Stats formula (1)

pmf - probability mass function
cdf - cummulative density function
(stats.binom.cdf(x,n,p)
(stats.binom.cdf(n, λ)

Formulas

- Mean $= \Sigma x . P(x)$
- Variance $= \Sigma (x-\bar{x})^2 . P(x)$
- Discrete Probability
  - Binomial Distribution > $P(x) = {}^nC_x p^x q^{n-x}$
    stats.binom.pmf(x,n,p)
  - Poisson Distribution > $P(X) = \frac{e^{-\lambda}\lambda^x}{x!}$ (where λ is mean)
    stats.poisson.pmf(n, λ)
- Standard deviation $= \sqrt{\Sigma (x-\bar{x})^2 . P(x)}$
-

	Binomial	Poisson
mean	np	λ
S.D	$\sqrt{npq}$	$\sqrt{\lambda}$

- Isf (Inverse survival function)
  stats.norm.isf(0.025)
  - Gives the 95% distribution range (Z-Score)

- Characteristics of a standard Normal Distribution ⇒ $\bar{X} = 0$ then S.D = 1

-1.96 to 1.96 → 95%
-1.64 to $+1.64$ → 90%
-2.58 to $+2.58$ → 99%

Empirical Formulas

- $\bar{\bar{x}} - 1SD$ to $\bar{\bar{x}} + 1SD$ } 68% CI, 32% Risk
- $\bar{\bar{x}} - 2SD$ to $\bar{\bar{x}} + 2SD$ } 95% CI, 5% Risk
- $\bar{\bar{x}} - 3SD$ to $\bar{\bar{x}} + 3SD$ } 99% CI, 1% Risk

Area Under the Normal Distribution

$\mu - \sigma \le x \le \mu + \sigma \rightarrow 0.6828$
$\mu - 2\sigma \le x \le \mu + 2\sigma \rightarrow 0.9545$
$\mu - 3\sigma \le x \le \mu + 3\sigma \rightarrow 0.9974$

- Sampling Error:-
  Difference b/w population mean and sample mean
- Standard Error :- $\sigma (S.D) / \sqrt{n \text{(sample size)}}$

- np.random.seed() - same random numbers will be generated.

-
  - Theoritical Approach :- we take many trials to calculate mean of the mean. → $\bar{\bar{x}} \pm 1.96$ (S.D)
  - Practical Approach :- we do only one round of sampling → $\bar{x} \pm 1.96 (SD/\sqrt{n})$ (standard error of the mean)

- Tolerance level of skewness is ±0.5

To find the confidence interval:-

$$CI = \text{sample mean} \pm (\text{margin of error})$$

$$\text{margin of error} = z\,\text{critical value} \times \frac{\text{Population S.D}}{\sqrt{\text{sample size}(n)}}$$

• Confidence interval is also known as acceptance zone for Ho.

★ Formula for 1 sample T test (Test of mean) is $\Rightarrow \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ (how much the sample mean $\bar{x}$ is away from hypothesized mean μ)

$\boxed{\text{ttest\_1samp(samp\_array, 300)}}$ (μ)

★ Formula for two sample T test (Test of mean) unpaired (Independent groups) $\Rightarrow \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

$\boxed{\text{t\_test\_ind}(g_1, g_2)}$

• Table to reject Ho

	tstat (dist b/w two means)	Pvalue (area)
95%	>1.96	<0.05
99%	>2.58	<0.01
90%	>1.64	<0.1

np.std[] by default divides by n.
df['vol'].std() by default, divides by n-1.
• ddof = 1, means the denominator is n-1.
ddof (delta degrees of freedom.

• when population S.D is not known use T-Test, if known, use z test

• when n = 30 and above, $s \simeq \sigma$, then tstat = Zstat

★ Formula for two sample T test (Test of mean) paired (Dependent groups)

$$\Rightarrow \frac{\Sigma D/n}{\sqrt{\dfrac{\Sigma D^2 - \dfrac{\Sigma D^2}{n}}{(n-1)n}}}$$

$\boxed{\text{ttest\_rel}(g_1, g_2)}$

rel means relativity
D - Difference
D^2 - Squared Difference

★ Two sample Proportion_z test formula
(This comes under Two sample T Test (Test of Proportion)

Formula:- $\dfrac{P_1 - P_2}{\sqrt{P_{Pooled}(1 - P_{Pooled}) \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Zstat on Z data

proportion_z test $([x_1, x_2], [n_1, n_2])$

PPooled means pooled proportion

★ One Sample Proportion Test comes under One sample T test
prop_z test (x, n, prop). (prop is the expected proportion)

~~This concept has many other flavours~~

we have four flavors of proportion test:-

- one sample
  - 2 categories
  - >2 categories
- Two sample prop, df = 1
- chi sq test, df ≥ 2

★ Chisquare :- we have two groups with degrees of freedom 2 and above, then chi2_contingency () is used.

★ Two Sample Proportion :- prop_z test $([x_1, x_2], [n_1, n_2])$

★ One sample (2 categories) :- prop_z test (x, n, expected proportion)

★ One sample (>2 categories):- chi square ([ ], [ ])
(we should put either numerical list (or) numpy array)