Inferential Statistics - II

Sample Size
Sampling Techniques
Distributions
T-test

Population vs Sample

What is a population?

- The total number of units one is interested to study.

What is a sample?

- A subset of units from the interested larger set of units (this is needed for inference making)

Unit of analysis vs sampling unit

- The units you are interested in studying verses a unit (or set of units) considered for sampling.

Parameter vs Statistics

- What we use to characterize a population versus how we characterize a sample

Population vs Sample (notations)

Population/Sample	Term	Notation	Formula
Population (X_1, X_2, X_3, ……, X_N)	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N} (X_i - \mu)^2}{N}$
Sample (X_1, X_2, X_3, ……, X_n) (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n} (X_i - \bar{X})^2}{n-1}$
Sampling Distribution of the Sample Mean ($\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, ……, $\bar{X}_k$) (k Sample Means)	Sampling Distribution's Size	No convention (We have used k, but that is not a norm)	
	Sampling Distribution's Mean (mean of sample means)	$\mu_{\bar{X}}$	$\mu_{\bar{X}} = \mu$
	Sampling Distribution's Standard Deviation	S.E. (Standard Error)	S.E. = $\sigma / \sqrt{n}$

Sampling Techniques - Probabilistic

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Random Sample
- Clustered Random Sample

Simple Random Sampling

```
seed = 0
treatment1 = data_for_sample_size[
    data_for_sample_size.avg_amount_payable_ceiled > 100.0].sample(n=4, random_state = seed)

treatment1
```

	city	yyyymmdd	weekday	pickup_cluster	time_period	surge_strategy_label	avg_dynamic_surge_pct	avg_amount_payable	median_dynamic_surge_pct
19	Chennai	20220730	Saturday	Anna Nagar	morning_peak	non_experiment	20.0	140.433333	20.0
31	Chennai	20220802	Tuesday	Anna Nagar	morning_peak	non_experiment	20.0	133.036364	20.0
49	Chennai	20220806	Saturday	Vadapalani	morning_peak	non_experiment	20.0	130.942308	20.0
26	Chennai	20220731	Sunday	Vadapalani	morning_peak	non_experiment	20.0	179.087558	20.0

Stratified Random Sampling

```
## see total counts of sample with normalization = True

proportional_stratified_sample['pickup_cluster'].value_counts(normalize = True)
```

```
T Nagar         0.5
Nungambakkam    0.5
Name: pickup_cluster, dtype: float64
```

```
## Compare with population
data_for_sample_size_strat['pickup_cluster'].value_counts(normalize=True)
```

```
T Nagar         0.5
Nungambakkam    0.5
Name: pickup_cluster, dtype: float64
```

Clustered Random Sampling

tinct_customers_treatment	treated_customers	rr_count	gross_orders	net_orders	FE_RR	G2N	avg_amount_payable_ceiled	orders_not_fulfilled	stress_category
177	333	54	68	46	0.1622	0.6765	98.0	22	bad
250	519	89	91	67	0.1715	0.7363	90.0	24	bad
175	364	65	82	57	0.1786	0.6951	82.0	25	bad
187	455	58	67	44	0.1275	0.6567	96.0	23	bad
173	435	49	55	27	0.1126	0.4909	99.0	28	bad
77	160	18	22	21	0.1125	0.9545	101.0	1	best
40	74	10	8	6	0.1351	0.7500	129.0	2	best
57	118	19	23	21	0.1610	0.9130	61.0	2	best
22	52	4	4	3	0.0769	0.7500	131.0	1	best
18	30	0	4	4	0.0000	1.0000	141.0	0	best

Difference Between Stratified and Clustered Sampling

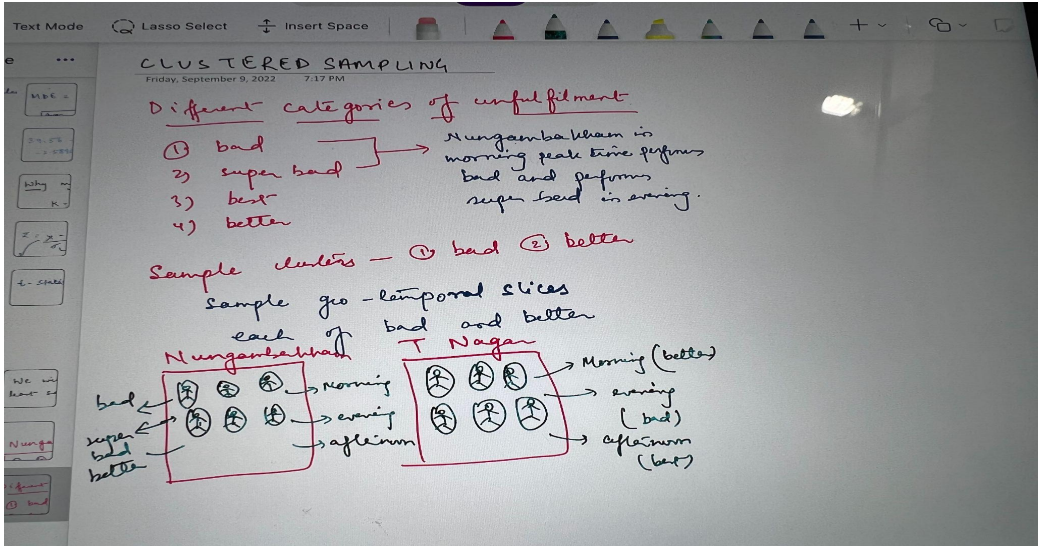
Stratified Random Sampling

- Split the population into subgroups
- Use simple random sampling on each subgroups

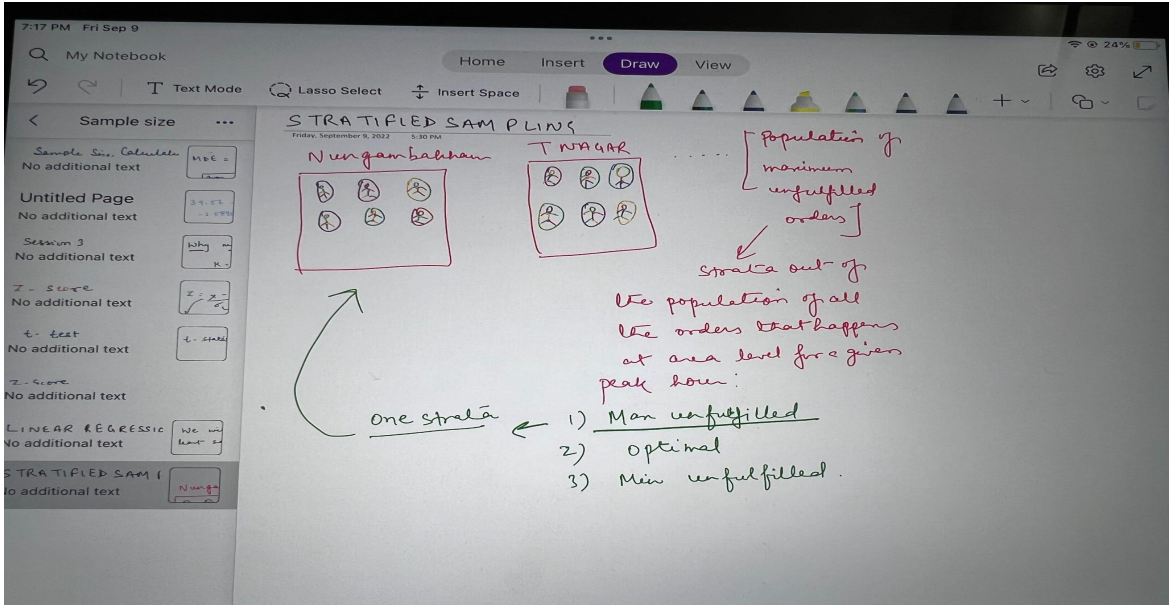
Clustered Random Sampling

- Use simple random sampling to select few subgroups
- Use simple random sampling on those subgroups

Clustering sampling



Stratified sampling



Non-Probability Samples

- Convenience samplings - Jimmy Kimmel Sampling
- Typical-case samples (Purposive)
- Expert Sampling (Purposive)
- Proportional and Non-proportional Quota Sampling
- Heterogeneity Sampling -> Sampling for ideas (Purposive)
- Snowball Sampling -> Start with one, then ask for recommendations

Sample Size

Why we need sample size?

- To avoid overpowering and underpowering

Sample Size can be calculated using two methods:

- Estimating Sample Size Based on a Proportion.
- Estimating Sample Size Based on a Mean.

Estimating Sample Size based on mean

$$n = \frac{(1.96)(1.96)}{d*d}\sigma^2$$

n = sample size
d = degree of precision
σ = standard deviation

Estimating Sample Size based on proportion

$$n = \frac{(1.96)(1.96)}{d*d} pq$$

b = sample size
d = degree of precision
p = proportion of the population having the characteristics of interest.
q = 1-p

Distributions

- Uniform Distribution
- Normal Distribution
- Poisson Distribution

Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

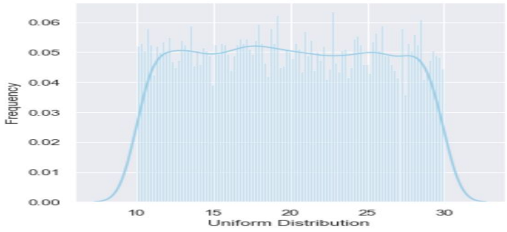
Uniform distribution

```
# random numbers from uniform distribution
n = 10000
start = 10
width = 20
data_uniform = uniform.rvs(size=n, loc = start, scale=width)
```

```
ax = sns.distplot(data_uniform,
                  bins=100,
                  kde=True,
                  color='skyblue')
ax.set(xlabel='Uniform Distribution ', ylabel='Frequency')
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packages/seaborn/
eWarning: `distplot` is a deprecated function and will be removed in a future version.
se either `displot` (a figure-level function with similar flexibility) or `histplot` (
istograms).
  warnings.warn(msg, FutureWarning)
```

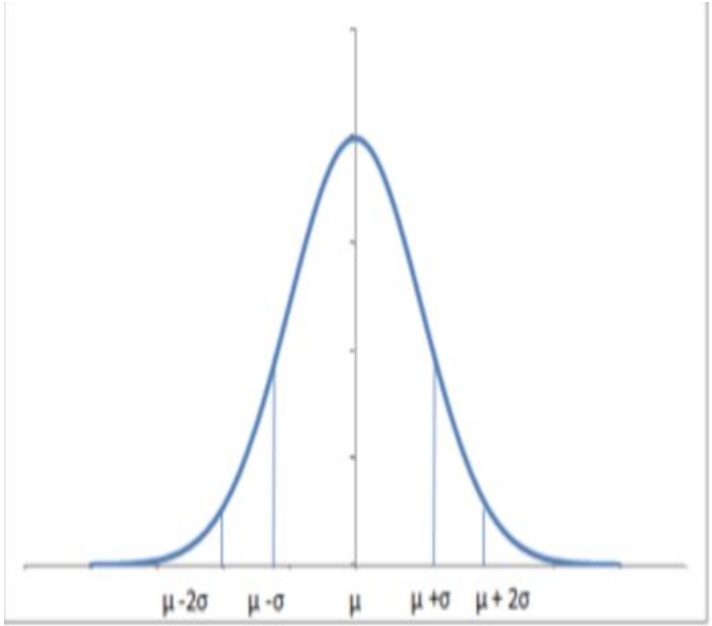
```
[Text(0.5, 0, 'Uniform Distribution '), Text(0, 0.5, 'Frequency')]
```



Normal Distribution

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 66% of the observations lie within 1st standard deviation.
- 95% of the observations lie within the 2nd standard deviation
- 99.7% of the data lie within the 3rd standard deviation



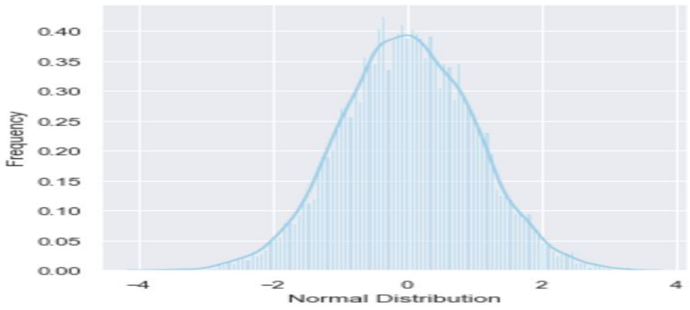
Normal Distribution

```
from scipy.stats import norm
# generate random numbers from N(0,1)
data_normal = norm.rvs(size=10000,loc=0,scale=1)
```

```
ax = sns.distplot(data_normal,
                  bins=100,
                  kde=True,
                  color='skyblue')
ax.set(xlabel='Normal Distribution', ylabel='Frequency')
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/sit
eWarning: `distplot` is a deprecated function and will be removed i
se either `displot` (a figure-level function with similar flexibili
istograms).
  warnings.warn(msg, FutureWarning)
```

```
[Text(0.5, 0, 'Normal Distribution'), Text(0, 0.5, 'Frequency')]
```



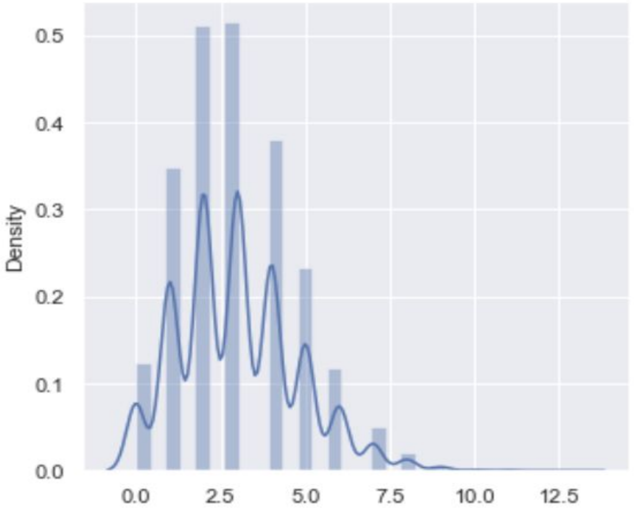
Poisson Distribution

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Application in Poisson regression.

$$f(D) = e^{\alpha S + \beta}$$

```
from scipy.stats import poisson
data_poisson = poisson.rvs(mu=3, size=10000)
```



Z-score calculation

$$Z = \frac{X - \mu}{\sigma x}$$

σ	Standard Deviation	μ	Mean
n	Number of observations		
X	Observation value		

Z-score calculation

Question: **If Y is distributed N(1,4), find Pr(Y <= 3)**

μ = 1 ,

$\sigma 2$, Variance = 4 σ = 2 X = 3

Pr (Y <= 3) = $\Pr\left(Z \leq \left(\frac{3-2}{2}\right)\right) = \Pr(Z \leq 0.5) = 0.6915$

Using Z-score for Z-test

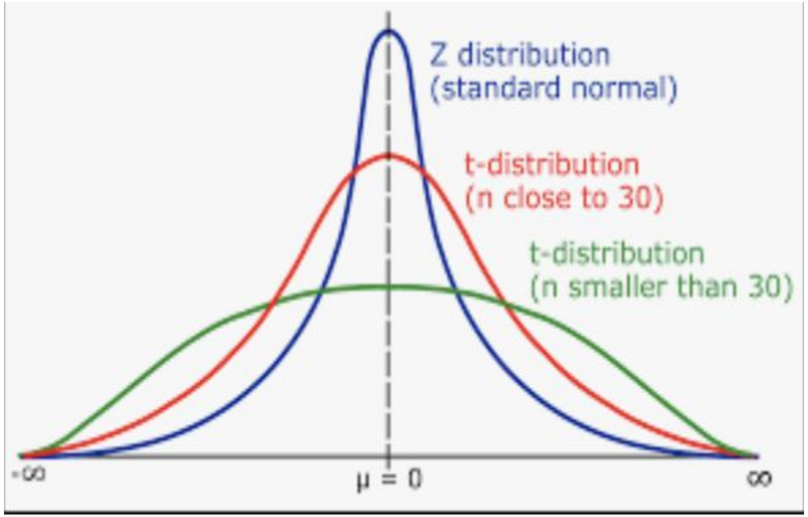
Calculating **p-value**

Left tailed test: ϕZ_{score}

Right tailed test: **1 -** ϕZ_{score}

Two -tailed test: **2 - 2** ϕZ_{score} OR $2\phi\text{-}|Z_{score}|$

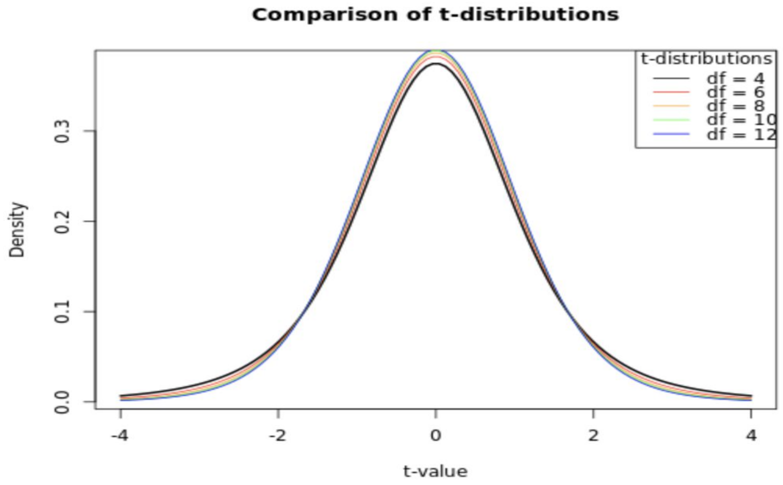
t-distribution



Source: [Source Link]

Difference between Normal distribution and t - distribution

Plotting t-distribution for small sample size



T - distribution plot at different degrees of freedom (here dfs = no of observations -1)

T-test one sample

$$t = \frac{X - \mu}{\sigma x}$$

X = Sample Mean

$$\text{where} \quad \sigma x = \frac{\sigma}{\sqrt{n}}$$

Standard Error

σ , Standard deviation

n , no of observations

T-test two sample

$t_{act} = \mu_{tn} - \mu_{b} - d_{0} / SE\ (\mu_{tn} - \mu_{b})$

$SE\ (\mu_{tn} - \mu_{b}) = root\ (S_{tn}^{2}/n_{tn} + S_{b}^{2}/n_{b})$

T-critical values at different alpha level

Confidence Interval	Alpha	t-critical Value
99%	0.01	2.58
95%	0.05	1.96
90%	0.1	1.64

t-critical Values At Different CI

Estimating Confidence interval

$$\overline{X} \pm 1.96\ \overline{SE}$$

At alpha 0.05 level, 95% CI

Central limit Theorem

https://onlinestatbook.com/2/sampling_distributions/clt_demo.html

Central limit theorem states that as we increase the sample size the distribution tends to achieve a normal distribution.

References

For Z - table: https://www.simplypsychology.org/z-table.html

For student t-table: https://www.tdistributiontable.com/

https://www.kaggle.com/datasets/henslersoftware/19560-indian-takeaway-orders