STATS

- to get useful information from the data.

2 types

DESCRIPTIVE | INFERENTIAL

↑ To describe data in hand

↑ To infere something from data for a larger population.

Areas of Descriptive Stats

* Measures of central tendency - mean (EV), median, mode, quartiles.

* Measures of dispersion/ variation - SD, variation, range.

EV - Expected Value - same as average

(For Casinos, portfolio managing strategies for investment)

- Depending on no. of values resulted in mode → Unimodal, bimodal, uniform, multimodal.

* Variance vs SD → Square root of variance.

↓

Average of squared differences from mean.

↓

Squaring so that +ve - ve values wont cancel out.

Both SD and variance indicates how much data is spreaded across w.r.t Mean.

✗ SD's unit same as the underlying data.

$$\sigma^2 = \Sigma (x_i - \bar{x})^2 / N$$

Conditional Probability.

AND → *

OR → +

Eg:- Picking a house in Street1 given we already picked a house in same street. $(P_1 * P_2)$

PMF vs PDF

Prob. Mass Function

Eg:- Barcharts / Piecharts / Tally charts

For discrete random variables

Distribution Func.

For continuous random varia

DISCRETE PROBABILITY DISTRIBUTIONS.

- Binomial Distributions → Eg:- Coin toss.
- Poisson Distributions → Eg:- A Particular cell within a large population will acquire a mutation.
- Geometric Distributions
- Negative Binomial Distributions

CONDITIONS:

1) Total. no. of trials fixed.
2) Each trial is binary
3) Prob. of success is same in all trials.

FORMULA $nC_r (P)^r (1-P)^{n-r}$

Poisson - To describe distribution of rare events in a large population.

- In another way, it shows how many times an event is likely to occur within a specified period of time. $e^{-\lambda} \lambda^x / x!$

Normal Distribution / Gaussian

- distribution that is symmetric about the mean.

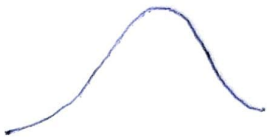
1-2-3 rule. w.r.t SD of mean.

68% 95% 99.7%

Skewness → measure of symmetry

Kurtosis → measure of whether data heavily-tailed or lightly tailed.

+ve Skew (Right Skewed Distribution)



-ve Skew (Left Skewed Distribution)

STANDARD $-1.96 \leq S.K \leq +1.96$.

Sample - Subset of population.

Sampling Distribution - Distribution of means of different samples.

Central Limit Theorem (CLT)

- Sampling distribution mean = Population mean.
- For $n > 30$, Sampling distribution will be normally distributed.

To find out Skewness

- Univariate plots.
- Shapiro-Wilks test (from scipy.stats import shapiro)
- .skew() method from pandas.

Handling Skewness

* Square-Root Transformation
* Log Transformation
* Reciprocal Transformation
* Box cox transformation
* Yeo-Johnson

- Confidence Interval (CI) - Range of possible values for an unknown parameter.

An associated confidence level with this (that indicates certainity).

$$CI = \left(\bar{x} - \frac{z^* s}{\sqrt{n}}, \bar{x} + \frac{z^* s}{\sqrt{n}} \right)$$

$\bar{x}$ → sample mean

s → sample SD

n → sample size

z^* → z-score associated with Confidence Level.

NOTE: Z-score is the value associated with a particular Cumulative probability and Z^* is z-score associated with confidence level.

Sampling Types.

Random Sampling (RS)
* Simple RS with replacement
* Simple RS without replace
* Stratified RS
* Cluster Sampling
* Systematic Sampling

Non-Random
* Convenience
* Quota
* Judgemental
* Snowball.

Sampling Useful in:

→ Quality Control
→ Pilot Testing
→ Market Research
→ Market Campaign Efficacy.

HYPOTHESIS TESTING - method of statistical inference

Null Hypothesis (H_0): status-quo / default position / no relation b/n variables.

Alternate Hypothesis (H_1): alternate to Null.

So 2 cases:
- Rejection of Null Hypothesis
- Failure to Reject Null Hypothesis.

NOTE:- H_0 assumed to be true and statistical evidence required to reject it in favour of Alternate Hypo.

Type of Errors:

		Truth about population	
		H_0 true	H_a true
Decision based on sample.	Reject H_0	Type-1 Error	Correct Decision
	Accept H_0	Correct Decision	Type-2 Error

Type-1 error → Rejection of True Null Hypothesis.

Type-2 error → Fail to reject a False Null Hypothesis.

In terms of Confusion Matrix,
Type-1 ≅ False Positives
Type-2 ≅ False Negatives.

Methods that supplement Critical Hypothesis:

- CRITICAL VALUE METHOD → * Find the z-critical with the significance level (α)
  * Find upper & lower critical values with z-critical with $\mu \pm (Z_c * \sigma_{\bar{x}})$
- P-VALUE METHOD.

* p-value is the probability of obtaining extreme-test result, assuming Null hypothesis is correct.
* If p-value less than α, then reject Null Hypothesis.

T-tests: → If sample size less than 30 and population SD/variance not known

Different types:
- 1-sample t-test → scipy.stats.ttest_1samp()
- 2-sample t-test → scipy.stats.ttest_ind()
- Paired t-test. → stats.ttest_rel()

ANOVA
(Analysis of Variance)
- To check statistical similarity of 2 or more groups.
- measured using F-Ratio.
- Other forms of ANOVA → MANOVA, ANCOVA.

CHI-SQUARE TESTS.
- Used when we have single categorical variable.
- 2 types: Test of Independence
  Goodness of Fit Test.