *all* the possible values
- Weight, height, time, etc.

The probability density function *f(x)* describes the distribution of a continuous random variable

- For any number a, the area under the curve of *f(x)* between the negative infinity to a gives the probability that $X \leq a$: $$\int_{-\infty}^{a} f(x)dx = P(X \leq a)$$
- The density at any point x must not be less than zero: $f(x) \geq 0$
- The area under the curve of *f(x)* must sum to one:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Continuous Random Variables

Some notation

- X denotes the random variable
- x denotes a particular value / realization of X

Continuous Random Variables

Some notation

- X denotes the random variable
- x denotes a particular value / realization of X
  - E.g.: The value of crop (X) can be \$100M or \$500M (x)

Continuous Random Variables

Some notation

- X denotes the random variable
- x denotes a particular value / realization of X
  - E.g.: The value of crop (X) can be $100M or $500M ($x$)
- Given some x
  - $P(X = x)$: probability that $X = x$

Continuous Random Variables

Some notation

- X denotes the random variable
- x denotes a particular value / realization of X
  - E.g.: The value of crop (X) can be \$100M or \$500M (x)
- Given some x
  - $P(X = x)$: probability that $X = x$
  - E.g.: The crop's value equals \$100M with probability 1/3
    and equals \$500M with probability 2/3

Continuous Random Variables

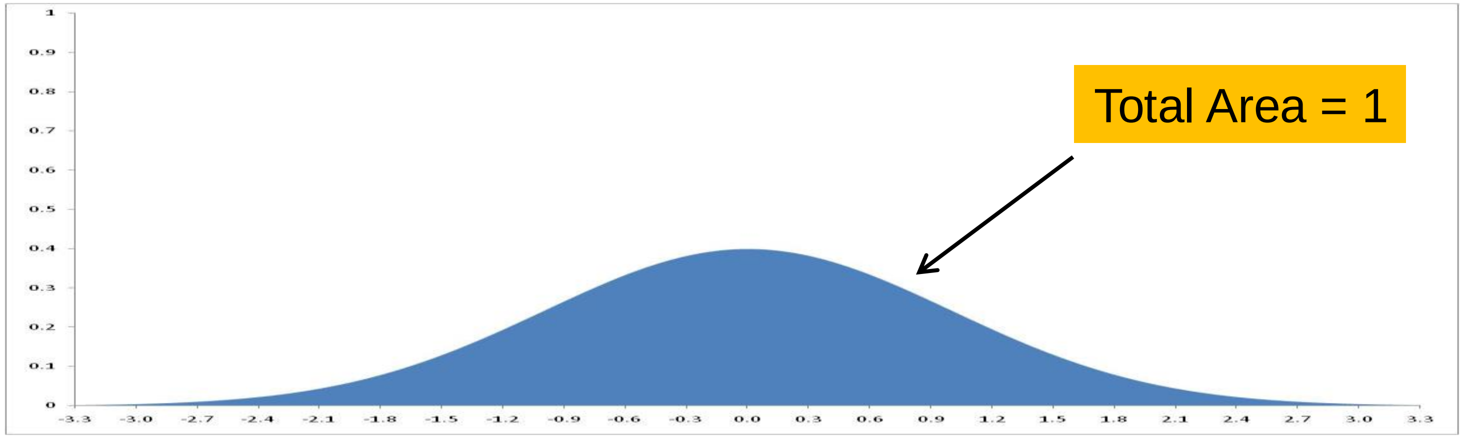
Some notation

- X denotes the random variable
- x denotes a particular value / realization of X
  - E.g.: The value of crop (X) can be \$100M or \$500M (x)
- Given some x
  - $P(X = x)$: probability that $X = x$
  - E.g.: The crop's value equals \$100M with probability 1/3
    and equals \$500M with probability 2/3

$$\begin{cases} P(\mathrm{X} = 100M) = 1/3 \\ P(\mathrm{X} = 500M) = 2/3 \end{cases}$$

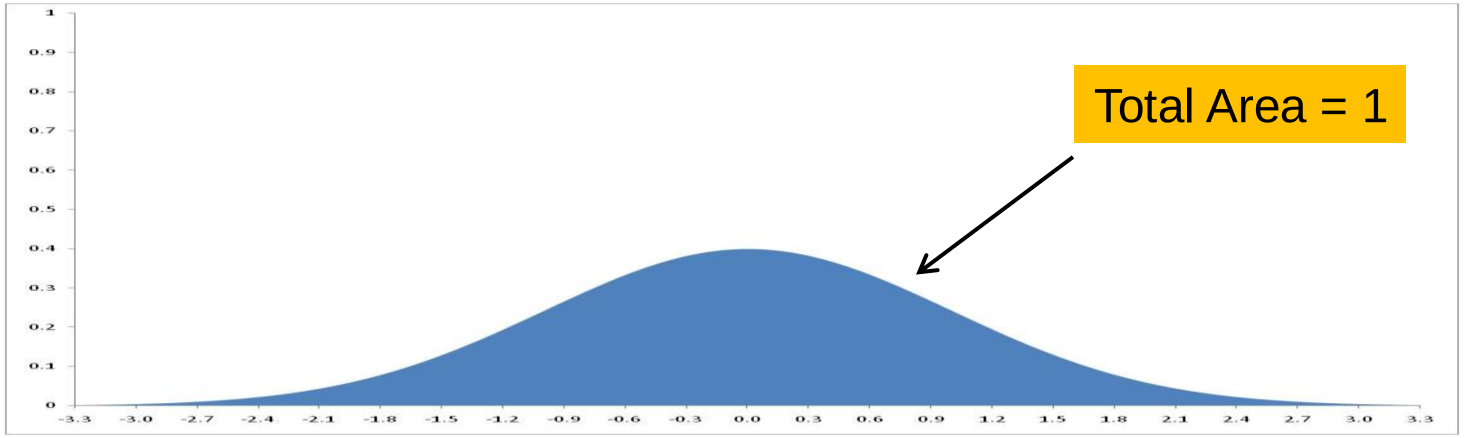
Continuous Random Variables

Probability Density Function of X : $f(x)$

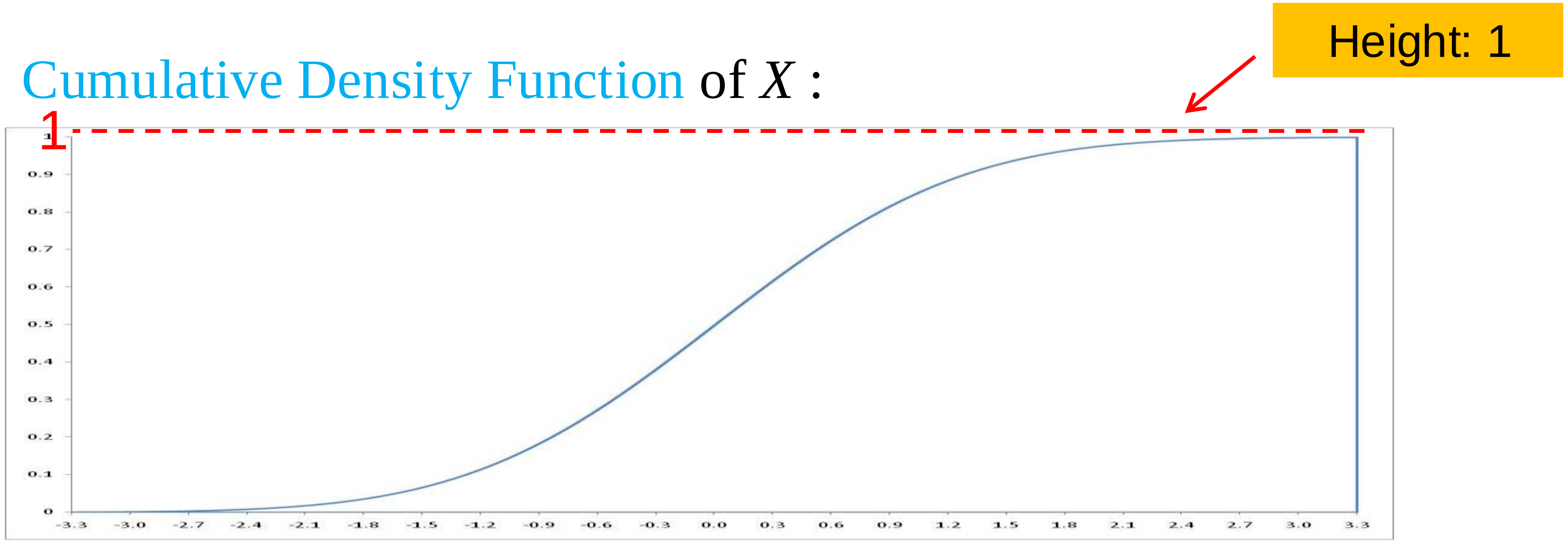


Continuous Random Variables

Probability Density Function of X : $f(x)$

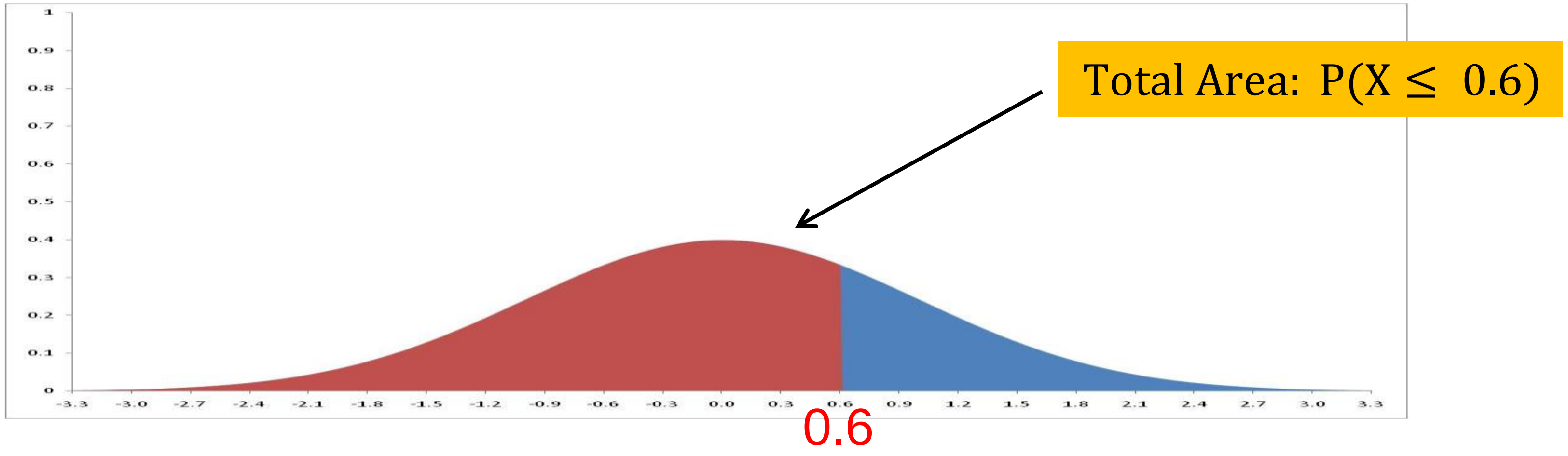


Cumulative Density Function of X :

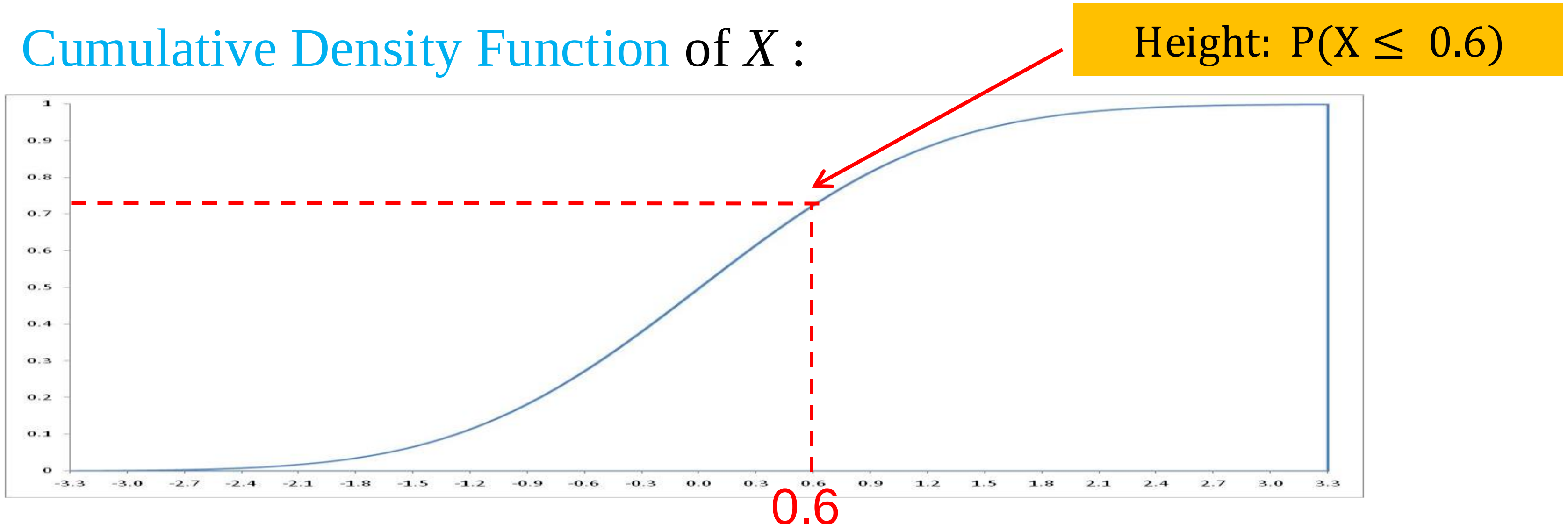


Continuous Random Variables

Probability Density Function of X : $f(x)$



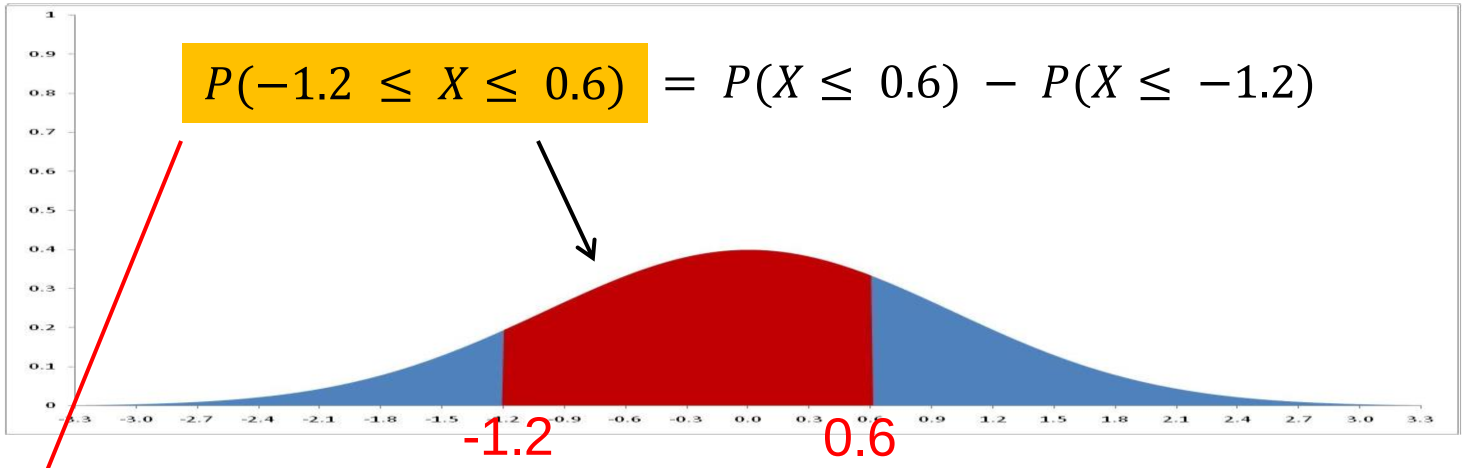
Cumulative Density Function of X :



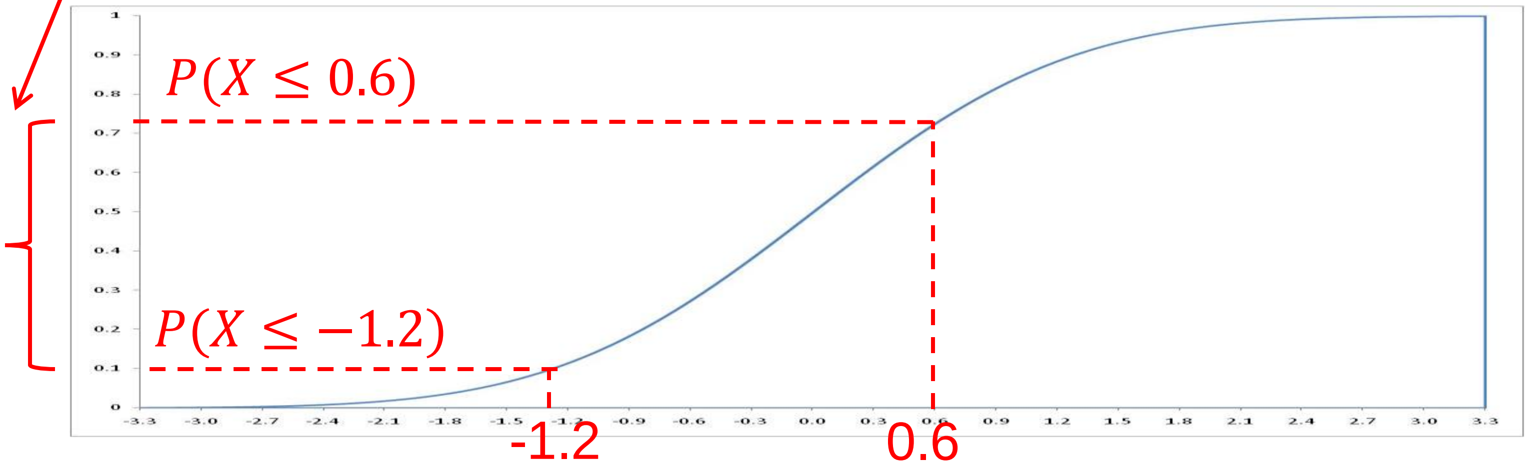
Continuous Random Variables

Probability Density Function of X : $f(x)$

$$P(-1.2 \leq X \leq 0.6) = P(X \leq 0.6) - P(X \leq -1.2)$$



Cumulative Density Function of X :



Continuous Random Variables

All the concepts and rules for discrete random variables apply to continuous random variables.

- Expected value: μ_X → measures the "center" of the distribution

- Variance: σ_X^2

- Standard deviation: σ_X

measure the dispersion of the distribution

Continuous Random Variable: numerical valued outcomes, can not list all possible values it may take

Linear Combinations of R.V.

- If random variable Y is a linear function of random variable X,
  - If $Y = a \cdot X$, then the expected value of Y:

$$E(Y) = a \cdot E(X)$$

  - If $Y = a \cdot X + b$, then the expected value of Y:

$$E(Y) = a \cdot E(X) + b$$

Linear Combinations of R.V.

- If random variable Y is a linear function of random variable X,
  - If $Y = a \cdot X + b \cdot Z$, then the expected value of Y:

$$E(Y) = a \cdot E(X) + b \cdot E(Z)$$

  - If $Y = a \cdot X + b \cdot Z + c$, then the expected value of Y:

$$E(Y) = a \cdot E(X) + b \cdot E(Z) + c$$