ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 3.1 | Location, Dispersion, Random Variables

Statistics

Why not just use visuals?

> *A picture is worth a thousand summary stats.*

> *But sometimes we want something more precise and concrete.*

Q. What is the ‘middle’ age in the class?

- *Measures of Location: **Mean, Median, Mode***

Q. How spread out are the ages in the class?

- *Measures of Dispersion / Spread: **Variance, Standard Deviation, Range***

Measure of Central Tendency (Location)

What is the "center" of the data?

Mode: *the value that appears most often*

Median: *the value separating the higher and lower halves*

- *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 two*

Mean: the center of mass

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

Central Tendency (Location): Class Age Example

What the center age in the class?

> see notebook

Central Tendency (Location): Tennis Example

Where should you stand on the court?

> see notebook

Measure of Dispersion

How spread out is the data?

Measure of Dispersion: Tennis Example

How far do you have to run?

> see notebook

Measure of Dispersion

How spread out is the data?

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

- *Simple but doesn't respond to changes near the middle of the distribution*

Mean Deviation: difference between each value and the average

$$\sum \frac{x_i - \bar{x}}{n} = \frac{X - \bar{X}}{n}$$

- *Simple but the average of the difference is zero…*

Mean Absolute Deviation: absolute value of the difference from the average

$$\sum \frac{|x_i - \bar{x}|}{n} = \frac{|X - \bar{X}|}{n}$$

- *The mean isn’t zero*
- *A little more complex and isn’t so nice mathematically*

Measure of Dispersion

How spread out is the data?

Variance: average squared difference from the mean

$$Var(X) = \sum \frac{(x_i - \bar{x})^2}{n} = \frac{X - \bar{X}}{n}$$

- *Treats negatives appropriately*
- *The mean isn't zero*
- *Mathematically nice*
- *Units are uninformative*

Measure of Dispersion

How spread out is the data?

Standard Deviation: a sort of average deviation from the mean

$$\sigma_X = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n}} = \sqrt{\frac{X - \bar{X}}{n}}$$

- *Treats negatives appropriately*
- *The mean isn't zero*
- *Mathematically nice*
- *Units are roughly average deviation from the mean*

Random Variables

What is data? … some fancy definitions….

Random Variable: *a function that assigns a number to each possible outcome of a random process (discrete or continuous)*

> *the random variable is like a deck with any collection of cards*

Probability Mass/Density Function: *a function that assigns probabilities to each possible outcome*

> *the probability function is like which cards are in the deck*

Observation: *a realization of a random variable . . .*

> *the observation is the card you drew*

Sample: *a collection of observations*

> *the sample is the record of cards you've drawn*

> *are the ages in the survey a random variable or observations?*

Random Variables

What is data? A sample.

Random Variable: *a function that assigns a number to each possible outcome of a random process (discrete or continuous)*

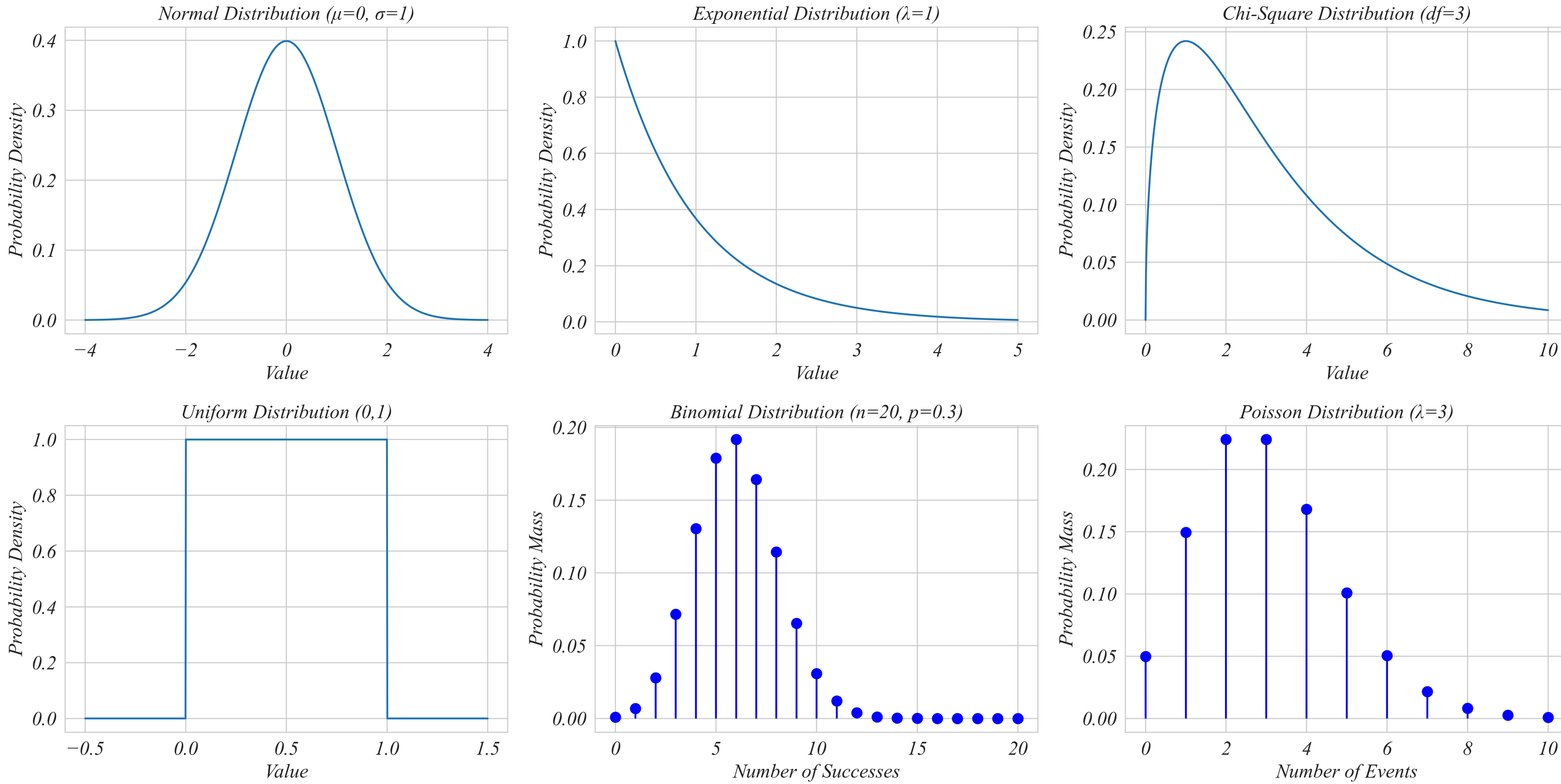
Probability Mass/Density Function: *a function that assigns probabilities to each possible outcome*

Observation: *a realization of a random variable*

Sample: *a collection of observations*

Some Known Distributions

... some well understood random variables



Random Variables: Known Distribution

What is data?

If we know the distribution:

- *We can compute mean, standard deviation, etc.*
- *We can easily answer questions about the population.*

Random Variables: Known Distribution

Example: Rich Person Bet

We’ll toss a coin once:

- *If it’s heads, you get $10 million*
- *If it’s tails, you pay $1 million*

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

> lets solve this theoretically and by simulation in python

> *we’ll find that the “sampling error” matches the distribution*

Random Variables: Known Distribution

Example: Bus Times

1. If I arrive at a random time, what's my expected wait time until the next bus?

> find the expected value of wait time

2. What's the probability I'll have to wait more than 20 minutes?

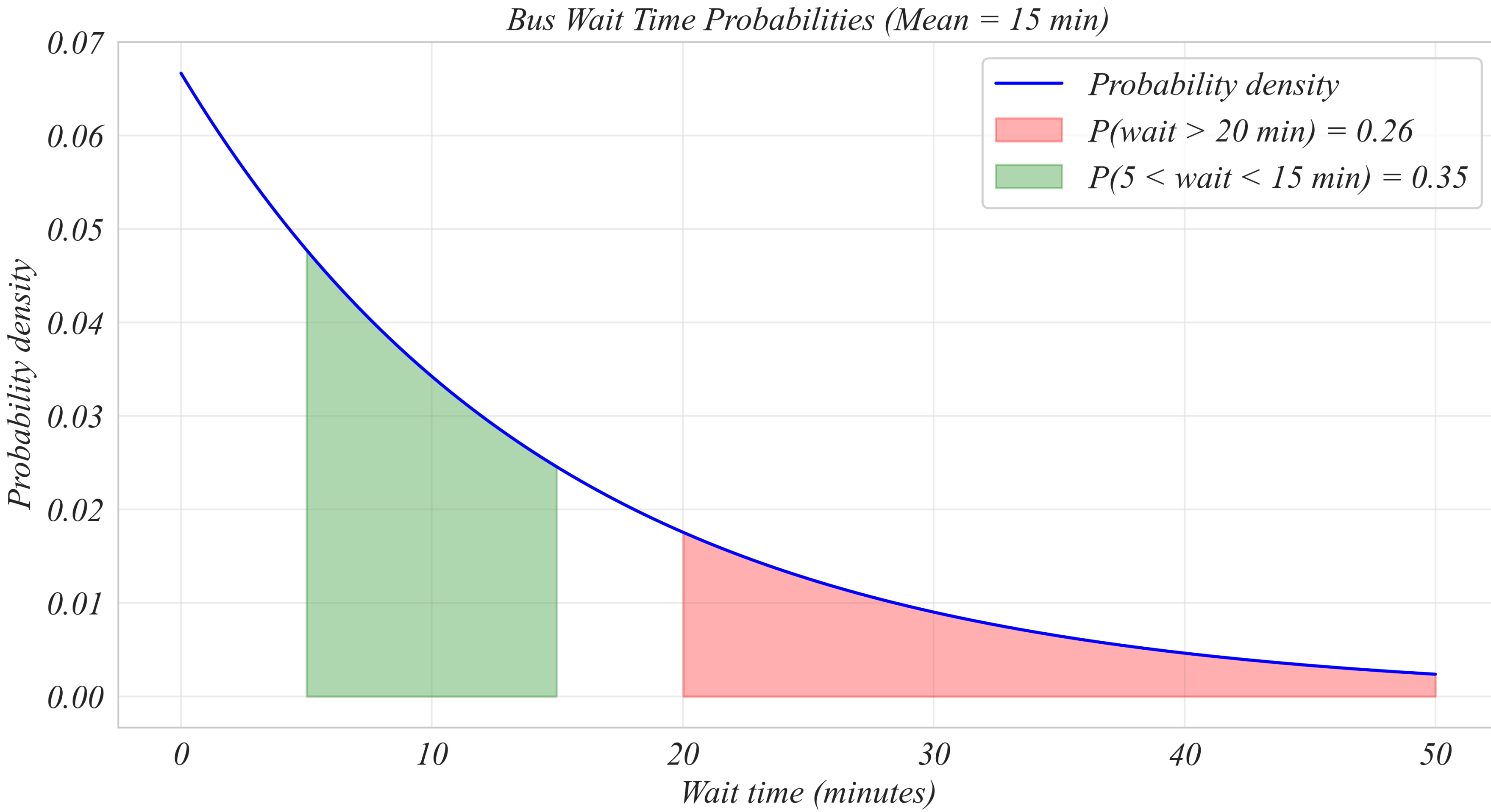
> find the area under the curve beyond 20 minutes

3. What's the probability my wait will be between 5 and 15 minutes?

> find the area under the curve between 5 and 15 minutes

Random Variables: Known Distribution

Example: Bus Times



Random Variables: Known Distribution

Example: Manufaturing Defects

Manufacturing defects of a part follow a normal distribution *(in cm)* with:

- `target_length = 100`
- `standard_error = 0.1`

We expect this to be normally distributed:

- *Multiple independent factors contribute to each defect.*
- *Small defects are more common than large ones.*
- *Positive and negative defects are equally likely.*

Random Variables: Known Distribution

Example: Manufaturing Defects

Manufacturing defects of a part follow a normal distribution *(in cm)* with:

- `target_length = 100`
- `standard_error = 0.1`

1. How many parts will be shorter than 99 cm?

> find the area under the curve below 99 cm

2. How many parts will have defects greater than 1/2 cm?

> find the area under the curve between 100 - 1/2 and 100 + 1/2

Random Variables

... main definitions

If we know the distribution:

- *We know the distribution, mean, standard deviation, etc.*
  1. *Probability function (f)*
  2. *Mean (μ)*
  3. *Standard deviation (σ)*

If we don't know the distribution:

- *We can compute*
  1. *Sample size (n)*
  2. *Sample mean ($\bar{x}$)*
  3. *Sample standard deviation (S)*
- *But how might we find the distribution?*

Random Variables: Unknown Distribution

What is the distribution of ages in this class?

- *Sample size (n):*
- *Sample mean ($\bar{x}$):*
- *Sample standard deviation (S):*

> *these are descriptives of the* **sample** *not the distribution ($\mu \neq \bar{x}$)*

> see notebook

> *we cannot see the distribution… we only observe realizations from it*

> *what* **can** *we say about ages when we don't know the distribution?*

Random Variables: Unknown Distribution

What can we know when we don’t know the distribution?

> *... next time!*