Day 6

26 May 2024 18:50

$y = mx + c$ (general)

$m_1x_1 + m_2x_2 + \cdots + c$

$y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_0$

$\boxed{y \simeq X \text{ variables}}$

function $\{b_0, b_1, b_2, b_3 \cdots\}$

objective:- find $b_0, b_1, b_2 \cdots$ in such a way that $\sum$ error (residual) is minimum.

$m = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$ ← (1) statistical method. (closed form of solution) linear data small size data.

(2) open form soln. ml method. (gradient based method)

Neural nets, deep learning → ml method

$y = b_0 + b_1x_1$.

ensure error is minimum.

adv X_1	y (salry).
1000	10
5000	15
3000	12
⋮	⋮

→ statistical approach. → b_0, b_1.

$b_0 + b_1x_1$.

x	y	y (pred).	residual	residual^2
–	10	9	1	81
–	15	6	9	16
–	14	8	4	64
–	19	11	8	⋮
	23	18	5	
	⋮	⋮	⋮	

$\sum$ error $= \frac{1}{n}\sum = SSE$ $= 70$

R^2 (how good is the model)

→ $b_0, b_1, \cdots$

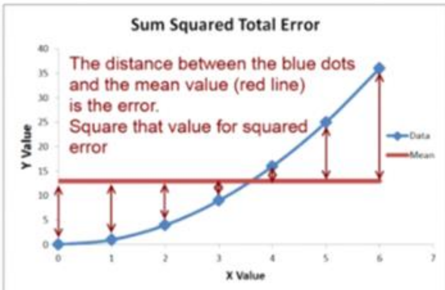
→ error (min)

→ ... model

→ $b_0, b_1, \ldots$

→ error (min)

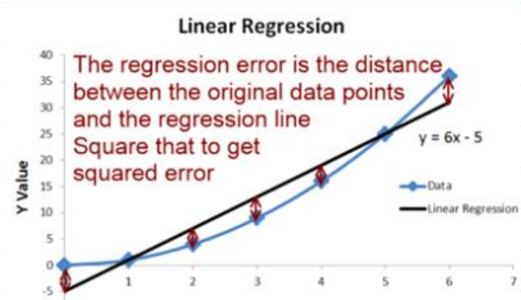
→ Quantify the model performance.



no regression model w/ a ~~eq.~~ ~~del~~.

X	Y	$\hat{Y}$	residual
	10		10
	15	20	5
	25		5
	30		10

$SSE = 100$ (SST).



← with reg model. $b_0, b_1, \ldots$

X	Y	$\hat{Y}$	residual
	10	12	–
	15	13	–
	25	22	–
	20	18	–

$\sum = SSE = 80.$

$$\frac{SSE}{SST} = \frac{80}{100} = 0.8$$

80% of error which existed with the avg model ↳ still exist.

$$1 - \frac{SSE}{SST} = 1 - 0.8 = 0.2$$

↑ $R2$

$R2 = 0.2$

... remained 20% of

$R2 = 0.2$

reg model remand 20% of errors with the avg model.

(0 — 1) ideally.

0 .5 .7 → 1.

Probability

Medical: Disease Diagnosis Probability	
Example: Calculating the probability of a patient having a certain disease given their symptoms and test results.	**Method**: Bayes' theorem is commonly used in medical diagnosis to calculate probabilities. This theorem incorporates prior probabilities (such as the prevalence of the disease in the population) and conditional probabilities (such as the likelihood of certain symptoms given the presence or absence of the disease) to determine the probability of the disease given observed evidence.

```
Index(['mean radius', 'mean texture', 'mean perimeter', 'mean area',
       'mean smoothness', 'mean compactness', 'mean concavity',
       'mean concave points', 'mean symmetry', 'mean fractal dimension',
       'radius error', 'texture error', 'perimeter error', 'area error',
       'smoothness error', 'compactness error', 'concavity error',
       'concave points error', 'symmetry error', 'fractal dimension error',
       'worst radius', 'worst texture', 'worst perimeter', 'worst area',
       'worst smoothness', 'worst compactness', 'worst concavity',
       'worst concave points', 'worst symmetry', 'worst fractal dimension',
       'target'],
```

x_1	x_2	x_3	- - - -	x_{30}	target
					0
					0
50K					0

$x' \rightarrow$ ML model $\rightarrow y = \{0, 1\}$

.55 .45

Example: Predicting the probability of a stock's price increasing or decreasing within a certain time frame.

Method: Statistical models such as time series analysis, Monte Carlo simulation, or option pricing models (e.g., Black-Scholes model) can be used to calculate probabilities in finance.

X'

cat.

ML

Trained 1000 differ types of image.

dog
cat
horse
car
:
:

$\hat{Y}$ = 1000 different nos. (prob).

[.01, 0.001, 0.7, .95 ......]

cat

next word prediction (DL).

new	delhi
new	york
new	delhi
new	york
new	mexico
–	–
–	–

statistical model.

text

new → delhi, york, mexico.

current word

$$p(delhi) = \frac{\#(new + delhi)}{\# new} = \frac{77}{100} = prob$$

$$p(york) = \frac{\#(new + york)}{\# new} = \frac{23}{100} =$$

$$p(mexico) = \frac{\#(new + mexico)}{\# new} = \frac{11}{100}$$

Conditional Probability

- Fundamental from DS/ML angle also
- Prob of some event occurring GIVEN that another event has taken place P(A|B)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Color	Shape	Ripe
Red	Round	Yes
Green	Round	No
Yellow	Oval	Yes
Red	Oval	No
Green	Round	Yes
Yellow	Oval	No

- **Event A:** Fruit is red.
- **Event B:** Fruit is round.
- **Event C:** Fruit is ripe.

- calculate $P(C \cap A \cap B)$, the joint probability of a fruit being ripe, red, and round.

- $P(C \cap A \cap B) = \frac{\text{Number of occurrences of ripe, red, and round fruits}}{\text{Total number of fruits}}$
- $P(C \cap A \cap B) = \frac{1}{6}$

Sample	Class	Long	Sweet
1	Orange	Yes	Yes
2	Orange	Yes	Yes
3	Banana	Yes	No
4	Orange	No	Yes
5	Orange	Yes	Yes
6	Banana	Yes	Yes
7	Banana	No	Yes
8	Banana	No	No
9	Orange	No	Yes
10	Banana	No	Yes
11	Banana	Yes	No
12	Orange	Yes	Yes
13	Banana	Yes	No
14	Banana	No	No
15	Orange	No	No
16	Orange	Yes	Yes
17	Orange	Yes	Yes
18	Orange	No	Yes
19	Banana	No	No
20	Orange	No	Yes

P(Banana) =
(Count(Banana) + 1) / (Total Samples + Number of Classes)
= (10 + 1) / (20 + 2) = 11/22

P(Orange) =
(Count(Orange) + 1) / (Total Samples + Number of Classes)
= (10 + 1) / (20 + 2) = 11/22

Prior probabilities of features - For 'Long'

P(Long=Yes | Banana) = (Count(Long=Yes, Banana) + 1) / (Count(Banana) + 2) = (6 + 1) / (10 + 2)
= 7/12

P(Long=No | Banana) = (Count(Long=No, Banana) + 1) / (Count(Banana) + 2) = (4 + 1) / (10 + 2)
= 5/12 (**)

P(Long=Yes | Orange) = (Count(Long=Yes, Orange) + 1) / (Count(Orange) + 2) = (4 + 1) / (10 + 2)
= 5/12

P(Long=No | Orange) = (Count(Long=No, Orange) + 1) / (Count(Orange) + 2) = (6 + 1) / (10 + 2)
= 7/12 (**)

Prior probabilities of features - For 'Sweet'

P(Sweet=Yes | Banana) = (Count(Sweet=Yes, Banana) + 1) / (Count(Banana) + 2) = (8 + 1) / (10 + 2) = 9/12 (**)

P(Sweet=Yes | Orange) = (Count(Sweet=Yes, Orange) + 1) / (Count(Orange) + 2) = (7 + 1) / (10 + 2) = 8/12 (**)

P(Sweet=No | Banana) = (Count(Sweet=No, Banana) + 1) / (Count(Banana) + 2) = (2 + 1) / (10 + 2) = 3/12

P(Sweet=No | Orange) = (Count(Sweet=No, Orange) + 1) / (Count(Orange) + 2) = (3 + 1) / (10 + 2) = 4/12

Take a test fruit , Long = 'No', Sweet = 'Yes'
Calculate the posterior probabilities for each class:

- For Banana:
  - P(Banana | Long=No, Sweet=Yes)
  - = P(Long=No | Banana) * P(Sweet=Yes |Banana) * P(Banana)
  - = (5/12) * (9/12) * (11/22) = 0.142
- For Orange:
  - P(Orange | Long=No, Sweet=Yes)
  - = P(Long=No | Orange) * P(Sweet=Yes | Orange) * P(Orange)
  - = (7/12) * (8/12) * (11/22) = 0.212

Conditioning ??

So, the test fruit is more likely to be classified as an Orange.

$P(F|E) = \frac{P(F \cap E)}{P(E)}$

Where:

- $P(F \cap E)$ is the joint probability of both events F and E.
- $P(E)$ is the probability of event E.

Bayes theorem

(TBD)

Inferencing
- Making some statement or predictions about a population, based on sample data

Methods for inferring
- Hypothesis testing
- Regression analysis

Null Hypothesis (*H0*):

- statement that there is no significant difference or effect. It often includes an equal sign (=).
- Example: *H0*: μ=50 (population mean is equal to 50).

Alternative Hypothesis (*Ha*):

- statement that contradicts the null hypothesis, indicating a significant difference or effect.
- Example: *H1*: *μ not equal* 50 (population mean is not equal to 50).

Examples – Market research

- Suppose we have a sample of 100 consumers who participated in the market research study.
- categorize their age groups into three categories: Young (18-30), Middle-aged (31-50), and Older (51 and above).
- For screen size preference, we'll consider two categories: Small and Large.

- **TASK** : the relationship between screen size preference and age group

Consumer ID	Age Group	Screen Size Preference
1	Young	Large
2	Middle-aged	Small
3	Older	Large
4	Young	Large
5	Middle-aged	Small
6	Middle-aged	Large
7	Older	Small
8	Young	Large
9	Young	Large
10	Older	Small
...	...	...
100	Middle-aged	Large

Example - Employee Productivity

- A business leader wants to determine if there is a significant difference in productivity between two teams within the organization.
- use a hypothesis test to compare the average productivity scores of Team A and Team B based on specific metrics (e.g., sales volume, project completion time).
- test would help the leader infer whether the observed difference in productivity between the two teams is statistically significant.

```
Team A Sales Volume: [100, 120, 110, 90,  105, 115,  95, 105, 115, 100]

Team B Sales Volume: [110, 105, 115, 100, 125, 115, 120, 110, 115, 105]
```

```
The customer satisfaction scores before training are represented by the list
[80, 75, 85, 70, 75, 78, 82, 79, 80, 77].

The customer satisfaction scores after training are represented by the list
[85, 82, 88, 75, 80, 84, 87, 86, 85, 82].
```

Example - Training Effectiveness

- A business leader invests in a training program for customer service representatives
- To assess the effectiveness of the training program, the leader could conduct a hypothesis test by comparing the customer satisfaction scores before and after the training.
- test would help infer whether there is a significant improvement in customer satisfaction as a result of the training program

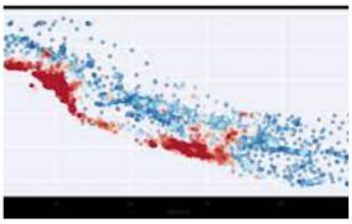
Example: Medical Treatment

- **Null Hypothesis (*H0*):** The new drug has no effect on patient recovery.
- **Alternative Hypothesis (*H1*):** The new drug improves patient recovery.
- **Interpretation:** Evaluating whether a medical intervention has a significant impact on patient outcomes.

Example: Educational Intervention

- **Null Hypothesis (*H0*):** The teaching method has no impact on student performance.
- **Alternative Hypothesis (*H1*):** The teaching method improves student performance.
- **Interpretation:** Investigating the effectiveness of a particular teaching approach on student learning.

Example



- **MedInc (Median Income)**: Median income of households within a district.
- **HouseAge (Housing Age)**: Median age of houses within a district.
- **AveRooms (Average Rooms)**: Average number of rooms per household within a district.
- **AveBedrms (Average Bedrooms)**: Average number of bedrooms per household within a district.
- **Population**: Total population of the district.
- **AveOccup (Average Occupancy)**: Average household occupancy within a district.
- **Latitude**: Latitude coordinate of the district's location.
- **Longitude**: Longitude coordinate of the district's location.
- **MedHouseVal (Median House Value)**: target/y Median house value for households within a district.

Example – test

01

Hypothesis Test -

Age of Houses and House Value:

02

Null Hypothesis (H0):

There is no significant correlation between the age of houses and median house values in California districts.

03

Alternative Hypothesis (H1): Older houses have significantly lower median values compared to newer houses.

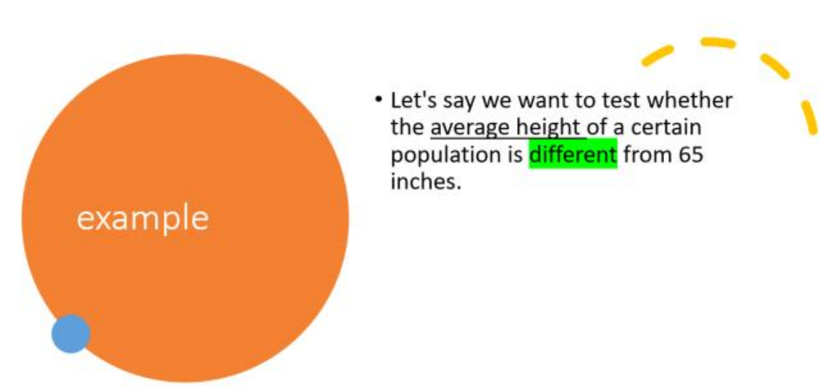
Process of Hypothesis testing

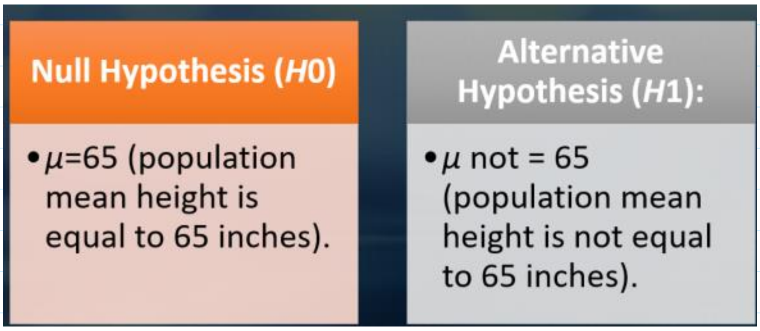
1. Objective setting
   a. Feature
      i. Independent
      ii. Dependent
2. Choose the statistical method
   a. t-test
   b. ANOVA
   c. Chi2
   d. ...
   e. ...

Test statistic
- Number/
- Output from the statistical fn (chosen)
- Used for concluding

p-value
- Number, prob
- Output
- Used for concluding

Critical value
- Threshold
  - Based on the level of significance you have chosen for your study

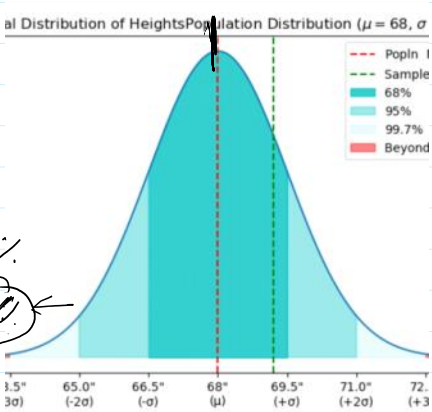




e-commerce conversion rate
- H0 - no change due to changes in website design

Churn analysis
- Introduced some loyalty program
- H0 : ...

Set the significance level
- A probability
- Threshold
- Beyond the threshold
  - Reject the H0



$\sigma = 1$

$\mu = 65$

$\pm 1\sigma = [64 - 66] \rightarrow 68\%$

$\pm 2\sigma = [63 - 67] \rightarrow 95\%$

$\pm 3\sigma = [62 - 68] \rightarrow 99\%$

Known to you. "68 - 95 - 99.7"

68% 1σ

95% 2σ

99.7 3σ

height = 70 inches ?

outlier Common (Something wrong with H_0).

$\alpha = 95\% \mid 99.5\%$

setting aside some room for wrong prediction.

how much?

Carry out the test

- We collect a sample of 100 people
- Measure heights
- Choose/execute statistical function
- obtain the test statistic (e.g., t-statistic), and
- obtain a p-value of 0.03.

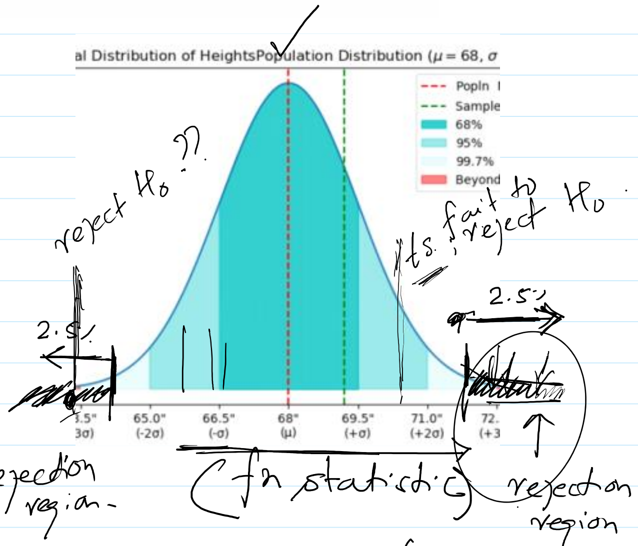
conclude

- **P-value Interpretation:**
  - If p-value < 0.05 (or any chosen significance level), we reject the null hypothesis.
  - In our example, p-value = 0.03 < 0.05, so we reject *H0*.
- **Test Statistic and Critical Value:**
  - We compare the test statistic to the critical value for our chosen significance level.
  - If the test statistic falls beyond the critical value, we reject *H0*.

$\alpha = 95\%$

$1 - \alpha = 5\%$

After the test,
- test statistic
- p-value



Conclude:

① if t_statistic > statistic for sig level
Critical value.

reject H_0.

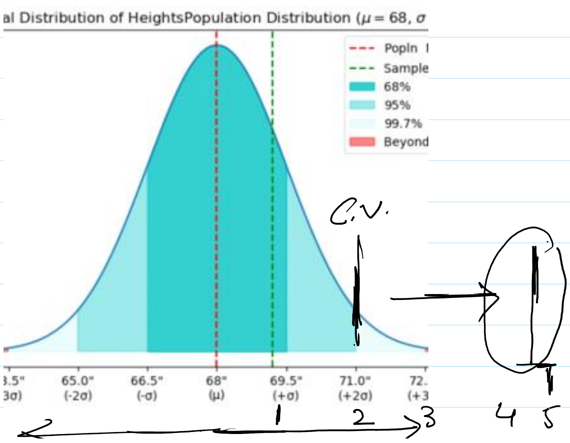
② if p-value < .05 95% (5%)

70

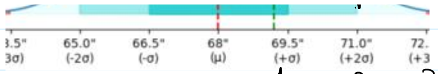
$$Z = \frac{70 - 65}{1} = 5$$

$\alpha = 95\%$ (5%)

critical value.



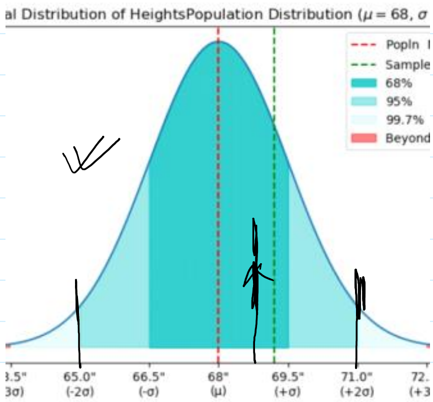
Critical value.



1 2 3 4 5

z-score

$h_s = 66$
likely to happen

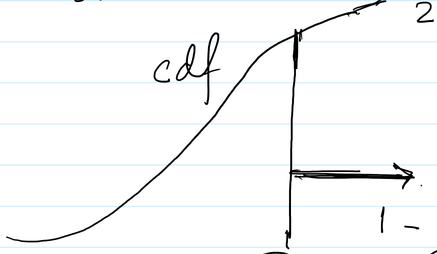


2.5% 2.5%

pdf
cdf.

$$\boxed{\begin{array}{l} h_s = 80'' \\ z = 7.6 \end{array}}$$

cdf

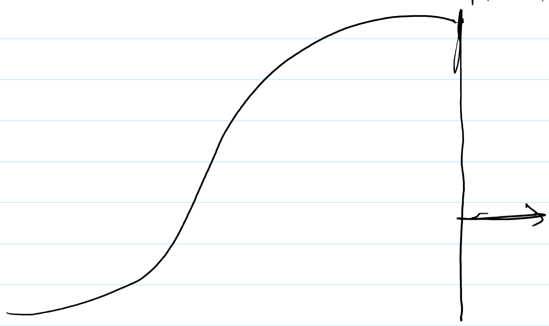


1 − 60%

66 (60%) → not extreme (common)

40% = .4 > .05

99.999%



100 − 99.999...

= .000001

< .05

$h_s = 80''$

t-test
- Check if there are differences in the data
  - 2 columns
    support team (CITES / ....)

Complaint types	time to resolve
p1	4
p1	6
	11
p1	9
p2	100
p2	70
	60
	2

$\cdot\ \mu_{p1}$

$\cdot\ \mu_{p2}$

Same / similar.

$\mu_1 = \mu_2$

p_2
p_2

60
2
150

μ_2

Suppose we want to test whether there is a significant difference in the average scores between two groups of students

who followed different study methods:

- Group A (traditional study) and
- Group B (online study).

Hypothesis

H0: $\mu A = \mu B$
(No significant difference)

H1: μA Not $= \mu B$
(Significant difference)

test

Data:	• Group A: [78, 85, 92, 88, 76] • Group B: [82, 90, 88, 79, 84]
Calculate T-Statistic:	• Use the statistical method to calculate the t-statistic.
Degrees of Freedom:	• For an independent samples t-test, degrees of freedom $= n1 + n2 - 2$.
P-Value:	• Use the t-statistic and degrees of freedom to find the p-value.

Calculate T-Statistic

the formula for calculating the t-statistic in an independent samples t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means of Group A and Group B, respectively.
- s_1^2 and s_2^2 are the sample variances of Group A and Group B, respectively.
- n_1 and n_2 are the sample sizes of Group A and Group B, respectively.

$$f(t;df) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\pi}\Gamma\left(\frac{df}{2}\right)}\left(1+\frac{t^2}{df}\right)^{-\frac{df+1}{2}}$$

where:

- t is the value of the random variable (t-statistic).
- df is the degrees of freedom.
- Γ denotes the gamma function.