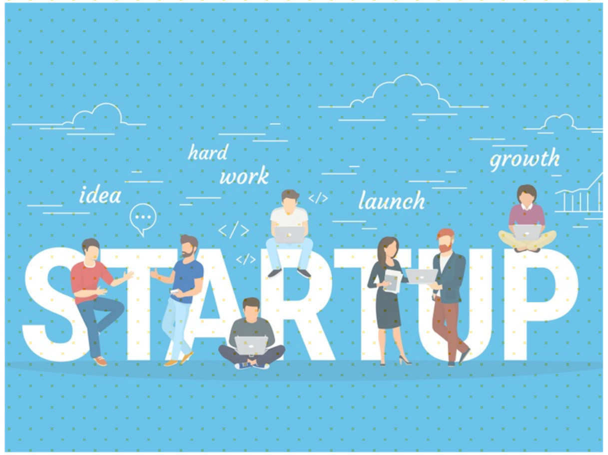
Introduction to Hypothesis Testing and its Application in Real-Life and Data Science

Fatemeh Asgarinejad



Announcements + schedule + links

- Lectures

  In person, SCHEDULE
- Discussions

  In person, SCHEDULE
- Office Hours

  SCHEDULES

COLAB

- Website| Piazza (Q&A) | Canvas (recordings + quizzes) | GradeScope (HW submission) | GitHub (code + slides) | Google Colab
- HWs, HW solutions, due dates and exams.

Course Plan: Where We Are Now

Week	Date	Topic	Activities
1	Monday – Jan 6	Combinatorics and Counting	HW 1 Programming 1 Exit Tickets 1-3
	Wednesday – Jan 8	Set Theory and Axioms of Statistics	
	Friday – Jan 10	Conditional Probability and Bayes' Rule	
2	Monday – Jan 13	Law of Total Probability, Independence, and Dependence	HW 2 Programming 2 Exit Tickets 4-6
	Wednesday – Jan 15	Conditional Independence, Chain Rule, Random Variables	
	Friday – Jan 17	Probability Mass Function, Cumulative Distribution, Expectation, and Variance	
3	Monday – Jan 20	Bernoulli and Binomial Distributions	Exam 1 Exit Tickets 7-9
	Wednesday – Jan 22	Geometric, Hypergeometric, and Poisson Distributions	
	Friday – Jan 24	Continuous Random Variables, Uniform Distribution, Normal Distribution	
4	Monday – Jan 27	Normal Distribution	HW 3 Programming 3 Exit Tickets 10-12
	Wednesday – Jan 29	Exponential Distribution and Its Memorylessness	
	Friday – Jan 31	Joint Distribution	
5	Monday – Feb 3	Conditional Joint Distribution, LOTUS	HW 4 Programming 4 Exit Tickets 13-15
	Wednesday – Feb 5	Covariance and Correlation	
	Friday – Feb 7	Independent Random Variables	

Week	Date	Topic	Activities
6	Monday – Feb 10	Central Limit Theorem	Exam 2 Exit Tickets 16-18 Project proposal
	Wednesday – Feb 12	Markov's Inequality, Chebyshev's Inequality, and Law of Large Numbers	
	Friday – Feb 14	Conditional Expectation, Linearity of Expectation, Inverse Transform Method	
7	Monday – Feb 17	Introduction to Statistics	HW 5 Programming 5 Exit Tickets 19-21
	Wednesday – Feb 19	Point Estimation and Maximum Likelihood Estimator	
	Friday – Feb 21	Interval Estimation, Confidence Interval	
8	Monday – Feb 24	Hypothesis Testing, z-Test, Fisher Test,p-Value	Project phase 1 Exit Tickets 22-24
	Wednesday – Feb 26	Project Phase 1 discussion	
	Friday – Feb 28	Chi-squared Distribution and Test, t-Distribution and t-Test, p-value cont.	
9	Monday – March 3	Nonparametric Tests, Permutation Test, p-Value Adjustment	HW 6 Programming 6 Exit Tickets 25-27
	Wednesday – March 5	Linear Regression	
	Friday – March 7	Bayesian Inference, Maximum a Posteriori Method, and Conjugate Priors	
10	Monday – March 10	Projects presentation and evaluation	Projects presentation and evaluation Final
	Wednesday – March 12	Projects presentation and evaluation	
	Friday – March 14	Final	

Today's Learning Goals

- Prerequisites: probability, distributions.
- Introduction to Hypothesis Testing (Using statistics to evaluate some hypothesis/statement about a population)
  - What is a Hypothesis?
  - Null and Alternative Hypotheses
  - Decision Making
    - Test Statistics
    - Accept or reject Null Hypothesis
    - P-value
  - Specific Statistical Tests
    - Z-test
    - Fisher Exact Test
  - Errors in Hypothesis Testing
    - Type I and Type II errors
  - More Applications
    - Hypothesis Testing in Data Science

Practical Use Cases of Statistics (Election)

Predicting which of the USA presidency candidates will win the election.
We ask a sample of population and estimate which candidate is more likely to win based on their opinions.

President Trump will earn more electoral votes.

Former Vise-President Harris's will earn more electoral votes.

Sample (poll here) can give us a good **estimate** about who is more probable to win the election.

images sources: first, second

Hypothesis

A **hypothesis** is an assumption (statement) about parameters of a population or distribution that can be True or Not True.

Note: parameter is a numerical or measurable characteristic of a population, e.g., $average\ (\mu), variance\ (\sigma^2), min, max, range, etc.$

Example: If we have a collection of individuals, hypothesize the max, average, etc. (e.g. average height of Americans is more than 5 feet 8 inches).

Example: We can have an assumption that the average GPA of first year students is > 3.0.

Example: We can hypothesize that men play more video games than women.

Example: We have a coin and our hypothesis is that it is unbiased ($P(Head) = \frac{1}{2}$).

Null and Alternative Hypothesis

We aim to evaluate the effectiveness of a medicine on **a group of patients (population)** using a randomized clinical trial. We give the **actual medicine** to half of the participants, while the other half receive **a placebo**.

6 people
took medicine

6 people
took placebo

- Medicine is not effective (has no effect on the outcome vs placebo) **Null Hypothesis (H_0)**
- Medicine is effective (there will be a difference between two populations). **Alternative Hypothesis (H_a or H_1)**

The assumption that is believed to be true until evidence is shown that it is not, is called **Null Hypothesis (H_0).**

The complementary (often) view is called **Alternative Hypothesis (H_a or H_1).**

Null and Alternative Hypothesis

Null Hypothesis ($\boldsymbol{H_0}$): The assumption that is believed to be true.
Alternative Hypothesis ($\boldsymbol{H_a}$ or $\boldsymbol{H_1}$): The complementary (often) view of Null hypothesis.


Discuss real-life scenarios of the null (H_0) and alternative hypotheses (H_1).

Example: **Null hypothesis ($\boldsymbol{H_0}$)** : Men play more video games than women.
Alternative hypothesis ($\boldsymbol{H_a}$) : Women play more video games than men.

Example: **Null Hypothesis ($\boldsymbol{H_0}$)** : People like Rock and pop music equally.
Alternative hypothesis ($\boldsymbol{H_a}$) : More than 50% of people prefer Rock music over Pop.

Example: **Null Hypothesis ($\boldsymbol{H_0}$)** : The coin used in a soccer game is unbiased $(P(Head) = \frac{1}{2})$.
Alternative hypothesis ($\boldsymbol{H_a}$) : The coin is biased $(P(Head) \neq \frac{1}{2})$.

Test Statistics

6 people
took medicine

6 people
took placebo

cured , # not cured

after 1 month

cured, # not cured

Which characteristic(s) should we study to measure the medicine's impact?

Test statistics is a function of sample data (mean, standard deviation, count, etc.) that helps us determine whether we should reject the Null Hypothesis or not.

Hypothesis Testing

6 people
took medicine

6 people
took placebo

4 cured , 2 not cured

after 1 month

1 cured, 5 not cured

Medicine is not effective. **Null Hypothesis (H_0)**

Medicine is effective (there will be a difference between two populations).
Alternative Hypothesis (H_a or H_1)

Hypothesis Testing: After collecting data, if the data is consistent with the null hypothesis, we **accept the null hypothesis (H_0)**.
Otherwise, if we have an strong evidence for the alternative hypothesis we reject the Null Hypothesis and **accept the Alternative Hypothesis**.

Hypothesis Testing

6 people
took medicine

4 cured , 2 not cured

6 people
took placebo

1 cured, 5 not cured

after 1 month

Should we allow the pharmaceutical company to release the medicine?

Is it possible to observe 4 cured , 2 not cured in the population who took placebo?
Or is the observed result statistically significant to reject the Null Hypothesis?

Should we set a threshold for rejecting the Null Hypothesis (H_0)?

P-value (probability value)

6 people
took medicine

6 people
took placebo

4 cured , 2 not cured

← **after 1 month** →

1 cured, 5 not cured

Let's show number of cured people with variable X.
We want to know how probable it is to observe 4 cured , 2 not cured or more extreme cases (5 or 6 cured) assuming H_0 was True: $P(X \geq 4 \,|H_0)$

Remember: Null Hypothesis (H_0): Medicine is not effective (there is no difference between two populations and any observed difference is due to chance or other factors rather than medicine)

Does the pharmaceutical company want $P(X \geq 4|H_0)$ be high or low?

Significance Level

6 people
took medicine

6 people
took placebo

4 cured , 2 not cured

after 1 month

1 cured, 5 not cured

Let's show number of cured people with variable X

We want to know how probable it is to observe 4 (or more) cured if there was no difference between the two populations (medicine was unimpactful) (H_0 was True):

$P(X \geq 4 \,|H_0)$

$$\boldsymbol{P(X \geq threshold \,|H_0) \leq \alpha}$$

significance level (α) is the probability of mistakenly rejecting H_0.

Note: significance level is often set to be **1%** or **5%.**

P-value

6 people
Took medicine

6 people
Took placebo

4 cured , 2 not cured ← **after 1 month** → 1 cured, 5 not cured

If $P(\boldsymbol{X \geq threshold} | \boldsymbol{H_0}) \leq \boldsymbol{\alpha}$, we can reject $Null\ Hypothesis\ (H_0)$ with $\mathbf{1 - \boldsymbol{\alpha}}$ **confidence**.

$P(X \geq x \mid H_0)$ is called p-value and is the probability of observing value x or more extreme values of x assuming H_0 holds.

Later in the slides: Fisher Exact Test

Testing the Hypothesis - P-value

Null Hypothesis ($\boldsymbol{H_0}$) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis ($\boldsymbol{H_a}$) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

Experiment: We flip the coin **20 times**

Test Statistic: Number of heads (we show it with variable X)

Intuitively we might say if $X = 9, 10 \; or \; 11$, we do not reject the **Null hypothesis ($\boldsymbol{H_0}$)** and we can conclude that the coin is not biased and otherwise we reject the $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_a}$

Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

Experiment: We flip the coin **20 times**

Test Statistic: Number of heads (we show it with variable X)

What **range** of X allows us to conclude that the coin is unbiased (accept H_0) with, e.g., 95% confidence?

Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

$P(X = 20 \mid H_0)$: If the coin was fair, how likely was it to observe 20 heads?

#Heads (X)	$P_{H_0}(X = x) = P(X = x \mid H_0)$
20	
19	
18	
17	
16	
15	
14	


Head Tail

Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

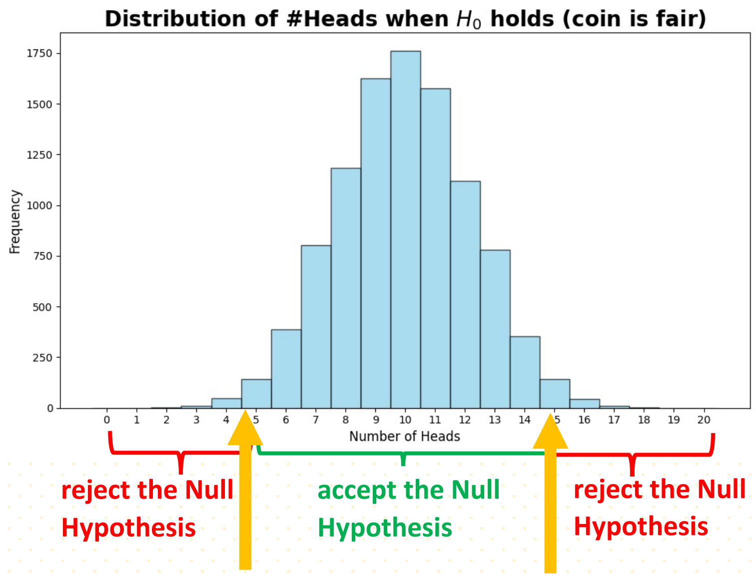
# Heads (X)	$P(X = x \mid H_0)$
20, 0	$\binom{20}{20} * \frac{1}{2^{20}} = 0.0001\%$
19, 1	$\binom{20}{19} * \frac{1}{2^{20}} = 0.0019\%$
18, 2	$\binom{20}{18} * \frac{1}{2^{20}} = 0.0181\%$
17, 3	$\binom{20}{17} * \frac{1}{2^{20}} = 0.1087\%$
16, 4	$\binom{20}{16} * \frac{1}{2^{20}} = 0.4621\%$
15, 5	$\binom{20}{15} * \frac{1}{2^{20}} = 1.4786\%$
14, 6	$\binom{20}{14} * \frac{1}{2^{20}} = 3.6964\%$

Testing the Hypothesis - P-value

Null Hypothesis ($\boldsymbol{H_0}$) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis ($\boldsymbol{H_a}$) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

# Heads (X)	$P(X = x \mid H_0)$
20, 0	$\binom{20}{20} * \frac{1}{2^{20}} = 0.0001\%$
19, 1	$\binom{20}{19} * \frac{1}{2^{20}} = 0.0019\%$
18, 2	$\binom{20}{18} * \frac{1}{2^{20}} = 0.0181\%$
17, 3	$\binom{20}{17} * \frac{1}{2^{20}} = 0.1087\%$
16, 4	$\binom{20}{16} * \frac{1}{2^{20}} = 0.4621\%$
15, 5	$\binom{20}{15} * \frac{1}{2^{20}} = 1.4786\%$
14, 6	$\binom{20}{14} * \frac{1}{2^{20}} = 3.6964\%$

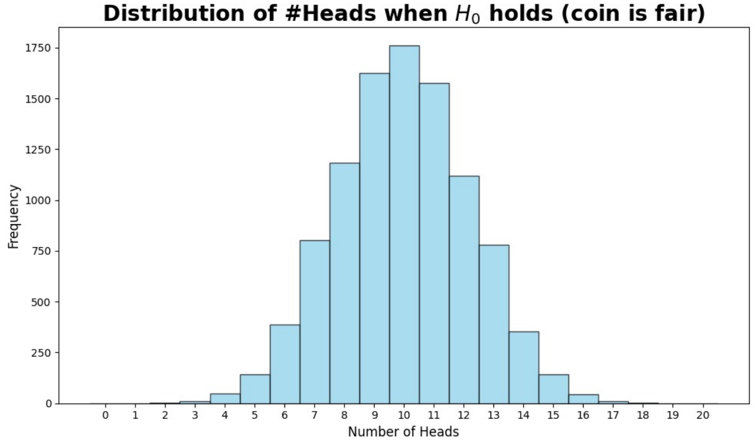


Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

# Heads (X)	$P(X = x \mid H_0)$	$P(X \geq x \text{ or } X \leq 20 - x \mid H_0)$
20	$\binom{20}{20} * \frac{1}{2^{20}} = 0.0001\%$	
19	$\binom{20}{19} * \frac{1}{2^{20}} = 0.0019\%$	
18	$\binom{20}{18} * \frac{1}{2^{20}} = 0.0181\%$	
17	$\binom{20}{17} * \frac{1}{2^{20}} = 0.1087\%$	
16	$\binom{20}{16} * \frac{1}{2^{20}} = 0.4621\%$	
15	$\binom{20}{15} * \frac{1}{2^{20}} = 1.4786\%$	
14	$\binom{20}{14} * \frac{1}{2^{20}} = 3.6964\%$	



Testing the Hypothesis - P-value

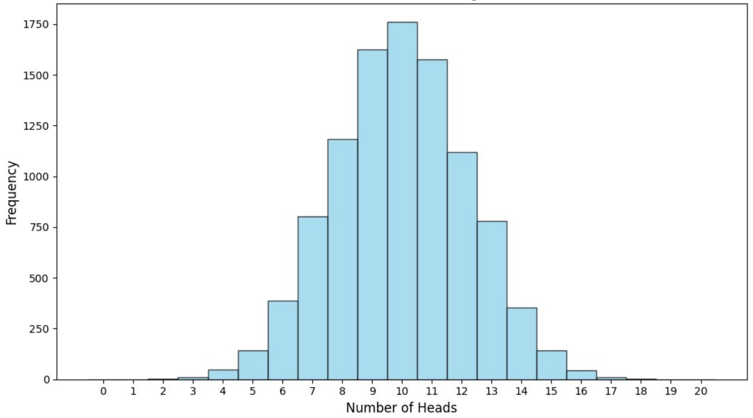
Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

# Heads (X)	$P(X = x \mid H_0)$	$P(X \geq x \text{ or } X \leq 20 - x \mid H_0)$
20	$\binom{20}{20} * \frac{1}{2^{20}} = 0.0001\%$	0.0001%
19	$\binom{20}{19} * \frac{1}{2^{20}} = 0.0019\%$	0.002%
18	$\binom{20}{18} * \frac{1}{2^{20}} = 0.0181\%$	0.0201%
17	$\binom{20}{17} * \frac{1}{2^{20}} = 0.1087\%$	0.1288%
16	$\binom{20}{16} * \frac{1}{2^{20}} = 0.4621\%$	0.5909%
15	$\binom{20}{15} * \frac{1}{2^{20}} = 1.4786\%$	2.0695%
14	$\binom{20}{14} * \frac{1}{2^{20}} = 3.6964\%$	5.7659%

Head Tail

Distribution of #Heads when H_0 holds (coin is fair)



Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

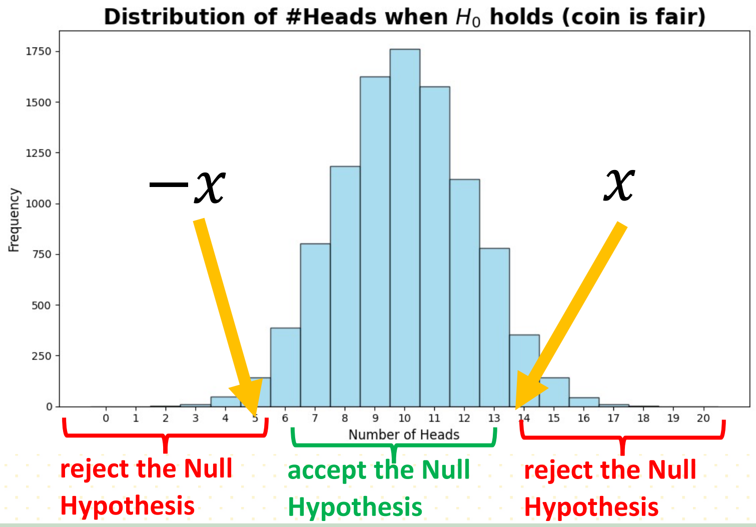
Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

#Heads (X)	$P(X = x \mid H_0)$	$P(X \geq x \text{ or } X \leq 20 - x \mid H_0)$
20	$\binom{20}{20} * \frac{1}{2^{20}} = 0.0001\%$	0.0001% ×2
19	$\binom{20}{19} * \frac{1}{2^{20}} = 0.0019\%$	0.002% ×2
18	$\binom{20}{18} * \frac{1}{2^{20}} = 0.0181\%$	0.0201% ×2
17	$\binom{20}{17} * \frac{1}{2^{20}} = 0.1087\%$	0.1288% ×2
16	$\binom{20}{16} * \frac{1}{2^{20}} = 0.4621\%$	0.5909% ×2
15	$\binom{20}{15} * \frac{1}{2^{20}} = 1.4786\%$	2.0695% ×2
14	$\binom{20}{14} * \frac{1}{2^{20}} = 3.6964\%$	5.7659%×2

$$P(|X| \geq x \mid H_0) \leq 5\%$$

What is the confidence interval for accepting the null hypothesis?

$$5 \leq x \leq 15$$



Testing the Hypothesis - P-value

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

#Heads (X)	$P(X \geq x \; or \; X \leq 20 - x \,|H_0)$
20, 0	0.0001% $\times 2$
19, 1	0.002% $\times 2$
18, 2	0.0201% $\times 2$
17, 3	0.1288% $\times 2$
16, 4	0.5909% $\times 2$
15, 5	2.0695% $\times 2$
14, 6	5.7659% $\times 2$

Note: If significance level is 5% and P-value is less than that, i.e., $P_{H_0}(||X| \geq x) < 5\%$, we can claim our results as significant results and can reject H_0 with 95% confidence.

Z-test (when the sample size is large enough)

Null Hypothesis (H_0) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis (H_a) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

$$\mathrm{Y_i} = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases} \qquad X = \sum_{i=1}^{n=100} Y_i$$

$$Y_i \sim Bernoulli\left(p = \frac{1}{2}\right) \quad \rightarrow \quad E(Y_i) = p = \frac{1}{2} \quad and \;\; Var(Y_i) = p(1-p) = \frac{1}{4}$$

$$E(X) = \mu = np = 100 * \frac{1}{2} = 50 \quad and \quad Var(X) = \sigma^2 = np(1-p) = 100 * \frac{1}{4} = 25$$

$$CLT: X \sim \text{Normal}(\mu = 50, \sigma = 5)$$

Now having a normal variable, we can use a statistics test called Z-test.

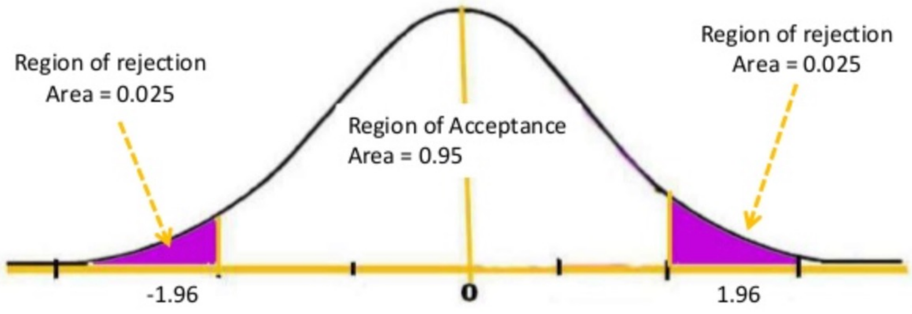
Z-test (when the sample size is large enough)

Null Hypothesis ($\boldsymbol{H_0}$) : The coin used in a soccer game is unbiased ($P(Head) = \frac{1}{2}$).

Alternative hypothesis ($\boldsymbol{H_a}$) : The coin is biased ($P(Head) \neq \frac{1}{2}$).

We first convert X to a standard normal variable $Z = \frac{X-\mu}{\sigma} = \frac{X-50}{5}$

$Z \sim Normal(0, 1)$



$P(|Z| < a) = 0.95$ (confidence level is 95%)

$a = |\Phi^{-1}(1 - 0.025)| = 1.96 \rightarrow$ $Confidence\ Interval$: $[-1.96, 1.96]$

$-1.96 < Z < 1.96$ and $Z = \frac{X-\mu}{\sigma} = \frac{X-50}{5}$. Hence, $-1.96 < \frac{X-50}{5} < 1.96 \rightarrow 40.2 < X < 59.8$

Note: $\Phi(z) = P(Z \leq z)$ and here $\Phi(1.96) = P(Z \leq 1.96) = 0.975$

Fisher’s Exact Test

6 people
Took medicine

3 cured , 3 not cured
50% of people got cured

after 1 month

6 people
Took placebo

2 cured, 4 not cured
33% of people got cured

If $P(X \geq threshold|H_0) \leq \alpha,$ we can reject H_0 with $1-\alpha$ confidence.

	cured	Not cured	Total
Took medicine	3 = a	3 = b	a + b
Took placebo	2 = c	4 = d	c + d
Total	a + c	b + d	n

$$P(observing\ this\ table) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{3+3}{3}\binom{2+4}{2}}{\binom{3+3+2+4}{3+2}} = 0.37$$

$$\boldsymbol{P(observing\ this\ table\ or\ more\ extreme\ results\ |H_0)}$$
$$= ???$$

Fisher's Exact Test

6 people
Took medicine

6 people
Took placebo

3 cured , 3 not cured

← **after 1 month** →

2 cured, 4 not cured

$P(\boldsymbol{observing\ this\ table\ or\ more\ extreme\ results}\ |H_0) \gg \mathbf{0.05}$ insufficient evidence to reject the null hypothesis

	cured	Not cured
Took medicine	**3 = a**	3 = b
Took placebo	2 = c	4 = d

	cured	Not cured
Took medicine	**5 = a**	1 = b
Took placebo	2 = c	4 = d

	cured	Not cured
Took medicine	**4 = a**	2 = b
Took placebo	2 = c	4 = d

	cured	Not cured
Took medicine	**6 = a**	0 = b
Took placebo	2 = c	4 = d

Fisher's Exact Test

6 people
Took medicine

5 cured , 1 not cured

← **after 1 month** →

6 people
Took placebo

0 cured, 6 not cured

If $P(X \geq threshold|H_0) \leq \alpha,$ we can reject H_0 with $1-\alpha$ confidence.

	cured	Not cured	Total
Took medicine	5 = a	1 = b	a + b
Took placebo	0 = c	6 = d	c + d
Total	a + c	b + d	n

$$P(observing\ this\ table\ or\ more\ extreme\ results) = \frac{\binom{5+1}{5}\binom{0+6}{0}}{\binom{12}{5+0}} + \frac{\binom{6+0}{6}\binom{0+6}{0}}{\binom{12}{6+0}} = 0.0152 < 0.05$$

sufficient evidence to reject the null hypothesis with $1 - 0.05 = 95\%$ confidence.

Type I and Type II Error

$$P(X \geq threshold \mid H_0) \leq \alpha$$

Note: significance level (α) is the maximum probability of making Type I error (mistakenly rejecting H_0) we allow to happen.

	Reality $= H_0$	Reality $= H_1$
Test result $= H_0$	Correct (TN)	Type II error FN
Test result $= H_1$	Type I error FP	Correct (TP)

False Positive

False Negative

	Reality $= H_0$	Reality $= H_1$
Test result $= H_0$	Correct (TN)	Type II error FN
Test result $= H_1$	Type I error FP	Correct (TP)

Fisher's Exact Test

6 people
Took medicine

3 cured , 3 not cured
50% of people got cured

← after 1 month →

6 people
Took placebo

2 cured, 4 not cured
33% of people got cured

Observation: $P(observing\ above\ or\ more\ extreme\ results\ |H_0) \gg\ 0.05$
Claim: Reject the Null Hypothesis

What type of error happens in above scenario if we reject the null hypothesis?

Note: To avoid Type I errors, we reject H0 **only** when $P - value \leq \alpha$.

A P-Value use-case in Data Science Feature Selection

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

from **50_Startups** dataset

Fit a multiple Linear Regression to the data:

$$\textbf{Profit} = \textbf{R\&D spend} * \mathrm{C}_1 + \textbf{Administraion} * C_2 + \textbf{Marketing Spend} * \mathrm{C}_3 + \textbf{State} * \mathrm{C}_4 + \beta \text{ (intercept)}$$

Which features are unlikely to contribute significantly to the **Profit** prediction?

A P-Value use-case in Data Science Feature Selection

Backward Elimination Steps:

1) Start with all features $f1, f2, f3, \ldots, fn$, run a multiple Linear Regression to get:

$$\mathbf{Profit} = \mathbf{R\&D\ spend} * \mathrm{C}_1 + \mathbf{Administraion} * C_2 + \mathbf{Marketing\ Spend} * \mathrm{C}_3 + \mathbf{State} * \mathrm{C}_4 + \beta\ (\text{intercept})$$

2) Identify the least significant feature (calculate features’ p-values and remove one with highest p-value)

3) Re-run the multiple linear regression with reduced number of features

e.g. if **State** is insignificant:

$$\mathbf{Profit} = \mathbf{R\&D\ spend} * \mathrm{C}_1 + \mathbf{Administraion} * C_2 + \mathbf{Marketing\ Spend} * \mathrm{C}_3 + \beta$$

4) Perform 2 and 3 iteratively and stop when all the remaining features are significant

How do we calculate P-value for each feature?

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

A P-Value use-case in Data Science Feature Selection

How do we calculate P-value for each feature?

Null Hypothesis: : The given feature has no effect in calculating the profit.
Alternative Hypothesis: The given feature has some effect in the target dependent variable (profit).

P(observing the given feature coefficient or more extreme values of it| H_0) $\leq \alpha$
If P-value $\leq \alpha$, then we can reject the null hypothesis.
If P-value $> \alpha$, then we can accept the null hypothesis and remove the feature.

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

COLAB

Exit Ticket – Due Saturday at 23:59 PM

SCAN ME

Resources

Fundamentals of Statistics and Probability, Sharif University of Technology, Professor Ali Sharifi Zarchi

Statistics and Data Science, UC San Diego, Professor Alon Orlitsky

When do we use statistics?

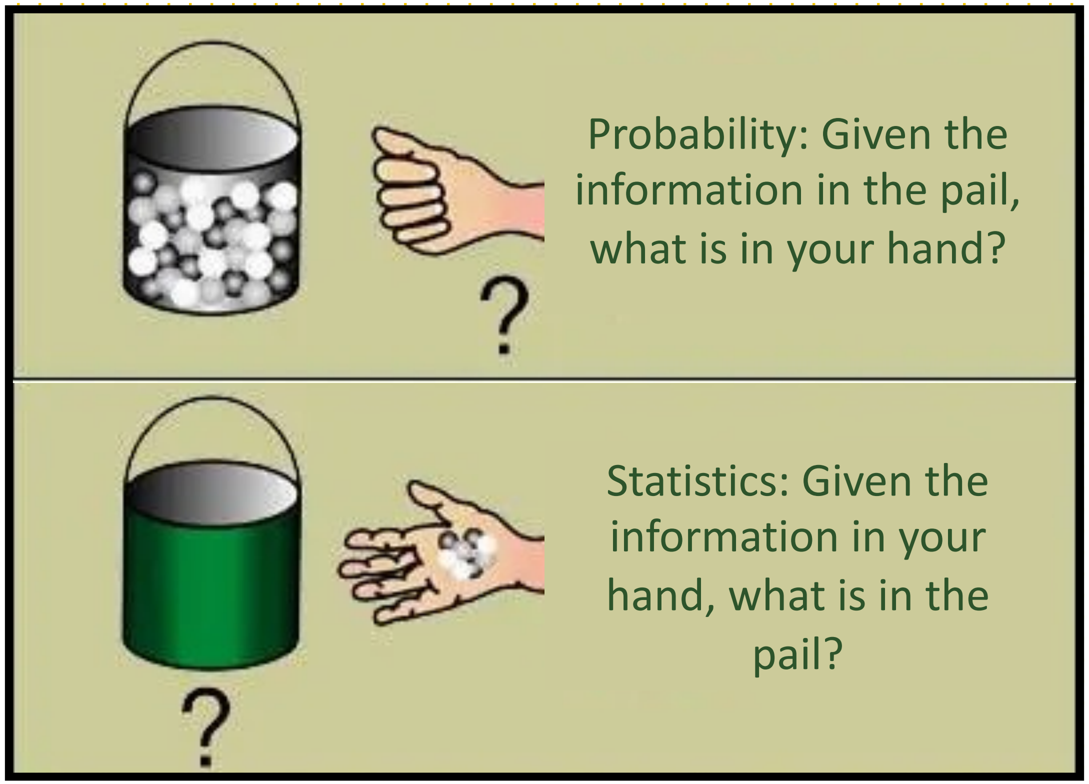


image source: https://teachingstatisticsandprobability.wordpress.com/

A P-Value use-case in Data Science Feature Selection

How do we calculate P-value for each feature?

Null Hypothesis: : The given feature has no effect in calculating the profit.
Alternative Hypothesis: The given feature has some effect in the target dependent variable (profit).

P(observing the given feature coefficient or more extreme values of it| H_0) $\leq \alpha$
If P-value $\leq \alpha$, then we can reject the null hypothesis.
If P-value $> \alpha$, then we can accept the null hypothesis and remove the feature.

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

COLAB

HW bonus question: Testing a hypothesis about a population

Choose a Null Hypothesis and prove/disprove it.

Example: election

Null Hypothesis (H_0) : Harris will win the election (will get more votes).

Alternative Hypothesis (H_a or H_1) : Trump will win the election.

Steps for testing a hypothesis:

1) Design an experiment (Test)
2) Define numerical outcomes (Test statistics to measure election outcomes) (e.g., the number of votes for each candidate, the ratio of votes (e.g. >51%))
3) Gather data (sample of population, please submit you hypothesis by tonight so we can gather data from your classmates)
4) Data consistent with null hypothesis or not?
   A. Calculate P(T=test statistics | H_0) $\leq$ significance value (often 5%, 1%, etc.)

      Yes. H_0 is inconsistent with data (reject Null)

      No. H_0 is consistent with data (accept Null)

HW bonus question example
Practical Use Cases of Hypothesis Testing (A/B Test)

We want to determine which version of our software performs better. We test a sample of users and estimate the preferred version based on their behavior or feedback before the final release.

Null Hypothesis (H_0) : Version A of the software performs better than Version B.
Alternative Hypothesis (H_a or H_1) : Version B of the software performs better than Version A