1️⃣ Population vs Sample

Population

- The **entire group** you want to learn about
- Usually large, varied, hard to measure

Sample

- A **subset of the population** actually observed

Example:

- Population: *All UT undergraduates*
- Sample: *10 UT students surveyed*

2️⃣ Parameters vs Statistics

Parameter

- A **numerical value describing a population**
- Usually unknown

Statistic

- A **numerical value calculated from a sample**
- Used to estimate parameters

Quantity	Population (Parameter)	Sample (Statistic)
Mean	μ (mu)	$\bar{x}$ (sample mean)
Std Dev	σ (sigma)	s
Proportion	p	$\hat{p}$

Example:

- μ = true average GPA of all UT students
- $\bar{x}$ = average GPA of 10 surveyed students

3️⃣ Research Question

- Clearly states **what you want to learn**
- Defines the **population of interest**
- Should not be **too broad or too narrow**

Example (Good): What factors impact attendance at UT home volleyball games?

4️⃣ Random Sampling

Random Sample

- Every individual in the population has an **equal chance** of being selected

Why it matters:

- Reduces bias
- Makes results more representative

Important: Humans are **bad** at choosing randomly → use random number generators.

5️⃣ Bias vs Sampling Error (Large and Random Samples are Good)

Bias

- Systematic error in sampling
- Comes from *how* data is collected
- **Does NOT go away** with larger samples

Sampling Error

- Natural variation from taking a sample
- **Decreases with larger sample size**

Examples:

- Bias: Surveying only TikTok users to represent all Austinites
- Sampling error: Two random samples give slightly different means

---

6️⃣ Variables & Observations

Variables

- Characteristics measured
- Go in **columns (x axis)**

Observations

- Individual units measured
- Go in **rows (y axis)**

Example:

- Variables: GPA, age, major
- Observation: One student's responses

7️⃣ Types of Variables

Numeric (Quantitative)

- Numbers with meaningful magnitude

Examples: Age, Number of texts per day

Categorical (Qualitative)

Describe group membership

Nominal (No order)

- No natural ranking

Examples: Race, Can do a handstand (Yes/No)

Ordinal (Ordered)

- Categories have an order

Examples: Exercise frequency (Never → Often), Education level

---

8️⃣ Independent vs Dependent Variables

Independent (Predictor)

- Variable that explains or predicts

Dependent (Outcome)

- Variable being explained

Example:

- Predictor: Weather
- Outcome: Attendance at UT volleyball games

9️⃣ Confounding Variables

Confounder

- A variable that **distorts or masks** the relationship between two variables
- Related to both predictor and outcome

Classic Example:

- Panhandlers ⬆️ and traffic accidents ⬆️
- Confounder: **Busy intersections**

➡️ Panhandlers don't cause accidents — traffic volume does.

🔟 Data Cleaning (Conceptual)

Before analysis:

- Remove impossible values (out of range)
- Fix inconsistent text
- Handle missing data
- Look for typos and entry errors

2️⃣ Categorical Data: Describing & Displaying

Frequency

- **Count** in each category

Relative Frequency

- **Proportion** in each category

Relative Frequency=Category Count / Total Observations

Example:

28 "Yes" responses out of 97 → 28/97=0.2887=28.87%

Frequency Table (Example)

Category	Frequency	Relative Frequency
Yes	28	28.87%
No	62	63.92%
It's complicated	7	7.22%

Bar Plot

- Used for **categorical data**
- Bars **do not touch**
- X-axis: categories
- Y-axis: frequency or proportion

3️⃣ Numeric Data: Describing & Displaying

Measures of Center

Mean (average)

$\bar{x}=1/n\sum x_i$ sum of x divided by the number of x's

Median

- Middle value when data are **ordered**

Example:

Data: 2, 4, 4, 5, 6, 6, 8, 9, 12

- Mean ≈ 5.9, Median = 6

Measures of Spread

Variance

$s^2=1/n-1\sum(x_i -\bar{x})^2$

Standard Deviation

$s=sqrt(s^2)$

Interquartile Range (IQR)

$IQR=Q3-Q1$

Quartiles

Q1: 25th percentile, **Q2**: Median, **Q3**: 75th percentile

Five-Number Summary

(Min,Q1,Median,Q3,Max)

Example:

Texts per day: (0, 15, 30, 70, 500)

---

4 Visualizing Numeric Data

Histogram

- Used for **numeric data**
- Bars **touch**
- X-axis: value ranges (bins)
- Y-axis: frequency or proportion

Boxplot

- Shows **5-number summary**
- Box = IQR
- Line = median
- Whiskers = range (or up to 1.5×IQR)

---

5 Shape of a Distribution

Symmetric

- Left and right sides are mirror images
- Mean ≈ Median

Right-Skewed (Positive Skew)

- Long tail to the **right**
- Mean > Median

Left-Skewed (Negative Skew)

- Long tail to the **left**
- Mean < Median

6️⃣ Choosing the Right Center & Spread

Distribution	Center	Spread
Symmetric	Mean	Standard Deviation
Skewed / Outliers	Median	IQR

7️⃣ How to Describe a Numeric Variable (EXAM RULE)

Always mention **ALL THREE**:

1. **Center** (mean or median)
2. **Spread** (SD or IQR)
3. **Shape** (symmetric / skewed)

8️⃣ Example Full Description

> The distribution of texts sent per day is **right-skewed**, with a **median of 30 texts** and an **IQR of 55 texts**.

📊 Describing Bivariate Data

1️⃣ Two Categorical Variables

Contingency Table

- Table of **counts** for category combinations

Marginal Distribution

- Distribution of **one variable alone** (row or column totals)

Conditional Distribution

- Distribution of **one variable given a category of the other (inside tables)**

Example:

Among students who **don't exercise**, what % are in a relationship?

count in category/(row/column total)

Graphs

- **Grouped bar plot**
- **Mosaic plot**

2️⃣ One Categorical + One Numeric Variable

Summary Statistics

- Compare **mean / median / SD / IQR** by category

Graphs

- **Grouped boxplots**
- **Grouped histograms**

Example: Median texts sent per day by relationship status

3️⃣ Two Numeric Variables

Scatterplot

- X-axis: predictor

- Y-axis: outcome

🔹 **Pearson Correlation (r)**

What it measures

- Strength and direction of **linear** relationship

$-1 \leq r \leq 1$

Value of r	Meaning
r ≈ 1	Strong positive
r ≈ −1	Strong negative
r ≈ 0	Weak/no linear relationship

When NOT to Use r

- Relationship is **nonlinear**
- **Outliers** strongly affect data

Interpretation Example

> The correlation between BMI and systolic blood pressure is **r = 0.62**, indicating a **moderate positive linear relationship**. Moderate, strong, weak correlation

🔹 **Probability Basics**

Probability

$0 \leq P(\text{event}) \leq 10$

Sample Space (S)

- Set of all possible outcomes

Event

- A subset of the sample space

Example (fair die):

- P(rolling a 4)=1/6 P(not rolling a 4)=5/6

🔹 **Probability Distribution**

- Lists **all possible outcomes** and their probabilities
- Probabilities must **sum to 1**

🔹 **OR Rule (Addition Rule)**

General Addition Rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Mutually Exclusive Events

- Cannot occur together

$P(A \text{ and } B) = 0$ $P(A \text{ or } B) = P(A) + P(B)$

Example:
60% have high cholesterol 70% have high blood pressure 50% have both
P(HC or BP)=0.6+0.7−0.5=0.8

🔹 **AND Rule (Multiplication Rule)**
General Multiplication Rule
$P(A \text{ and } B)=P(A)\times P(B \mid A)$

🔹 **Conditional Probability**
Definition
$P(B \mid A)=P(A \text{ and } B)/P(A)$
Interpretation: Probability that **B occurs given A already happened**

🔹 **Independence**
Events A and B are **independent if and only if**:
$P(B \mid A)=P(B)$
Equivalent test:
$P(A \text{ and } B)=P(A)\times P(B)$

🔹 **Complement Rule**
$P(\text{not } A)=1-P(A)$

🔹 **Law of Total Probability**
If event A can happen under different scenarios B1,B2,…,Bk
$P(A)=\sum_{i=1 \text{ to } k} P(Bi)\cdot P(A \mid Bi)$
Tiny Example (mice survival):
Environments:

- Low: $P(L)=0.30$, $P(S \mid L)=0.80$
- Mod: $P(M)=0.20$, $P(S \mid M)=0.30$
- High: $P(H)=0.50$, $P(S \mid H)=0.10$

$P(S)=0.30(0.80)+0.20(0.30)+0.50(0.10)$
$P(S)=0.24+0.06+0.05=0.35$

🔹 **Bayes' Theorem**
Flips conditional probability:
$P(A \mid B)=P(B \mid A)\cdot P(A)/P(B)$
Where:

- $P(A)$ = prior
- $P(B \mid A)$ = likelihood
- $P(B)$ = normalizing constant (often from Law of Total Probability)

Tiny Example (surviving mouse from high-stress):
$P(H \mid S)=P(S \mid H)\cdot P(H) / P(S)=0.10\cdot 0.500.35\approx 0.143$
Interpretation: About **14.3%** of surviving mice came from the high-stress environment.