# POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti
(Based on DSS Materials)

## Lecture 13 | Probability

# Before we start

**Announcements:**

- Quizzes and Participation: On Canvas.
- GitHub page: https://github.com/umbertomig/POLI30Dpublic
- Piazza forum: Not sure what the link is. Ask your TA!
- Note to self: Turn on the mic!

## Before we start

**Recap:** We learned:

- The definitions of theory, scientific theory, and hypotheses.
- Data, datasets, variables, and how to compute means.
- Causal effect, treatments, outcomes, and randomization.
- Sampling, descriptive statistics, and correlation.
- Prediction of a binary and a non-binary variable.
- Strengths and weaknesses of observational and experimental studies.

**Great job!**

- Do you have any questions about these contents?

## Plan for Today

- Probability
- Events and Random Variables
- Probability Distributions
  - Bernoulli Distribution
  - Normal Distribution
  - The Standard Normal Distribution
- Population Parameters vs. Sample Statistics
- Law of Large Numbers and Central Limit Theorem

# Probability

- There are two ways of interpreting probability:
- Frequentist: The probability of an event is the proportion of its occurrence among infinitely many identical trials
  - Example: the probability of heads when flipping a coin
- Bayesian: Probabilities represent one's subjective beliefs about the relative likelihood of events
  - Example: the probability of rain in the afternoon

# Events and Random Variables

- **Events**: Sets of outcomes that occur with a particular probability
- Most things in our lives can be considered **events**
  - Example: Being 6 feet or taller
- **Random Variables**: Assigns a numeric value to each mutually exclusive event produced by a trial
  - As soon as we assign a number to an event, we create a random variable.

# Events and Random Variables

- Random variable *tall*:

$$\text{tall}_i = \begin{cases} 1 \text{ if individual i is 6 feet or taller} \\ 0 \text{ if individual i is not} \end{cases}$$

# Probability Distribution

- Each random variable has a **Probability Distribution**:
- Likelihood of each value the variable can take.
- Probability distribution of *tall*:
- $\mathbb{P}(\text{tall})$ = probability of being tall
- $\mathbb{P}(\text{not tall})$ = probability of not being tall

# Probability

- Probabilities are always between zero and one:
  - $\mathbb{P}(\text{tall}) \in [0, 1]$
  - $\mathbb{P}(\text{not tall}) \in [0, 1]$
- Do you agree that a person can either be tall or not? If yes:
  - $\mathbb{P}(\text{tall}) + \mathbb{P}(\text{not tall}) = 1$
- Do you agree that a person can either be tall or not? If yes:
  - $\mathbb{P}(\text{neither tall nor not tall}) = \mathbb{P}(\emptyset) = 0$

# Probability Distributions

- We distinguished between binary and non-binary (random) variables
  - Binary variables are?
  - Non-binary variables are?
- We will focus on two different types of probability distributions
- *Bernoulli distribution:* probability distribution of a binary variable.
- *Normal distribution:* probability distribution we commonly use as a good approximation for many non-binary variables.

# Bernoulli Distribution

- Probability distribution of a binary variable.
- It is characterized by one parameter: $p$.
    - If $\mathbb{P}(X = 1) = p$, then $\mathbb{P}(X = 0) = 1 - p$.
    - If $\mathbb{P}(X = 1) = 0.8$, then what is $\mathbb{P}(X = 0)$?
- The *mean* of a Bernoulli distribution is $p$
- The *variance* of a Bernoulli distribution is $p(1 - p)$

# Example: Passing the class

- Random Variable: $pass = \{0, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$

$$\text{where: } pass_i = \begin{cases} 1 \text{ if student } i \text{ passed the class} \\ 0 \text{ if student } i \text{ didn't pass the class} \end{cases}$$

- Probability distribution: Bernoulli, where $p =?$
  - $\mathbb{P}(\text{pass}) = p$
  - $\mathbb{P}(\text{not pass}) = 1 - p$

## Example: Passing the class

$$pass = \{0, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$$

- What is the probability that a student passes the class? What is $p$?

$$\mathbb{P}(\text{pass}) = \frac{\text{number of students who passed}}{\text{total number of students}}$$

$$= \frac{\text{frequency of 1s}}{\text{total number of observations}} = \quad ?$$

- $\mathbb{P}(pass) = p = 0.9$.
  - Interpretation: The probability of passing the class is 90%.

# Example: Passing the class

$$pass = \{0, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$$

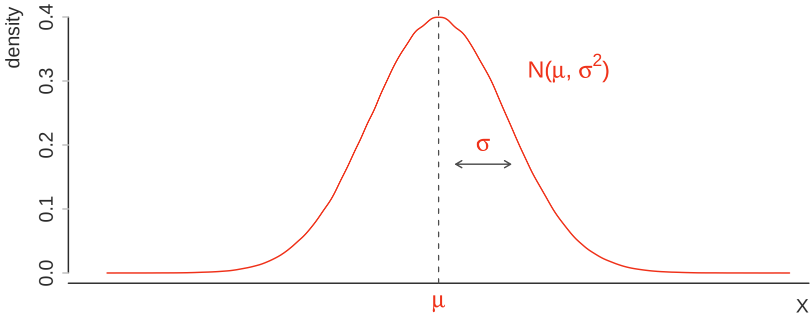
- What is the probability that a student will fail? What is $1 - p$?

$$\mathbb{P}(not\ pass) = \frac{\text{number of students who did not pass}}{\text{total number of students}}$$
$$= \frac{\text{frequency of 0s}}{\text{total number of observations}} = \ ?$$

- $\mathbb{P}(not\ pass)$ = 1-p = 1-0.90 = 0.1.
  - Interpretation: The probability of failing the class is 10%.

# Normal Distribution

- Probability distribution is commonly used as a good approximation for many non-binary variables.
- It is characterized by two parameters: $\mu$ (mu, the mean) and $\sigma^2$ (sigma-squared, the variance).

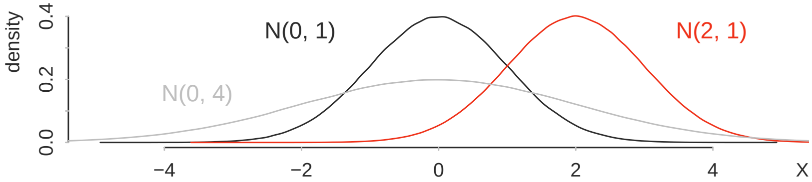


# Normal Distribution

- *Normal random variables* are variables *normally* distributed

$$X \sim N(\mu, \sigma^2)$$

- Examples of Normal distributions:

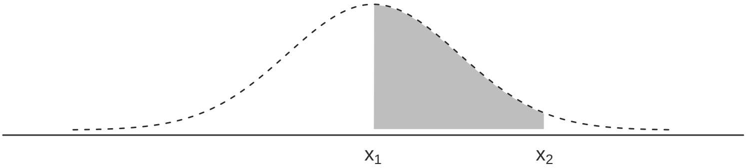



- What is the mean and variance of $N(0, 1)$?

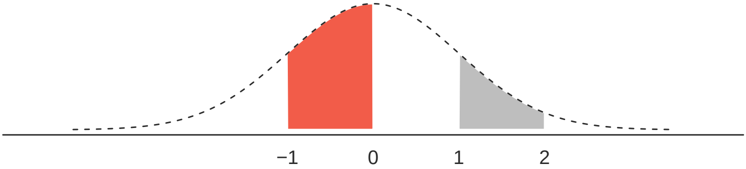
## Normal Distribution

- The probability density function of the normal distribution represents the likelihood of each possible value of a normal r. v. can take.
- We can use it to compute the probability that X takes a value within a given range:

$$\mathbb{P}(x_1 \leq X \leq x_2) = \text{area under the curve between } x_1 \text{ and } x_2$$



# Normal Distribution




$$\mathbb{P}(-1 \leq X \leq 0) \;\; < \text{or} > \;\; \mathbb{P}(1 \leq X \leq 2)?$$

# The Standard Normal Distribution

- Normal distribution with mean 0 and variance 1 (and standard deviation = 1)
- In mathematical notation, we refer to the standard normal random variable as $Z$ and write it as

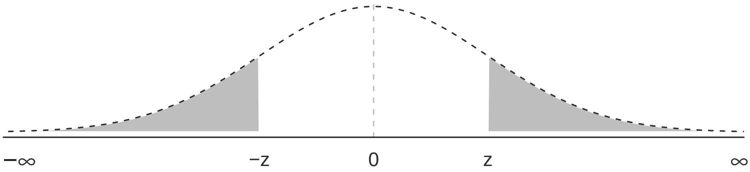
$$Z \sim N(0, 1)$$

- Note that this $Z$ has nothing to do with confounding variables
- $Z$ has two useful properties...

# Normal Distribution

First, since $Z$ is symmetric and centered at 0:

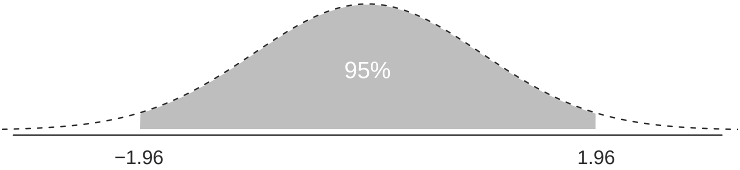
$$P(Z \leq -z) \;=\; P(Z \geq z) \qquad \text{(where } z \geq 0\text{)}$$



# Normal Distribution

Second, about 95% of the observations of $Z$ are between -1.96 and 1.96:

$$P(-1.96 \leq Z \leq 1.96) \approx 0.95$$



# How to Transform A Normal Random Variable Into the Standard Normal Random Variable

$$\boxed{\text{if } X \sim N(\mu, \sigma^2), \qquad \frac{X-\mu}{\sigma} \sim N(0, 1)}$$

- Example: If $X \sim N(4, 25)$, $\frac{X - ?}{?} \sim N(0{,}1)$

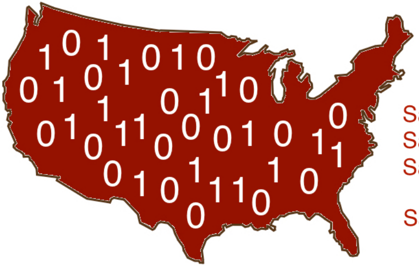
## Population Parameters vs. Sample Statistics

- When we analyze data, we are usually interested in the value of a **parameter** at the population level.
  - Proportion of candidate A supporters among all voters in a country.
- We typically only have access to **statistics** from a small sample:
  - Proportion of supporters among the voters who responded to a survey.
- The sample statistics differ from the population parameters because the sample contains noise.
  - This noise comes from **sampling variability**.

# Sampling variability

- The value of a statistic varies from one sample to another.
  - Each sample contains different observations drawn from the target population.
- This is true even when the samples are drawn using the same method, such as random sampling.
- Smaller sample size generally leads to greater sampling variability.

# What proportion of US voters supports candidate A?



$x \begin{cases} 1 = \text{support} \\ 0 = \text{no support} \end{cases}$

Sample 1: {1,0,0,1,0,0,1} $\bar{x}_1 = 3/7$

Sample 2: {1,1,1,0,1,0,0} $\bar{x}_2 = 4/7$

Sample 3: {0,0,1,0,1,0,1} $\bar{x}_3 = 3/7$

...

Sample k: {0,0,1,0,1,0,0} $\bar{x}_k = 2/7$

- If we repeatedly draw a random sample from the population, we will get different proportions of support ($\overline{X}$).

# Sampling variability

- How can we figure out what we want to know: the proportion of support among the whole population?
- The two large sample theorems help us understand the relationship between population parameters and sample statistics.
- As we will see next class, we can use them to draw conclusions about population parameters using data from a sample.

# Sampling variability

- **Population Parameters**:
- **Expectation or Population Mean, $\mathbb{E}(X)$**: The average value of the random variable X at the population level
- **Population variance, $\mathbb{V}(X)$**: The variance of the random variable X at the population level

# Sampling variability

- **Sample Statistics**:
  - **Sample mean, $\overline{X}$**
  - **Sample variance, $var(X)$**
- $\overline{X}$ and $var(X)$: Sample statistics:
  - Vary from sample to sample
- $E(X)$ and $V(X)$ are population characteristics:
  - Same (unknown) value.

# Sampling variability

POPULATION
X ~ distribution
centered: E(X)
spread: V(X)

what we want to know

SAMPLE
n = sample size
$\overline{x}$ = sample mean
var(x) = sample variance

what we can measure

# The Law of Large Numbers

*As the sample size increases,* $\overline{X} \rightarrow_p \mathbb{E}(X)$

As $n$ increases, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ converges to $\mathbb{E}(X)$

- Check R Script `LLNsims.R` for a simulation to see how it works.

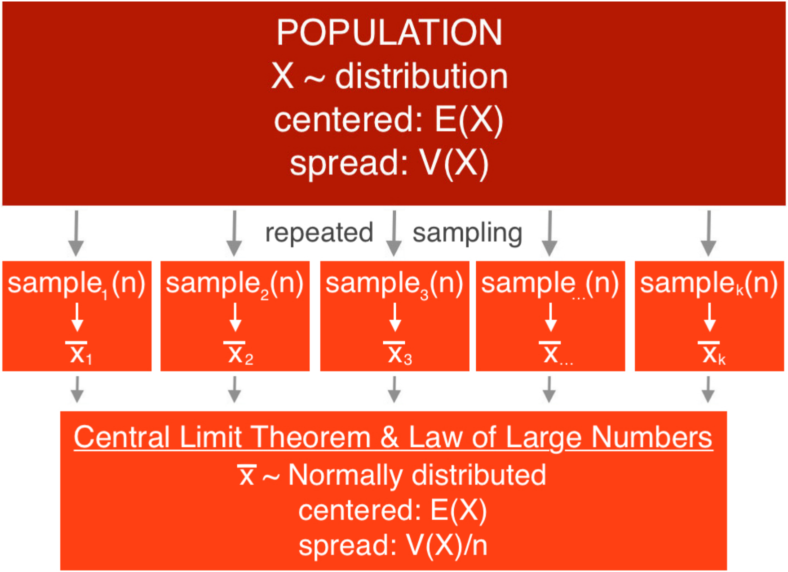
## The Central Limit Theorem

*As the sample size increases, the distribution of the sample means can be approximated by a normal distribution*

$$\text{as } n \text{ increases,} \quad \frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}} \overset{\text{approx.}}{\sim} N(0, 1)$$

- Check R Script `CLTsims.R` for a simulation to see how it works.

# The Central Limit Theorem



## The Central Limit Theorem

- Let multiple 1,000 observations samples from a random variable, with mean of the means at 10 and variance 0.002.
- What is our guess for the population parameters?
- Mean:
  - 10. $\overline{X} \approx \mathbb{E}(X)$
- Variance:
  - 2 $var(\overline{X}) \approx \frac{\mathbb{V}(X)}{n}$

## Summary

- **Today's Class:**
  - Probability theory
    - Probability, Events, Random Variables.
    - Distributions
    - Central Limit Theorem and Law of Large Numbers.
- **Next class**:
  - Hypothesis Testing

Questions?

See you in the next class!