ECON 0150 | Economic Data Analysis

The economist's data analysis skillset.

Part 3.1 | Populations and Random Variables

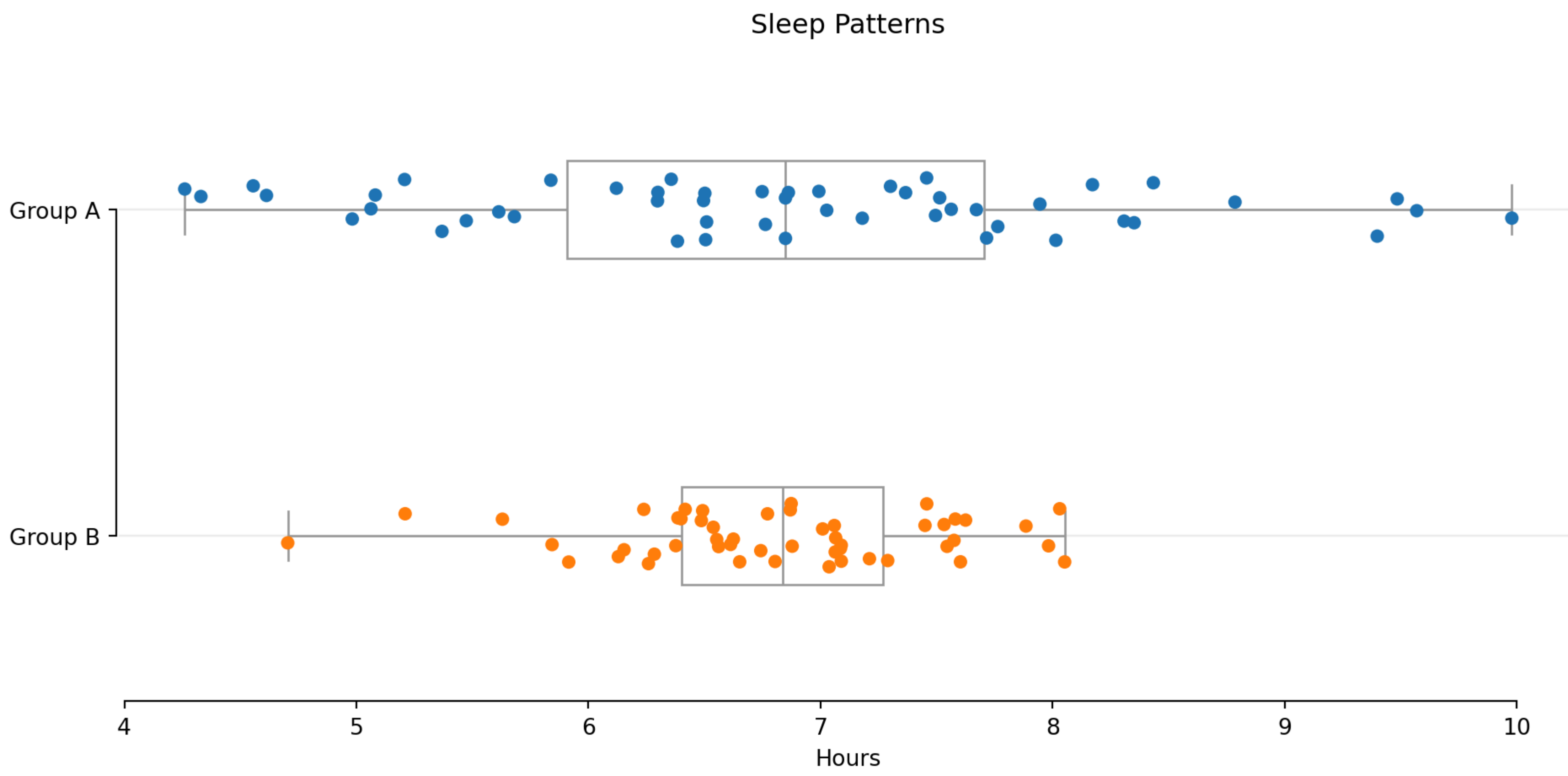
From Data to Understanding

We've spent Part 1 & 2 understanding our data… but what comes next?

- *We've mastered **visualizing** data*
- *We've developed skills to **summarize** and **transform** data*
- *But sometimes we need something more **precise** and **quantifiable***
- *And sometimes we want to say something about the **population**, not just our **sample***

Two Groups, One Question

Which group sleeps longer?



> the distributions overlap… how can we compare them precisely?

Measures of Location

Where is the "center" of each group?

Mean: The average value

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

Measures of Location

Where is the “center” of each group?

Mean: The average value

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
# Calculate means
mean_A = group_A.mean()
mean_B = group_B.mean()
```

```
Group A mean: 6.86 hours
Group B mean: 6.81 hours
```

> *group A sleeps longer* ***on average***

> *but notice the spread!*

Measures 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*

Measures of Dispersion

How spread out is the data?

Mean Deviation: difference between each value and the average

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

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

Measures of Dispersion

How spread out is the data?

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

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

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

Measures of Dispersion

How spread out is the data?

Variance: average squared difference from the mean

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

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

Measures of Dispersion

How spread out is the data?

Standard Deviation: A measure of spread

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

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

Measures of Dispersion

How spread out is the data?

Standard Deviation: A measure of spread

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

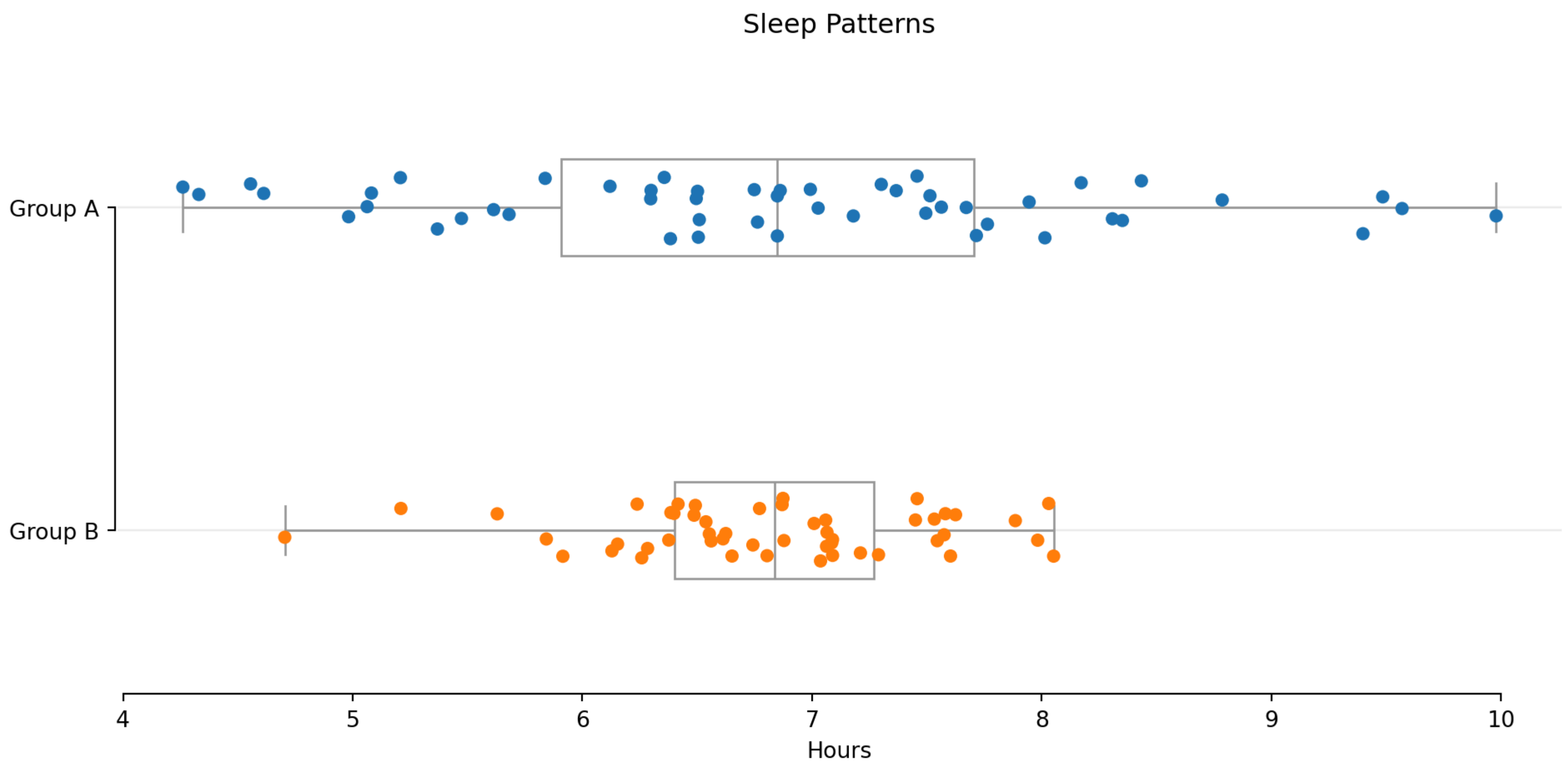
```
# Calculate standard deviations
std_A = group_A.std()
std_B = group_B.std()
```

```
Group A std dev: 1.39 hours
Group B std dev: 0.69 hours
```

> *Group A has **more variability** - some sleep much less, some much more*

Sample vs Population

What if I told you these groups are from different counties?

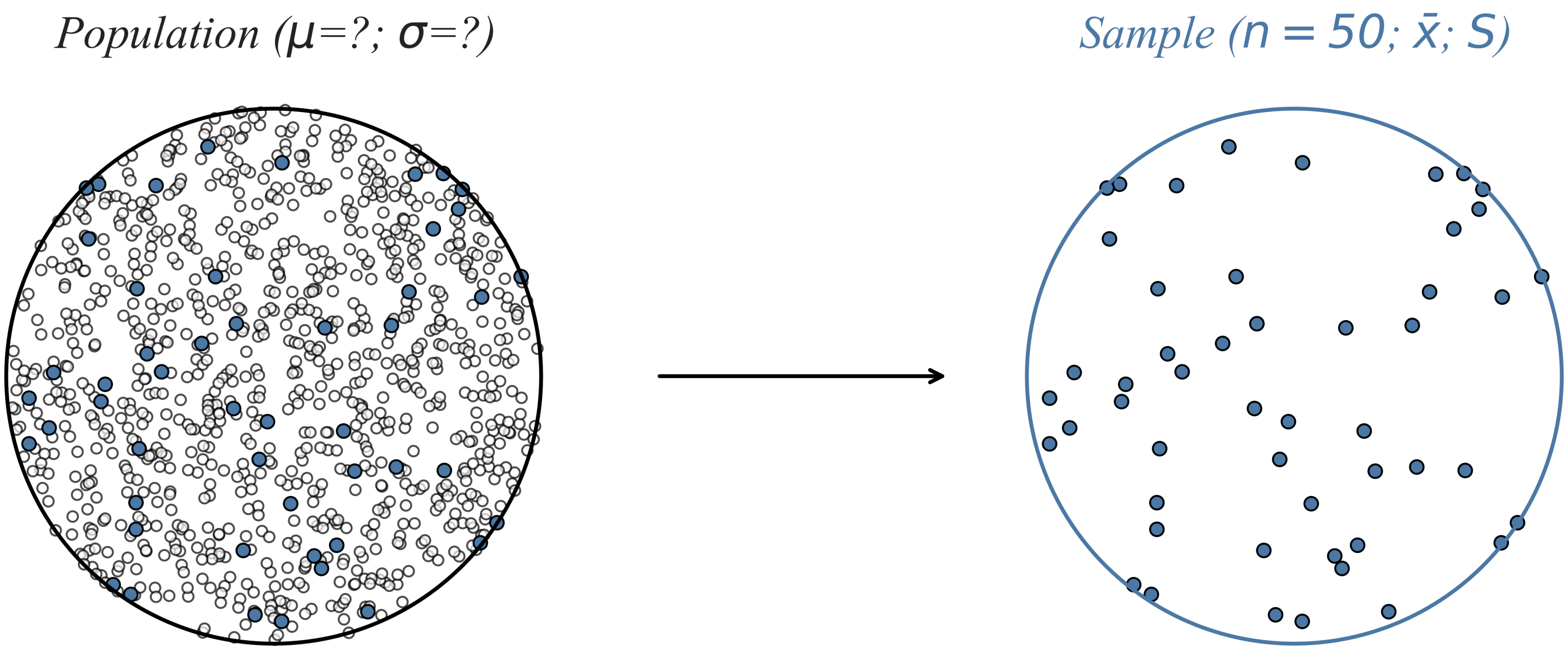


Old question: “Which **group** sleeps longer?” *(about the **data**)*

New question: “Which **county** sleeps longer?” *(about the **population**)*

Sample vs Population

The data is a sample drawn from a population.



Sample vs Population

What's the difference between our data and the population?

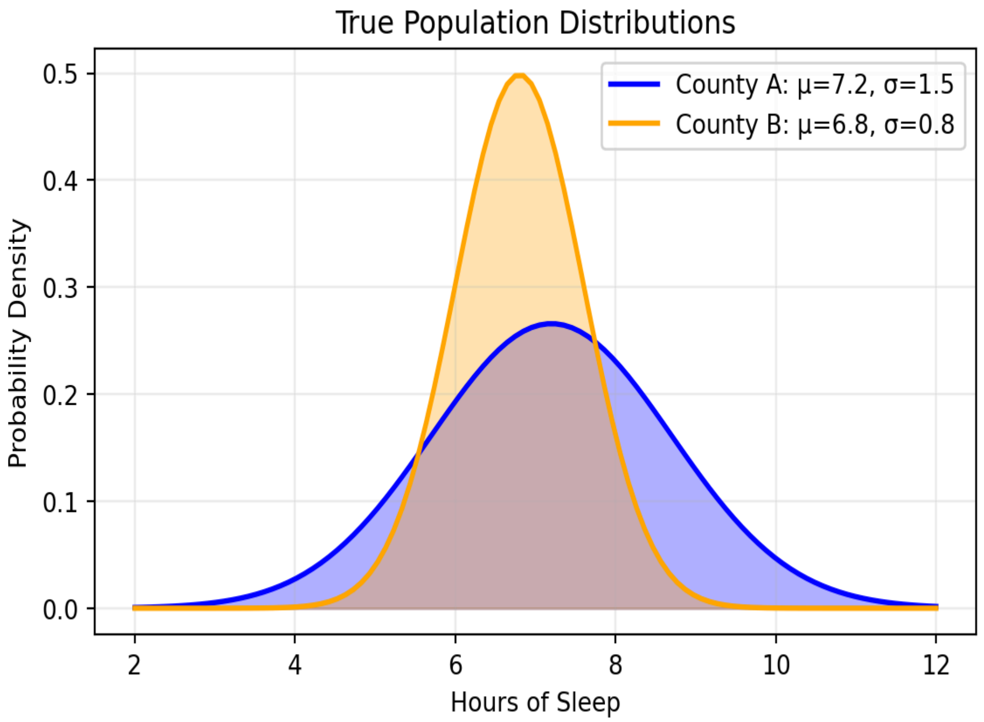
- ***The Data***: *50 individuals we happened to sample from each county*
- ***The Population***: *All people who could live in these counties*
  - *Even if we surveyed everyone today, tomorrow would bring new residents*
  - *The* ***population*** *is a theoretical concept - an infinite pool of possibilities*

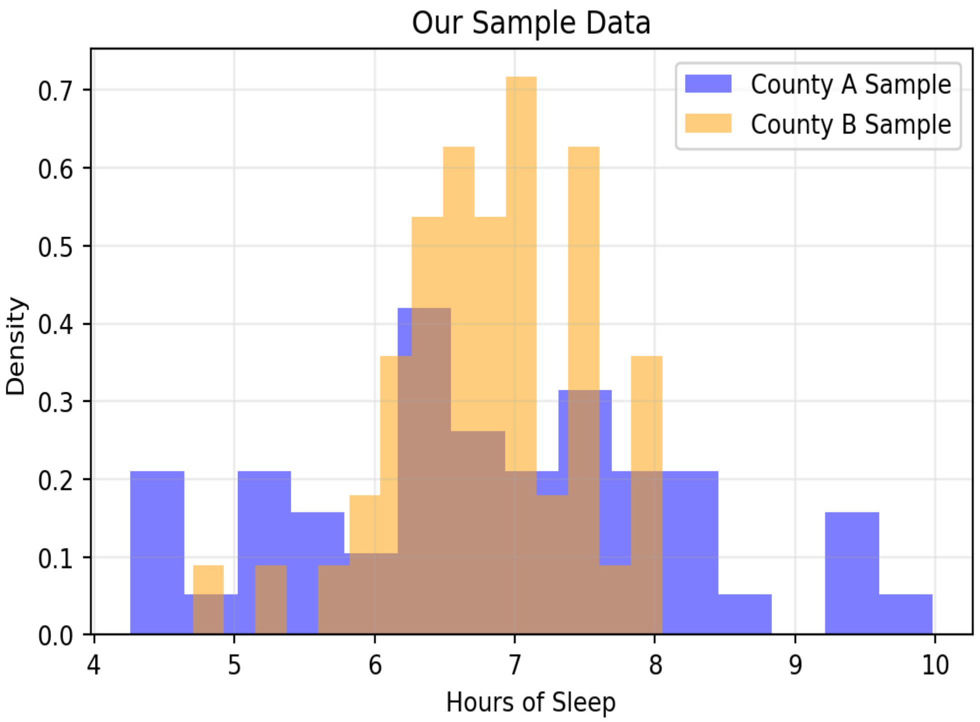
> *we observe* ***samples*** *but want to understand* ***populations***

Data is a Sample

A random variable generates our data.

Random Variable: A function that assigns numbers to random outcomes





*> data is a **sample** drawn from these a **random variable***

Known Distribution

What questions can we answer?

If we know County A:

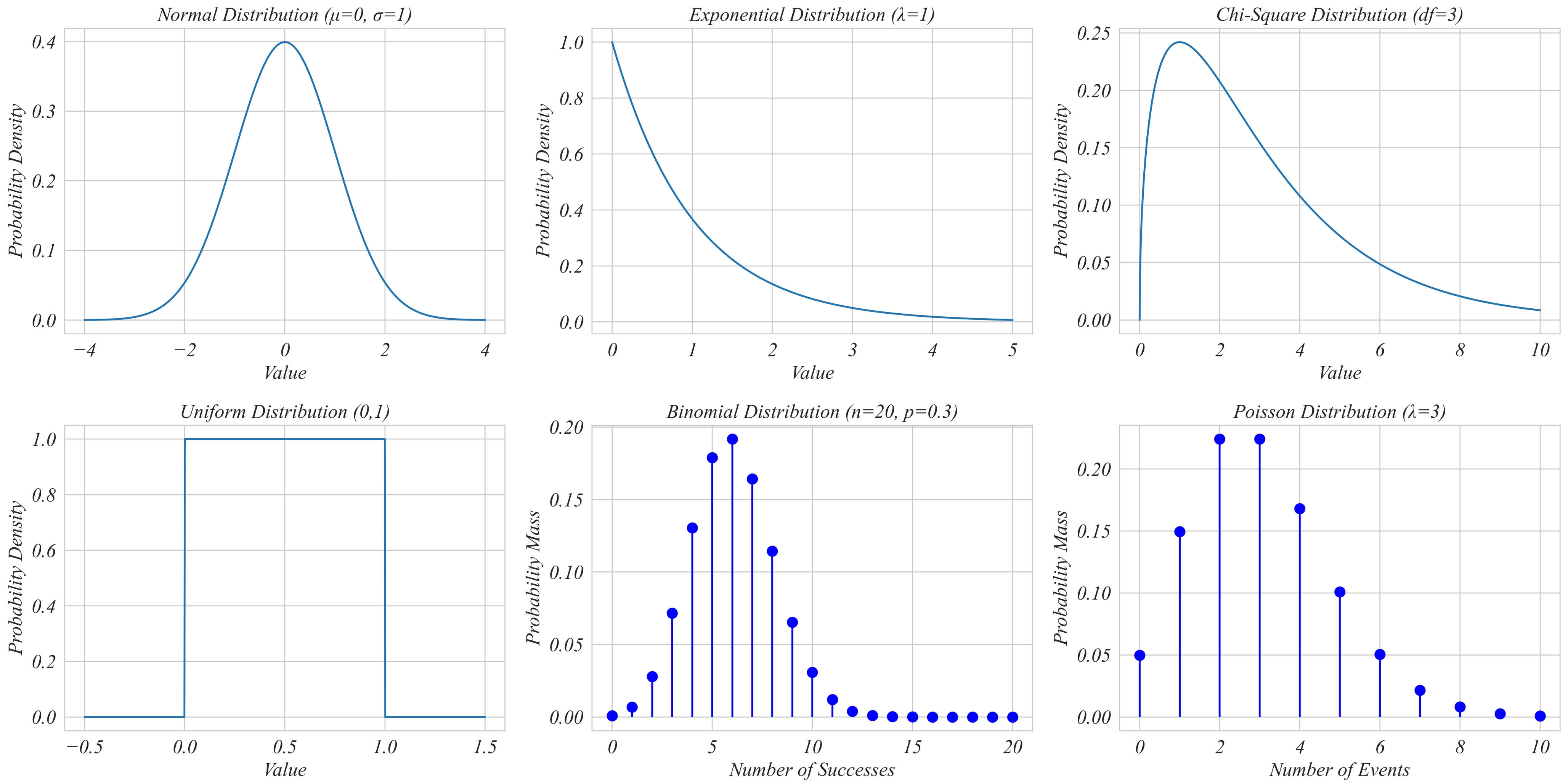
$$x_i \sim N(\mu = 7.2, \sigma = 1.5)$$

- *What proportion of the population sleeps less than 5 hours?*
- *What proportion of the population sleeps more than 9 hours?*
- *How much sleep does the middle 95% of the population get?*

> *with known distributions, we can answer* ***any*** *probability question!*

Random Variables Come in Many Shapes

Which distribution fits your data?



But What If We Don't Know the Distribution?

How do we move from sample to population?

What we observe:

- *Sample size:* $n = 50$
- *Sample mean:* $\bar{x} = 7.24$ *hours*
- *Sample standard deviation:* $s = 1.48$ *hours*

What we want to know:

- *Population mean:* $\mu = ?$
- *Population standard deviation:* $\sigma = ?$
- *Population distribution:* $f(x) = ?$

> *the sample statistics are* ***not*** *the same as population parameters!*

> $\bar{x} \neq \mu$, *and* $s \neq \sigma$

The Central Question

Can we say anything about the population when we only have a sample?

- *The **Central Limit Theorem** - our bridge to inference*
- ***Confidence intervals** - quantifying uncertainty*
- ***Hypothesis testing** - making decisions with data*
- *Moving from description to **inference***

> without seeing the population we can still answer questions about the population!