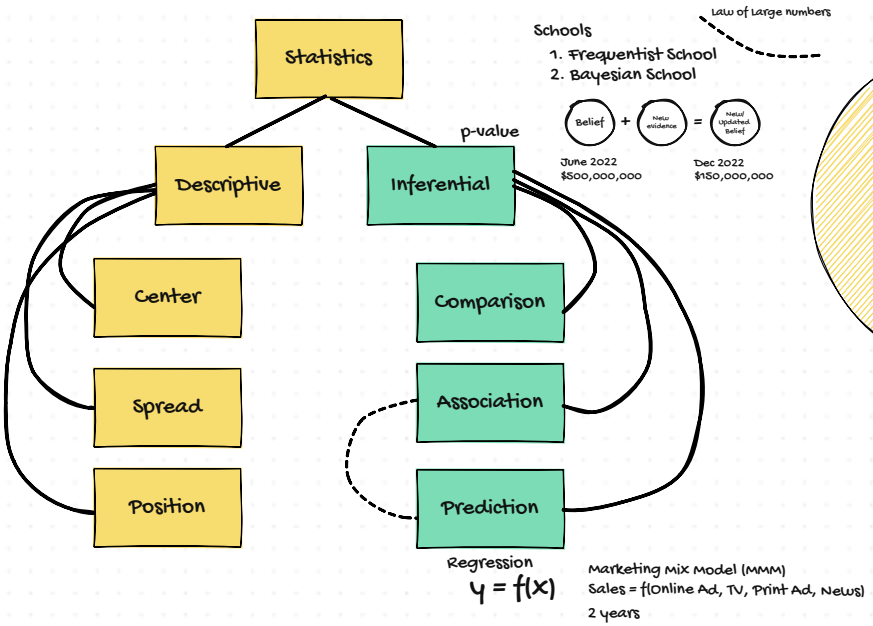
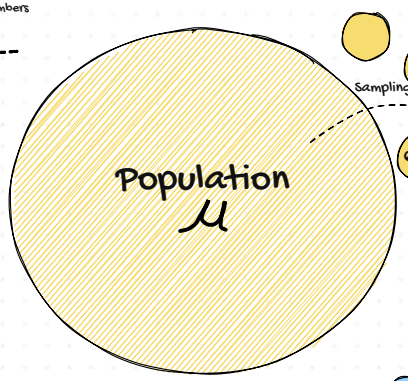
Intro to Statistics for Data Analyst





Precision
n=9000 70% Margin Error =1%
[69%, 71%]

Budget + Time

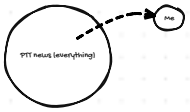
65% Sample

66% 68% 72%

Types of Sampling
1. Probability
- Random Sampling
- Systematic Random Sampling
- Stratified Sampling
- Cluster Sampling
2. Non-Probability
- Convenient Sampling
- Snowball sampling

1 → 2 → 3

Real Life Example



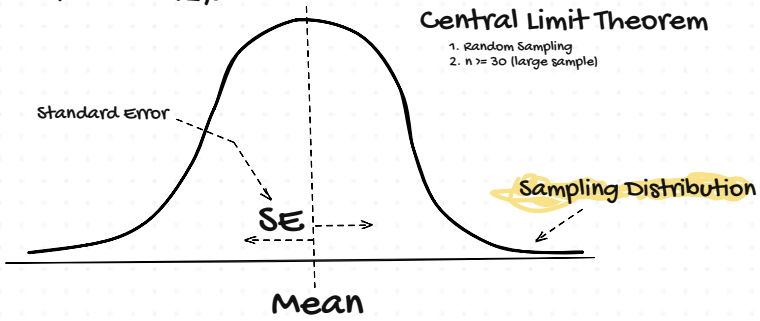
Sampling Variation

1 2 10000

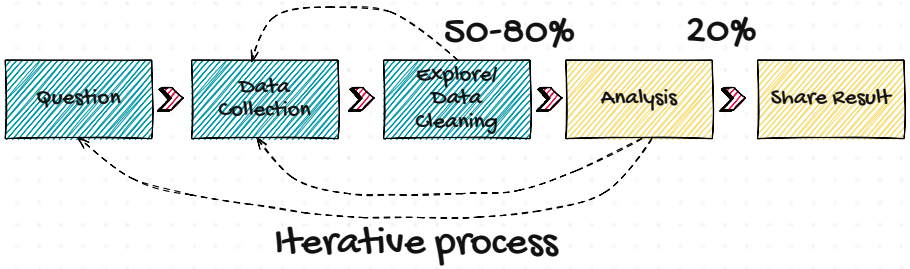
65% 68% 72%

Central Limit Theorem

1. Random Sampling
2. n >= 30 (large sample)



Statistics Process



Center
Central Tendency

- Mean
- Median
- Mode

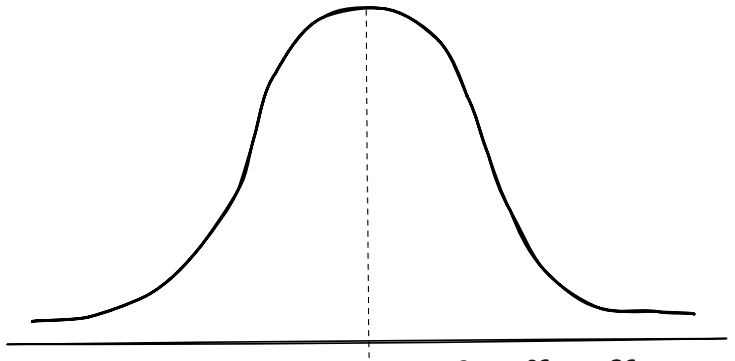
Spread
Variation

- Range
- SD
- Variance
- IQR

Position
How to find the data

- Min
- Max
- Percentile
- Quartile

Normal Distribution
Gaussian / bell shaped curve



Empirical Rule

68% +/-1 SD

95% +/-2 SD

99% +/-3 SD

Agenda

Morning - Concepts
Break 1 hour
Afternoon - Sheets/ R

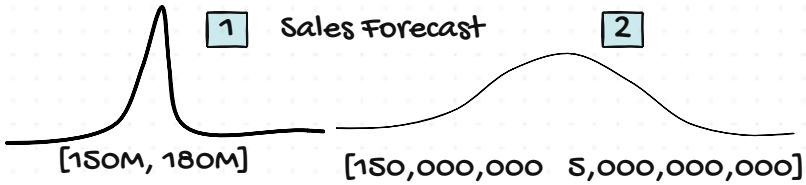
Spread

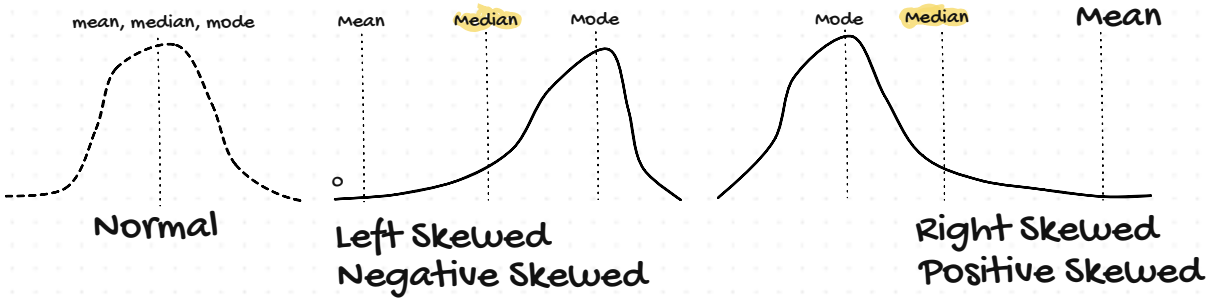
Variability/ Variation

- Range (max- min)
- Variance
- Standard Deviation
- IQR (Interquartile Range)

20, 25, 50, 100, 200

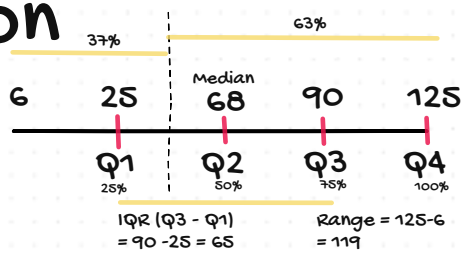
Range = 200-20 = 180
var = sum((x-x_bar)^2)) / (n-1)
sd = SQRT(sum((x-x_bar)^2)) / (n-1))





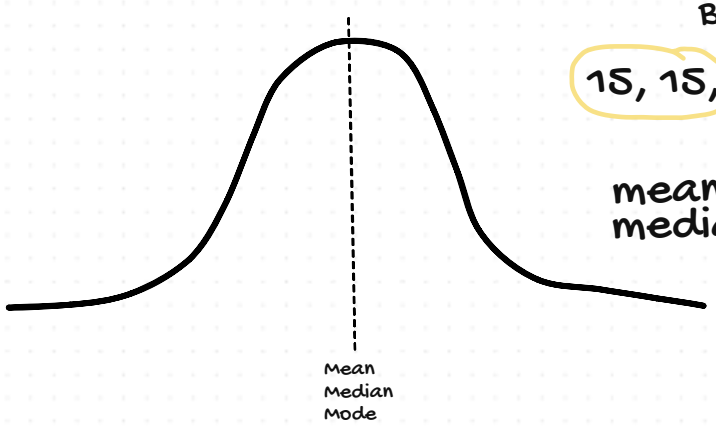
Position

Min
Max
Percentile
Quartile



Median < Mean

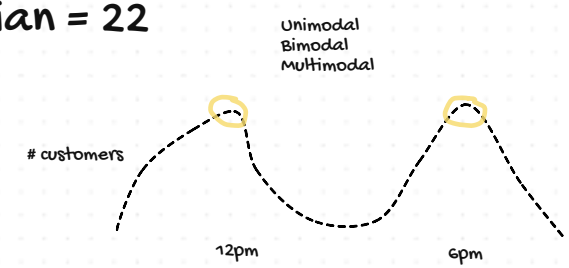
Central Tendency



Bimodal

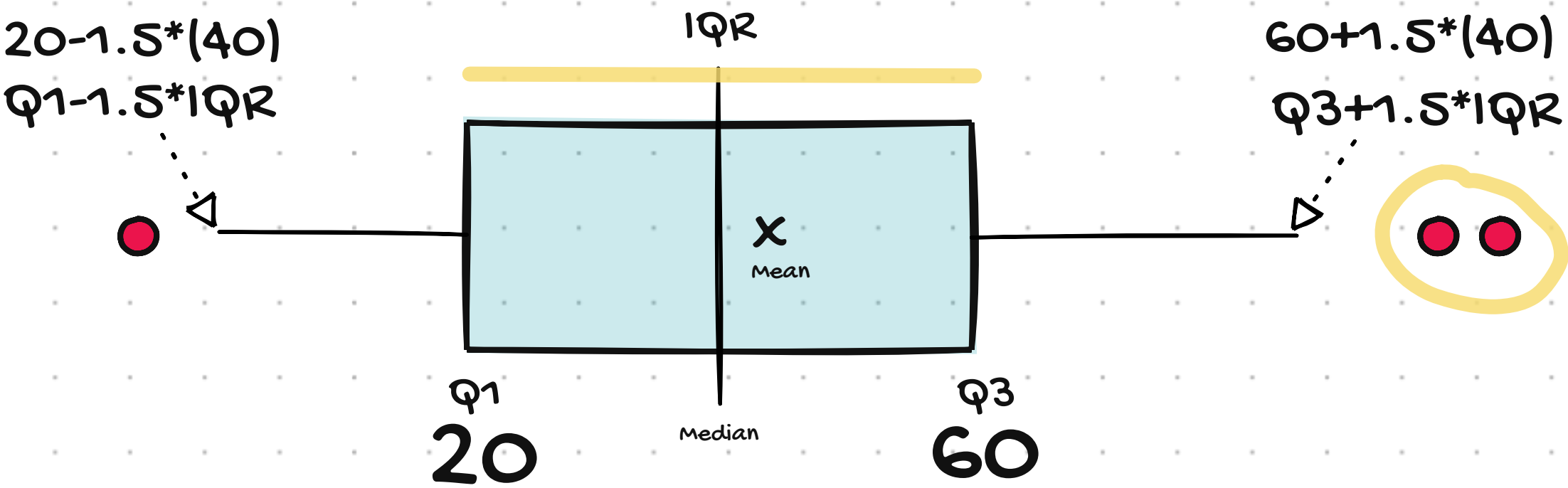
15, 15, 22, 25, 25, 50, 100

mean = 26.4
median = 22



Outliers (extreme values)

Boxplot + 5 numbers summary

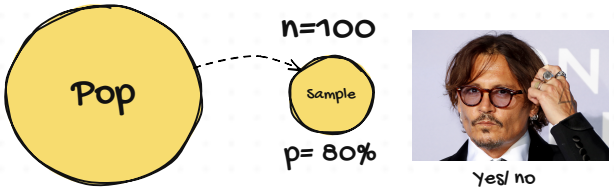


Confidence Interval (CLT)

Results:
80% [72%, 88%] CI 95%

Steps to Apply CLT

1. collect data
2. SE
3. ME = SE*1.96
4. Confidence Interval



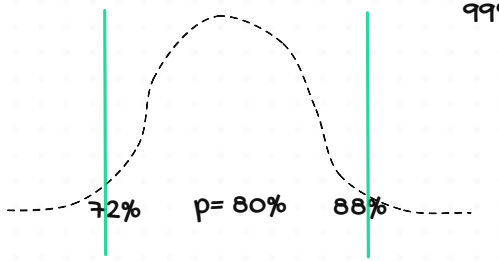
Yes/ no

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

$$SE = \sqrt{p * (1-p) / n}$$

z
90% 1.645
95% 1.96
99% 2.58



True Definition CI

Confidence 95%

19 / 20 = 0.95 ~ 95%
1 / 20 = 0.05 ~ 5%

SE for mean vs. SE for proportion

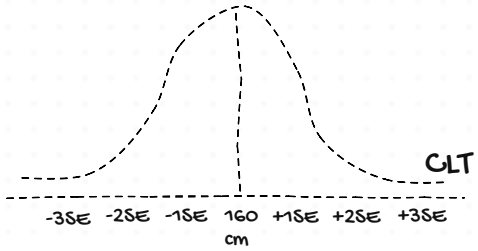
SE m = SD / SQRT(N)

SE p = SQRT(p*(1-p) / n)

n = 100
mean = 160 cm
sd = 5 cm

Our sample

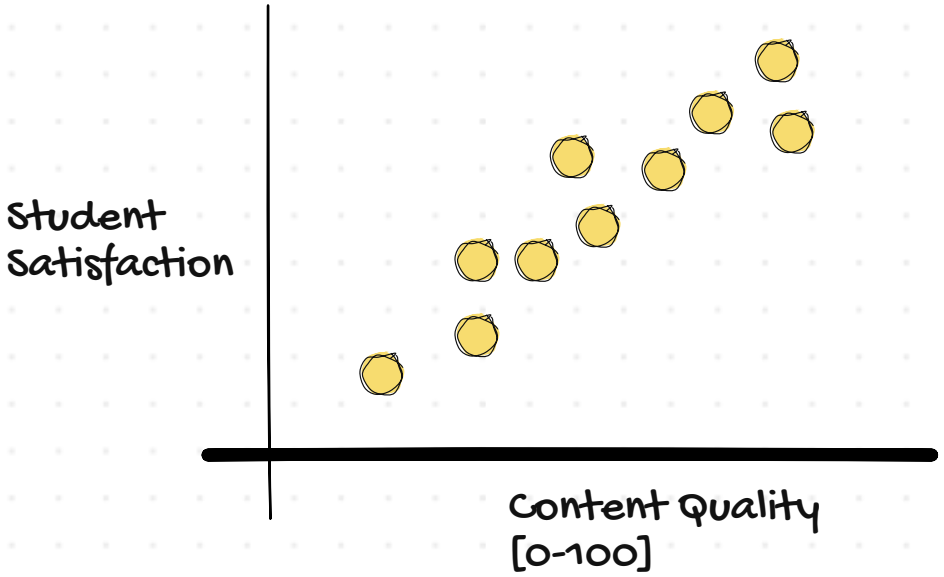
Height Thai Pop



1. SE = SD/ SQRT(N)
   SE = 5 / SQRT(100) = 5/10 = 0.5
2. Margin Error = SE * 2 = 0.5 * 2 = 1
3. Confidence Interval = [160-1, 160 + 1]

Correlation

Linear association



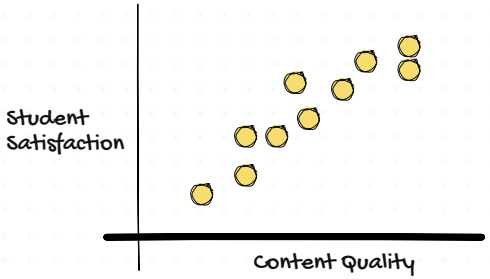
Correlation (r)

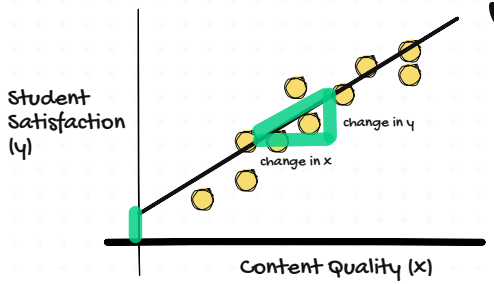
[-1 , +1]
+ positive
- negative
0 no relationship

=CORREL(x, y)

strong r > 0.7
moderate r >= 0.4
weak r < 0.4

Linear Regression





$y = mx+c$

m
= slope
= change in y / change in x
= 0.8 / 1
= 0.8

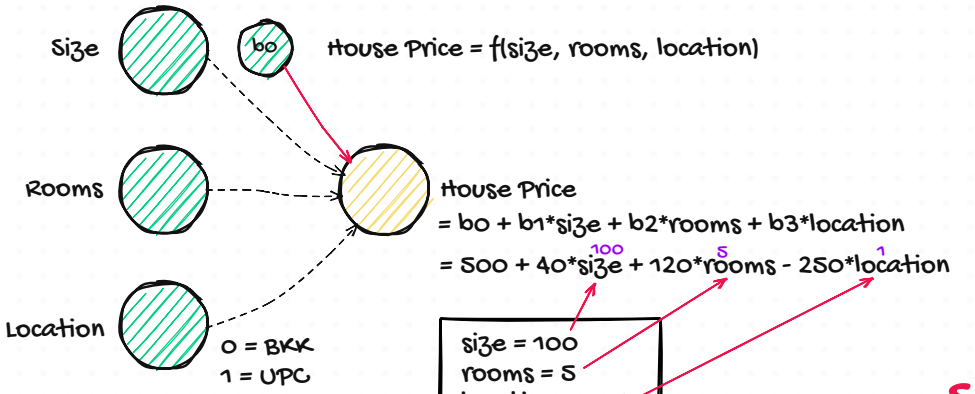
$y = f(x)$

$y = b_0 + b_1 * x$

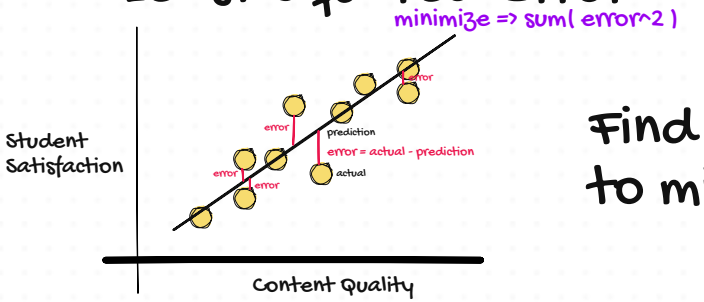
y intercept

slope

Coefficients
best fitted line



Least Squared Error



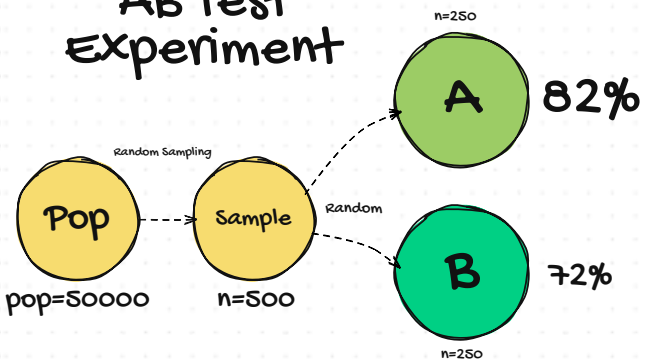
Find bo, b1
to minimize sum(error^2)

Significance Test

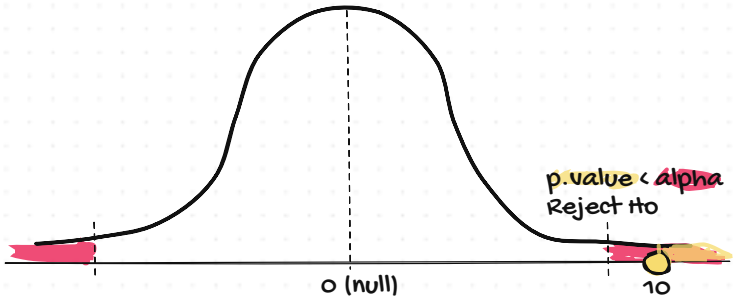
x is significance if p.value < 0.05

Break 10 min (2.51pm)

AB Test Experiment



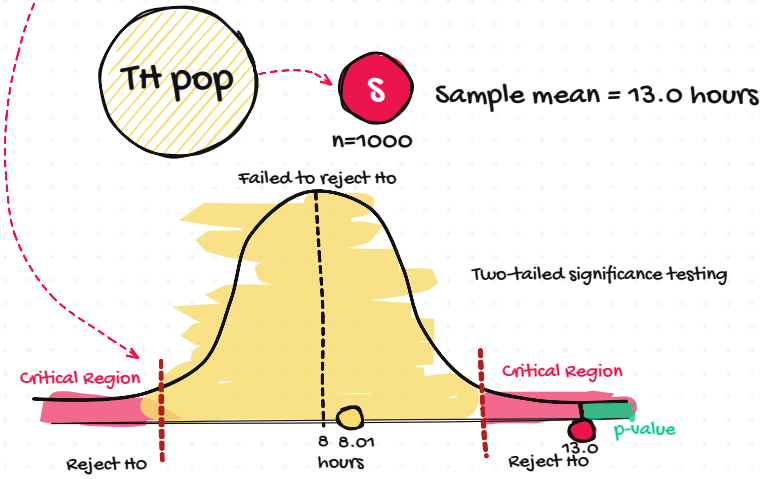
Ho: %like A - %like B = 0
Ha: %like A - %like B != 0

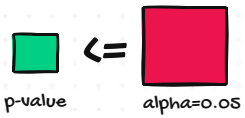


Foundations of Hypothesis Testing

Frequentist Approach

Ho: average Thai sleep = 8 hours
Ha: average Thai Sleep != 8 hours

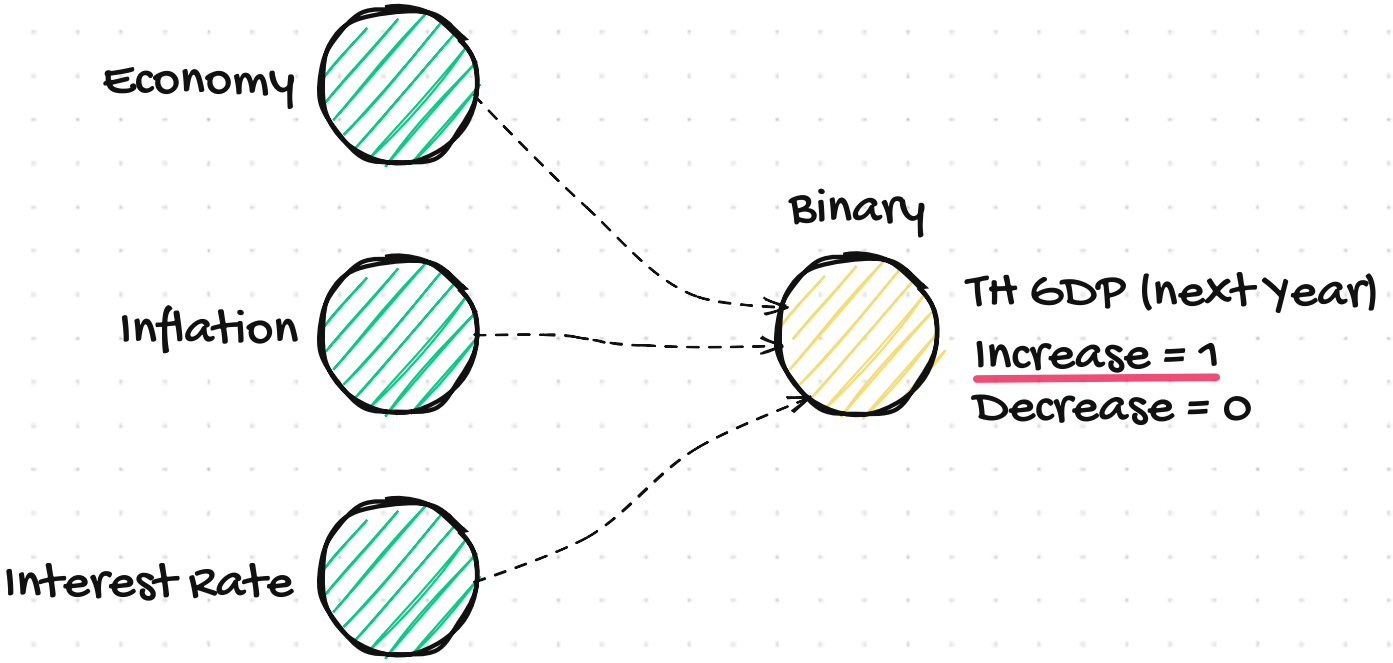




Reject Ho

p-value = p(observed data or more extreme | Ho is true)

Logistic Regression



GDP = sigmoid(f(economy, inflation, interest rate))

$p(Y=1) = \text{sigmoid}(b_0 + b_1 * \text{economy} + b_2 * \text{inflation} + b_3 * ir)$

$1 / (1+e^{-Z})$

[0, 1]

=IF(p(Y=1) > 0.5, 1, 0)

Intro Neural Network