① Producing Data

Population
Sample Frame
Sample

SAMPLING

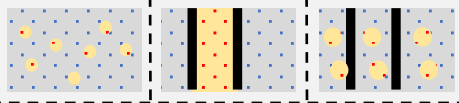
It is essential that the sample is representative of the larger population.

Types of samples:

chance of bias ↑

- **Volunteer Sample:** almost always biased
- **Convenience Sample:** at the right moment and time
- **Sampling Frame:** only a selected sub-group
- **Systematic Sampling:** all pop., but not completely random
- **Probability Sampling**:

Simple Random Samp. | Custer Sampling | Stratified Sampling

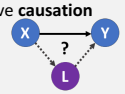


- **Multistage sampling**: consecutive forms of sampling

STUDY DESIGN

Observational:

Prospective or Retrospective:
- Potentially unlimited number of **lurking variables** make it difficult to prove **causation**



- **memory or honesty** not reliable

Survey:
- Open Q: too broad
- Close Q: too narrow
- Unbalanced response options
- Planting ideas in Q
- Complicated or sensitive Q

mitigate with anonymity and/or Random responses

Experiment:

Vocabulary:
- *X : Factor*
- *X1,X2,… : Treatments (ttt)*
- *Individuals: Subjects*

Hawthorne effect:
people behave differently in an experiment than in real life.

How to reduce bias:
- **randomly assign ttt :** *randomized controlled experiment*
- **blind or double-blind**
- Measure outside variables with **blocking** (~stratification)
- **Matched pairs** to compare responses for similar subjects

Population

ONE VARIABLE

Point Estimation

We estimate an **unknown parameter** using a **single number** calculated from the sample data.

p ⟵ ⟵ $\hat{p}$
x ⟵ ⟵ x

Interval Estimation

We estimate an **unknown parameter** using an **interval of values** likely to **contain the true value,** and state how **confident** we are this interval is correct.

Estimating the mean μ

If we can apply the **Central Limit Theorem** AND **σ known**:

$\mu \in x \pm (z^* \frac{\sigma}{\sqrt{n}})$

Confidence multiplier:
level 95% => z^*=2
level 90% => z^*=1.645
level 99% => z^*=2.576

If we know margin or error (m) we can define the sample size n:

$n = \left(\frac{z^* \sigma}{m}\right)^2$

① *Range Rule of Thumb:* $\sigma \leq {}^{range}/_{4}$

If we **cannot** apply **Central Limit Theorem** OR **σ unknown**:

$\mu \in x \pm t^* \frac{S}{\sqrt{n}}$ t^*depends of confidence level AND sample size n

① *for large value of* ***n****,* $z^* \sim t^*$

Estimating the proportion p

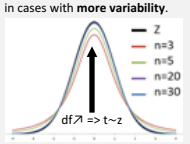
If we can apply the **Normal Approximation to the Binomial:**

$p \in \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$n = \frac{z^* \hat{p}(1-\hat{p})}{m^2}$ conservative approach $\hat{p} = 1/2$ $\Rightarrow n = 1/m^2$ $m = 1/\sqrt{n}$

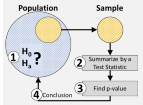
Normal Distribution vs t-distribution

t-distribution is more appropriate in cases with **more variability**.



t is assigned a **degree of freedom**, and gets similar to z for a high **df**.

Hypothesis Testing



We have some **claim** about the **population**, and we check whether or not the **data** obtained from the **sample** provides **evidence against** this claim.
Significance level of the test α, usually 5%

Testing for p	Testing for μ with σ	Testing for μ without σ

1 – Stating the claims

Null hypothesis: $H_0 : p = p_0$
Alternative Hypothesis : $H_a : p < p_0$ or $p > p_0$ (one-sided) or $p \neq p_0$ (two-sided)

2 – Test Statistic difference between $\hat{p}$ or x and p_0 or μ_0 in standard deviation

z-test for the population p
If **sample random** and can apply **Normal Distribution to the Binomial**

$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

z-test for the population μ
If **sample random** and can apply **Central Limit Theorem**

$z = \frac{x - \mu_0}{\sigma/\sqrt{n}}$

t-test for the population μ
If **sample random** and can apply **Central Limit Theorem**

$t = \frac{x - \mu_0}{s/\sqrt{n}}$ with df = n-1

① *for large value of* ***n****, we can use z-test*

3 – Find p-value probability of $\hat{p}$ when we assume H_0 **is true**

for $H_a : p < p_0$ => $P(Z \leq z)$	for $H_a : p < p_0$ => $P(Z \leq z)$	for $H_a : p < p_0$ => $P(T_{n-1} \leq t)$
for $H_a : p > p_0$ => $P(Z \geq z)$	for $H_a : p > p_0$ => $P(Z \geq z)$	for $H_a : p > p_0$ => $P(T_{n-1} \geq t)$
for $H_a : p \neq p_0$ => 2 * $P(Z \leq \lvert z \rvert)$	for $H_a : p \neq p_0$ => 2 * $P(Z \leq \lvert z \rvert)$	for $H_a : p \neq p_0$ => 2 * $P(T_{n-1} \leq \lvert t \rvert)$

4 – Conclusion based on p-value (or with null value vs alternative value **interval estimation**)

Determine if results are statistically significant (p-value< α?) to reject H_0 and accept H_a
if yes, give a 95% confidence interval for the alternative value.
Conclude, based on context (⚠statistical significance ≠ practical significance)

TWO VARIABLES

RESPONSE (/dependent) VARIABLE		
EXPLANATORY (/independent) VARIABLE	CATEGORICAL	QUALITATIVE
CATEGORICAL	**C → C**	**C → Q**
QUALITATIVE	**Q → Q** *(for linear relationships)*	

C → C

1 - Stating the claim
H_0: no relationships between two variables
H_a: relationship between the two variables
2a- Check conditions
(i) The sample is random
(ii) Large sample
- *rule of thumb:* all expected counts >1 AND 80% expected counts >5
- *conservative rule of thumb:* all expected counts >5

2b- Test statistic

$\text{expected count} = \frac{\text{column total} \times \text{row total}}{\text{table total}}$

Chi-Square test for independence

$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$ → χ^2contribution

C → Q

k=2 independent groups	k=2 "Matched-Pairs"	k>2 independ. groups
1- Stating the claim	1- Stating the claim	1 - Stating the claim
$H_0: \mu_1 - \mu_2 = 0$	$H_0: \mu_1 - \mu_2 = \mu_d = 0$	$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$
$H_a: \mu_1 - \mu_2 >< \neq 0$	$H_a: \mu_d >< \neq 0$	Ha: not all μ's are equal
2a- Check conditions	2a- Check conditions	2a- Check conditions
- Samples independent	- sample of d random	- sample groups indep.
- can apply CLT 2 groups	- can apply CLT	- can apply CLT to all samples
2b- Test statistic	2b- Test statistic	2b- Test statistic
Two-Samples t-test	*One-Sample t-test for the mean of the difference*	*ANalysis Of VAriation based on F-Distribution*
$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$t = \frac{\bar{x}_d}{s_d/\sqrt{n}}$	
4 - CI for $\mu_1 - \mu_2$: $\bar{y}_1 - \bar{y}_2 \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}$		

Q → Q *(for linear relationships)*

A/Explanatory Data Analysis
Scatterplot, retrieve correlation coefficient r
B/Inference
1- Stating the claim
H_0: no linear relationship between X and Y
H_a: linear relationship between X and Y

r: strength of relationship
r^2: evidence of relationship

2a- Check conditions
- observed data look linear
- observations independent
- no extreme outliers
- large sample size (n ≥ 30)

2b- Test statistic
t-test for the slope

Apply 4 steps of Hypothesis Testing
1. Stating the claim
2. Check conditions & Test statistic
3. Find p-value
4. Conclusion & Confidence Interval

specific to relationship types
standard

③ Inference

① Sampling Distribution

Statistics vary from sample to sample due to **sampling variability**, and therefore can be regarded as **random variables** whose distribution is called **sampling distribution** (very theoretical).

PARAMETER:
numbers that describe a **population**
- p (for categorical variable)
- μ (for quantitative variable)
- σ (for quantitative variable)

VS

STATISTIC:
numbers computed from a **sample**
- $\hat{p}$
- x
- s

Those are actually random variables

$\hat{p}$ **has a normal distribution with:**
mean of $\hat{p}$: $\mu_{\hat{p}} = p$
std dev of $\hat{p}$: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
↳ if we can use normal approximation of the binomial

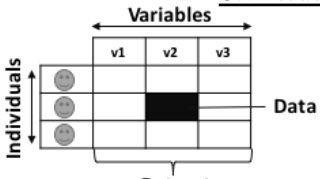
x **has a distribution with:**
mean of x : $\mu_x = \mu$
std dev of x : $\sigma_x = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem:
Normal distribution if:
- sample size is large enough (n≥30) OR
- variable varies normally (or "looks normal")

Sample ↳ Data

② Exploratory Data Analysis

ONE VARIABLE

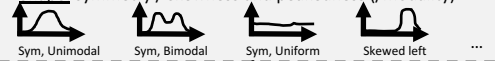


Variable: Categorical/Qualitative or Quantitative

Scale of measurement ↓ | can be transformed ↑
1. Nominal: no ranking, just differences
2. Ordinal: ordered differences
3. Interval: ordinal + equal intervals
4. Ratio: Interval + meaningful 0

Interpreting Histograms:

Shape: symmetry / skewness and peakedness (/modality)



Center:
- Mode: most frequently occurring value
- Mean: $x = \sum_{i=1}^{n} x_i/n$
- Median: M midpoint of the distribution

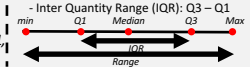
Reasonably symmetric distribution, no outliers: => Use x and s

Otherwise:
=> User 5-number summary *(min, Q1, M, Q3, Max)* with IQR and range

Spread:
- Standard Deviation:

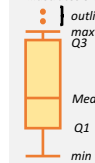
$s = \sqrt{variance} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}}$

- Range: Max – min
- Inter Quantity Range (IQR): Q3 – Q1



BoxPlot

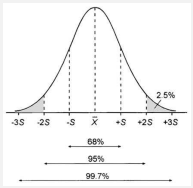
Illustrates 5-number summary:



1.5(IQR) criterion: outlier if
> Q3 + 1.5(IQR) OR <Q1 – 1.5(IQR)

Best used when comparing distributions of different groups

Standard Deviation Rule
(aka Empirical Rule)



TWO VARIABLES

RESPONSE (/dependent) VARIABLE		
EXPLANATORY (/independent) VARIABLE	CATEGORICAL	QUALITATIVE
CATEGORICAL	**C → C**	**C → Q**
QUALITATIVE	**Q → Q**	

C → C

	Y1	Y2	...	TOTAL
X1	%	%	%	100%
X2	%	%	%	100%
...	%	%	%	100%

Two-ways table Conditional %

or

Double Bar Chart

C → Q

Compare the distribution of the quantitative response for each category of the explanatory variable using side-by-side boxplots.

Q → Q

Linear Relationship Strength (/correlation coefficient)

$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - x)}{s_x}\frac{(y_i - y)}{s_y}$ r: -1 strong, 0 weak, 1 strong

Sensible to outliers inconsistent with linear direction

Linear Regression with method of the least square criteria

$Y = a + bX$
$b = r\,(S_y/S_x)$ *aka* ***regression coefficient***
$a = \bar{Y} - b\bar{X}$

Interpretation:
- Direction: positive, negative, neither
- Form: linear, curvilinear, cluster
- Outliers
- Strength (how close data points follow the form)

Scatter Plot

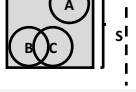
Probability

Probabilities can be determined either **theoretically** or **empirically** / observational (# event occurrence / # repetitions)

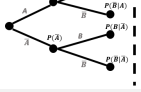
=> when we increase the # of repetitions, empirical and theoretical probabilities get closer.

Notation:
Random Experiment
▼
Sample Space S
▼
Events (A,B,C...)
▼
Probability of Events:P(A), P(B)...

Venn Diagram:



Probability Tree:



Basic Rules:
∀ event A, $0 \leq P(A) \leq 1$
$P(S) = 1$
$P(A) = 1 - P(A)$ *and* $P(A|B) = 1 - P(A|B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multiplication Rule: $P(A \cap B) = P(B|A) * P(A)$

Bayes Rule: $P(A|B) = \frac{P(A) * P(B|A)}{P(A) * P(B|A) + P(A) * P(B|A)}$

Independence Checks:
A and B independent if:
- P(B|A) = P(B) OR
- P(B|A) = P(B|A) OR
- $P(A \cap B) = P(A) * P(B)$

Random Variable: assigns a unique numerical value to the outcome of a random experiment

Discrete Random Variable

Table of outcomes:

OUTCOME	PROBA.	X
A	%	X_1
B	%	X_2
...	...	...

Proba. Distrib.

x	X_1	X_2	...
P(x=X)	%	%	...

$0 \leq P(x = X) \leq 1$
$\sum_x P(x = X) = 1$

$E(X) = \mu_X = \sum_{i=1}^{n} x_i * p_i$
$\sigma_X = \sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2 * p_i}$

$\mu_{a+bX} = a + b * \mu_X$
$\sigma^2_{a+bX} = b^2 * \sigma^2_X$
$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$

A value is **unusual** if ∉ $[\mu_X - 2\sigma_X; \mu_X + 2\sigma_X]$

Continuous Random Variable

$P(a \leq X \leq b) = \int_a^b f(x)dx$ with f density curve

Normal Distribution X_N: if follows standard deviation rule.
↳Standardized value with "Z-Score":
$z = \frac{x - \mu}{\sigma}$ to be used with normal tables

→ fixed number of n trials
→ each trial independent (/replacement)
→ just 2 outcomes success or fail
→ constant probability p
① Without replacement, can be used if n ≤ 0.10 N

Binomial R. Variable X_B

$P(x = X) = \binom{n}{x} p^x (1-p)^{n-x}$
$\mu_X = np$
$\sigma^2_X = np(1-p)$

Normal Approximation of the Binomial

if np ≥ 10 and n(1-p) ≥ 10 → X_B can be approximated by a X_n with same μ and σ

Continuity correction: $P(X_B \leq x) \simeq P(Z_N \leq x + 0.5)$
$P(X_B \geq x) \simeq P(Z_N \geq x - 0.5)$