Statistical Techniques for Data Science

Sampling and Distributions

Prof.Gangaboraiah PhD (Stats)

Learning objectives of this unit

- At the end of the session on unit 3, the student should be able to
  - Define between, population, sample, target population, sampling frame, sampling interval, and sampling unit
  - Distinguish between different sampling methods
  - Understand sampling variation, point estimates, interval estimates, and confidence interval.
  - Compute 100 (1-α)% CI for mean, proportion, variance and ratio of variances
  - Define central limit theorem

Sampling

Example 1

To know whether the rice is boiled or not in a in a cooker, it is enough to check a few grains randomly instead of checking the whole grains in the cooker

Example 2

To know the number of red blood cells in a person the Researcher should be satisfied with the estimate based on a few drops (sample) of blood; he cannot think of extracting all the blood (population) from the body.

- ***Population***: **set of people or entities to which findings are to be generalized. Should be defined explicitly before the sample is taken**

- ***Census or Enumeration:*** **collection of data from every person or entity in the population.**

Target Population

- The population which is in direct relation to the samples drawn
- Finite and Infinite target population
- Homogeneous and Heterogeneous target population

Sampling frame

a listing of all the elements in a population

Sample

a subset of the target population chosen so as to be representative of that population.

Sampling unit

a member of the sample

Sampling:

Is the process of selecting units from a population of interest so that by studying the sample, results can be generalized back to the Population from which they were chosen.

Why draw a sample?

- The entire group is too large to study
- Time efficient, cost effective and feasible
- Can provide a close approximation of the population
- Information actually be more accurate when based on carefully drawn samples
- Offer greater scope and flexibility than a census

Sampling Methods

Probability sampling

Procedure that assures that all the units in the population have some **probabilities** (chance) known in advance of being chosen in a sample

Non-probability sampling

Procedures in which units in the sample are collected with **no specific probability** structure

Probability Sampling

- Simple random sampling
- Systematic sampling
- Stratified random sampling
- Multistage sampling

Simple Random Sampling

- Target Population must be Homogeneous and finite
- Population is relatively small
- Sampling frame is complete and up-to-date.
- Samples are selected unit by unit
- Each sampling unit will have and equal chance of being selected
- The random selection from the sampling frame can be done using a table of random numbers table or Lottery method

Random Number Table

73735	45963	78134	63873
02965	58303	90708	20025
98859	23851	27965	62394
33666	62570	64775	78428
81666	26440	20422	05720
15838	47174	76866	14330
89793	34378	08730	56522
78155	22466	81978	57323
16381	66207	11698	99314
75002	80827	53867	37797
99982	27601	62686	44711
84543	87442	50033	14021
77757	54043	46176	42391
80871	32792	87989	72248
30500	28220	12444	71840

Systematic Random Sampling

- Target Population Homogeneous and finite/ infinite
- Compute the sampling interval k=(N/n)
- Samples are selected unit by unit
- Only first sample is selected at random.
- Subsequent samples are selected at an interval of k

Stratified Random Sampling

- Target Population Heterogeneous and finite
- Divide the heterogeneous target population into different stratum ensuring homogeneity within the stratum
- Using Probability Proportional to Population Size (PPPS), Samples are selected from each stratum

Stratified Random Sampling

- Calculate the estimates for each stratum separately
- Combine the estimates to generalize the results to the whole population from where samples were drawn.

Non-Random Sampling

- Purposive/ Judgment sampling
- Convenience sampling
- Quota sampling
- Snowball technique

Sampling variation

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

Sample 1

3000 2486 820 1678 2070 2638 2490 1865 1000 2090 596 3200

Sample 2

2840 2858 3000 2490 2998 3050 2070 2896 3200 2490 3280

Sample 3

2858 3240 2497 2865 656 2093 934 1861 868 795

Sample 4

2086 1000 2497 596 656 875 2085 934 1313

Sample 5

820 1313 3000 2640 596 2640 2600 2495 934 2500

Sample 6

2840	2499	1327	1861	2495	3024	3038	2497

Sample 7

2858	2490	868	1670	1480	2643	1480	1680	2085	2490

Sample 8

2495	2858	1861	2092	2499	3000	2660	1000	1679	926	2660

Sample 9

795	791	3200	2085	2638	2497	2486	1159	2640

Sample 10

3019	3240	3200	3050	3000	3015	2900	2896	2998

$$\frac{3000+2486+820+2070+2638+2490+1865+1000+2090+596+3200}{12}=1994.42$$

$$\frac{2840+2858+3000+2490+2998+3050+2070+2896+3200+2490+3280}{11}=2830.14$$

$$\frac{2858+3240+2497+2865+656+2093+934+1861+868+795}{10}=1866.70$$

$$\frac{2086+1000+2497+596+656+875+2085+934+1313}{9}=1338.00$$

$$\frac{820+1313+3000+2640+596+2640+2600+2495+934+2500}{10}=1953.80$$

$$\frac{2840+2499+1327+1861+2495+3024+3038+2497}{8}=2447.63$$

$$\frac{2858+2490+868+1670+1480+2643+1480+1680+2085+2490}{10}=1974.40$$

$$\frac{795+791+3200+2085+2638+2497+2486+1159+2640}{9}=2032.33$$

$$\frac{2495+2858+1861+2092+2499+3000+2660+1000+1679+926+2660}{11}=2157.27$$

$$\frac{3019+3240+3200+3050+3000+3015+2900+2896+2998}{9}=3035.33$$

<u>True</u> (not observable) value

Mean FBS=2162.24

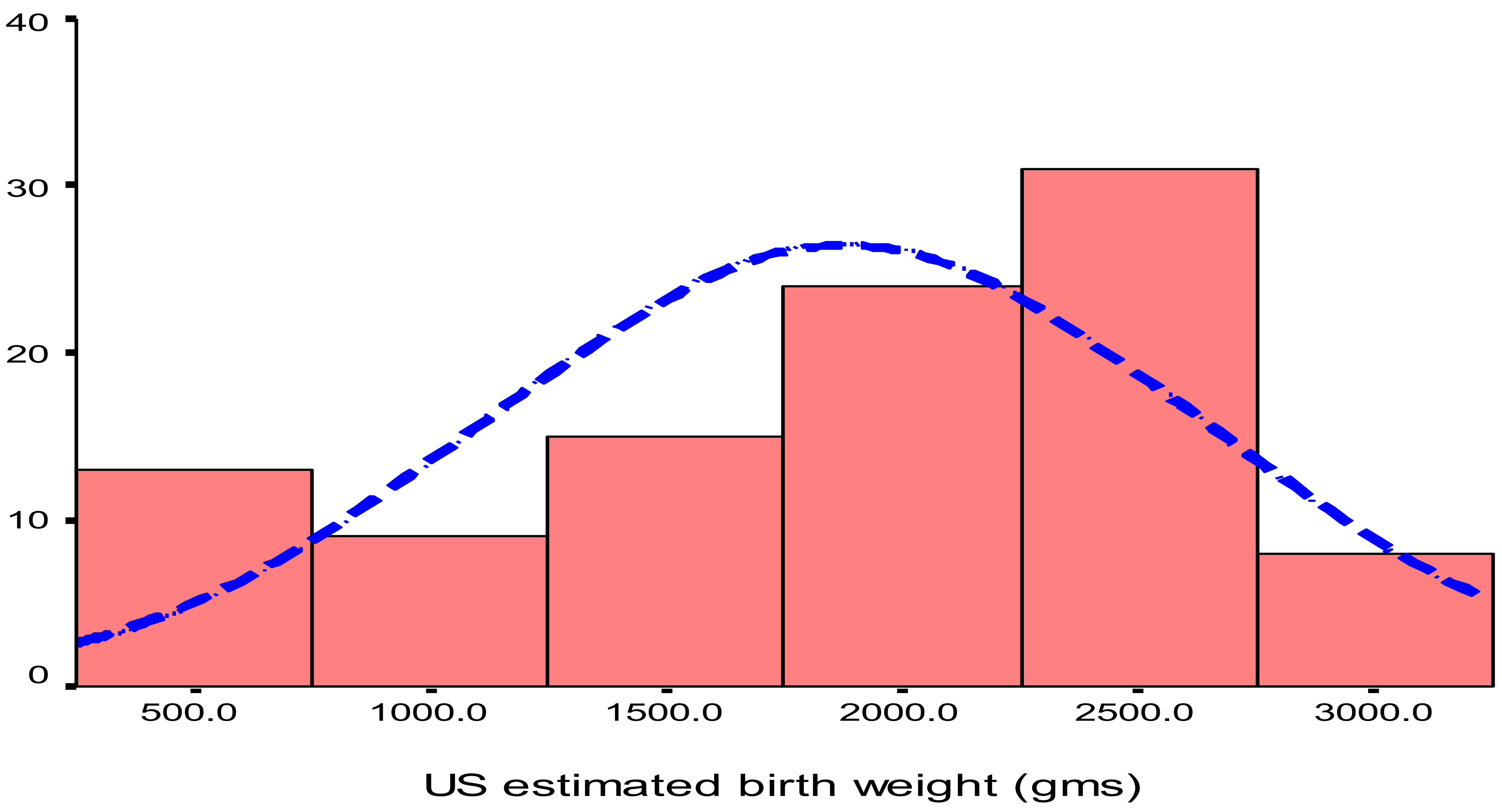
Sample No.	Sample size	Mean	SD
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
Overall	100	2162.24	732.26

Sampling variation

Variation in Sample estimates, even if the samples drawn are from same population

Distribution of Ultrasound estimated birth weights at different gestational age

US estimated birth weight (gms)	Frequency	Percent
501-1000	13	13
1001-1500	9	9
1501-2000	15	15
2001-2500	24	24
2501-3000	31	31
3001-3500	8	8
Total	100	100


40
30
20
10
0
500.0
1000.0
1500.0
2000.0
2500.0
3000.0
US estimated birth weight (gms)

US estimated birth weight (gms)	Values
Mean	2162.24
Median	2492
Mode	1861
Minimum	590
Maximum	3240
Range	2650
Variance	536204.20
Std. Deviation	732.26

Do you consider these sample means and sample SDs as variable?

If yes, should we not describe the distribution of these variables?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

The most important one to be computed from these sample estimates is the **standard deviation** of **sample mean**, **sample proportion**, **Sample correlation** etc. as are computed for individual observations

Sampling distribution

Definition

- The probability distribution of a statistic (sample estimate) is called sampling distribution.
- The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.

Sampling distribution of means

- Suppose that a random sample of size n taken from a normal population with mean μ and variance σ^2. Now each observation in a sample X_1, X_2, …, X_n is a normally and independently distributed random variable with mean μ and variance σ^2 . Then by the reproductive property of normal distribution

Sampling distribution of means

The sample mean $$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

has a normal distribution with mean $$\mu_{\bar{x}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$$

and variance $$\sigma^2_{\bar{x}} = V(x) = V\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Sampling distribution of means

If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean μ and variance σ^2/n if the sample size n is large. This is one of the most useful theorems in statistics called central limit theorem

Sampling distribution of proportions

Suppose that a random sample of size n taken from a binomial population with mean $\mu=np$ and variance $\sigma^2=npq$. By defining Z=(Estimator-mean)/SD, and with mean $\mu=np > 5$ and as increases, the binomial distribution converges to standard normal distribution. Hence, the sampling distribution of sample proportion is distributed as standard normal distribution.

Sampling distribution of variance

- Like the sampling distribution of proportion and mean, the sampling distribution of sample variance can also be found. Since the variance S^2 cannot be negative, the sampling distribution of S^2 is not normal. In fact it is related to Gamma distribution.
- Define $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is a random variable having Chi-square distribution with v = n-1 degrees of freedom.

Sampling distribution of ratio of variances

Suppose that we have two independent normal population with unknown variance σ_1^2 and σ_2^2 respectively. We have two random sample of sizes n_1 and n_2 respectively, from these two populations and let S_1^2 and S_2^2 be the two sample variances. Then we can find 100(1-α)% CI for the ratio of the two variances $\frac{\sigma_1^2}{\sigma_2^2}$

Sampling distribution of ratio of variances

The sampling distribution of the ratio of the two variances is distributed as Fisher's F-distribution with (n_2-1, n_1-1) degrees of freedom (df), that is,

$$F = \frac{\dfrac{s_2^{\,2}}{\sigma_2^{\,2}}}{\dfrac{s_1^{\,2}}{\sigma_1^{\,2}}}$$ is distributed with (n_2-1, n_1-1) df.

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and variance σ^2, and if $\bar{X}$ is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

as $n \to \infty$, is the standard normal distribution with mean 0 and variance 1

Standard error or mean

Standard error of mean is

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of observations in the observed sample and n is the number of observations in the sample.

Standard error of proportion

Standard error of a proportion is

$$SE(p) = \sqrt{\frac{pq}{n}}$$

where p is the proportion of occurrence of an event in the observed sample, q = (1 – p) and n is the number of observations in the sample

Standard error of difference between two means if the two sample sizes are not equal

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}\right)}$$

where

- s_1: the sample SD of group 1

- s_2: the sample SD of group 2

- n_1: the sample size of group 1

- n_2: the sample size of group 2

Standard error of difference between two proportions when two sample sizes are same

$$SE\,(p_1 - p_2) = \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}, \quad q_1 = 1 - p_1, \quad q_2 = 1 - p_2$$

where

- p_1 : The sample proportion of occurrence of an event in the group 1
- p_2 : The sample proportion of occurrence of an event in the group 2
- n_1: The sample size of group 1
- n_2: The sample size of group 2

Point estimation

Point estimate is a single number, calculated from available sample data that is used to estimate the value of an unknown parameter

Point estimation

The statistic

♣ Mean ($\bar{x}$)

♣ Variance (s^2)

♣ Proportion (p)

♣ Correlation (r) etc.,

computed from sample observations estimates of population parameters μ, σ^2, P, and ρ

Sample estimates		Population parameters
Sample Mean ($\bar{x}$)	$\longrightarrow$	**Population Mean (μ)**
Sample S D (s)	$\longrightarrow$	**Population S D (σ)**
Sample Proportion(p)	$\longrightarrow$	**Population Proportion(P)**
Sample Correlation Coefficient (r)	$\longrightarrow$	**Population Correlation Coefficient(ρ)**

Interval estimation

Interval estimate is an interval that provides a lower and upper bound for a specific unknown parameter.

Undoubtedly, the most powerful type of inference.

Confidence Interval

Computation of 100 (1-α)% confidence interval is the most common way of finding the interval estimate, where α is the probability of type I error.

Confidence Interval

Confidence interval is an interval of numbers believed to contain the parameter value.

- The probability, the method produces an interval that contains the parameter is called the confidence level. Most studies use a confidence level close to 1, such as 0.95 or 0.99.

Most CIs have the form

Point estimate ± Margin of error

with margin of error based on spread of sampling distribution of the point estimator; e.g., margin of error $\cong$ 2 (standard error) for 95% confidence.

Finding Confidence Interval in practice

The100 (1-α)% confidence interval for mean is

$$\text{Sample estimate} \pm z_{\alpha/2}\ \text{SE (estimate)}$$

$$\bar{x} \pm z_{\alpha/2}\ \text{SE}\,(\bar{x})$$

For α=0.05, $z_{\alpha/2} = 1.96$

For α=0.01, $z_{\alpha/2} = 2.58$

Example

Suppose a regional computer center wants to find the performance of its disk memory system. One measure is the average time between failure of its disk drive. The mean time between failure of a random sample of 20 disk drive is 1762 hours with a population standard deviation of 215. Construct (i) 95% CI and (ii) 99% CI for mean time between failure of its disk drive.

The SE (mean)= $\dfrac{\sigma}{\sqrt{n}} = \dfrac{215}{\sqrt{20}} = \dfrac{215}{4.47} = 48.10 \text{ hours}$

95% CI= 1762 ± 1.96 (48.1) = (1762 - 94.28, 1762 + 94.28)
= (1667.72, 1856.28)

99% CI= 1762 ± 2.58 (48.1) = (1762 – 124.10, 1762 + 124.10)
= (1637.90, 1886.10)

Finding Confidence Interval for proportion

The100 (1-α)% confidence interval for proportion is

$$p \pm z_{\alpha/2}\, SE\,(p)$$

For α=0.05, $z_{\alpha/2} = 1.96$

For α=0.01, $z_{\alpha/2} = 2.58$

Example

It was observed that 4 out of every 20 persons own an Audi car in a city. Construct (i) 95% CI and (ii) 99% CI for the proportion of persons owning an Audi car.

The SE (proportion)= $\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.2 * 0.8}{20}} = \sqrt{\frac{0.16}{20}} = 0.09$

95% CI= 0.2 ± 1.96 (0.09) = (0.2 – 0.18, 0.2 + 0.18)
= (0.02, 0.38)

99% CI= 0.2 ± 2.58 (0.09) = (0.2 – 0.23, 0.2 + 0.23)
= (-0.03, 0.43)

Finding Confidence Interval for difference between two means

The100 (1-α)% confidence interval for difference between two means

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\, SE(\bar{x}_1 - \bar{x}_2)$$

A taxi company is trying to decide whether to purchase brand A or brand B tires for its fleet of taxis. To estimate difference in two brands, an experiment is conducted using 30 of each brand. The tires are run until they wear out. Construct

(i) 95% CI and (ii) 99% CI for mean difference in mileage between two brands.

Brands	Sample size	Mean (kms)	SD (kms)
A	30	36300	5000
B	40	38100	6100

The SE (Diff. in mean) = $SE(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

where $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Solution: $s_p = \sqrt{\dfrac{29 \text{ x } 5000^2 + 39 \text{ x } 6100^2}{30 + 40 - 2}} = 5657.10$

$$SE(\bar{x}_1 - \bar{x}_2) = 5657.10 \text{ x } \sqrt{0.03 + 0.025} = 1496.42$$

(i) 95% CI for difference between means is

= (36300 -38100) ± 1.96 x 1496.42

= - 1800 ± 2932.98

= (- 4732.61, 1132.98)

(i) 99% CI for difference between means is

= (36300 -38100) ± 2.58 x 1496.42

= - 1800 ± 3524.61

= (-5660.76, 2060.76)

Actual difference between means is 1800 kms

Twice the SE (diff. in mean)= 2x 1496.42 = 2992.84

Inference: There is no difference in the mileage between the two brands of tires. Hence, either of the brand can be chosen

Inference:

- If the actual difference between means < Twice the SE (diff. in mean), then no difference between the two group means.
- If the actual difference between means ≥ Twice the SE (diff. in mean), then there is difference between the two group means.

Finding Confidence Interval for proportion

The100 (1-α)% confidence interval for proportion is

$$p \pm z_{\alpha/2}\, SE\,(p)$$

The100 (1-α)% confidence interval for difference between proportions

$$(p_1 - p_2) \pm z_{\alpha/2}\, SE\,(p_1 - p_2)$$

The following data relates to the two judges who have declared innocent defendants as guilty (false positives) due to lack of evidence. Construct 95% and 99% CI for difference in proportions

Judges	No. of defendants (n)	No. of false positives	False positive rate
1	2500	22	0.88%
2	3000	90	3.00%

The SE (Diff. in Proportion) =

$$SE\,(p_1 - p_2) = pq\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$p = \sqrt{\frac{(n_1 - 1)p_1 + (n_2 - 1)p_2}{n_1 + n_2 - 2}}, \quad q = 1 - p$$

The pooled proportion =

$$p = \sqrt{\frac{2499 \text{ x } 0.0088 + 2999 \text{ x } 0.03}{5498}} = 0.143$$

and

$$q = 1 - p = 1 - 0.143 = 0.857$$

The SE (Diff. in Proportion) =

$$0.143 \text{ x } 0.857 \sqrt{0.0004 + 0.0003}$$

$$= 0.026 \text{ x } 0.122 = 0.003$$

(i) 95% CI: $(p_1-p_2) \pm z_{0.025}\, SE\,(p_1-p_2)$

$= (0.88-3.00) \pm 1.96 \times 0.003$

$= -2.12 \pm 0.006$

$= (-2.126, -2.114)$

(ii) 99% CI: $(p_1-p_2) \pm z_{0.005}\, SE\,(p_1-p_2)$

$= (0.88-3.00) \pm 2.58 \times 0.003$

$= -2.12 \pm 0.008$

$= (-2.128, -2.112)$

Actual difference between proportion

= 3.00-0.88

= 2.12

Twice the SE (diff. in proportion) = 0.006

Inference: Observed difference (2.12 > 2 SE(p_1-p_2)). Hence, the two judges have differ in with respect to false positive rate.

- Large samples have narrower widths than small samples
- Higher confidence levels have wider intervals than lower confidence levels
- Narrow widths and high confidence levels are desirable, but these two things affect each other

Sampling distribution of variance

- The 100(1-α)%CI for variance of normal distribution σ^2 is given by

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},\ n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},\ n-1}}$$

Sampling distribution of ratio of variances

Then we can find 100(1-α)% CI for the ratio of the two variances is given by

$$f_{1-\frac{\alpha}{2},\, n_2-1,\, n_1-1} \leq F = \frac{S_2^{\,2}/\sigma_2^{\,2}}{S_1^{\,2}/\sigma_1^{\,2}} \leq f_{\frac{\alpha}{2},\, n_2-1,\, n_1-1}$$

ie.,

$$\frac{S_1^{\,2}}{S_2^{\,2}} f_{1-\frac{\alpha}{2},\, n_2-1,\, n_1-1} \leq \frac{\sigma_1^{\,2}}{\sigma_2^{\,2}} \leq \frac{S_1^{\,2}}{S_2^{\,2}} f_{\frac{\alpha}{2},\, n_2-1,\, n_1-1}$$

THANK YOU

Manipal ProLearn
#7, Service Road, Pragathi Nagar, Electronic City,
Bengaluru 560100
contact@manipalprolearn.com | manipalprolearn.com

Highlights of the program

- **First-of-its-kind program in the Data Science space**, equipping learners with competencies and skills boosting their visibility and credibility for future employment prospects.
- **Industry-relevant course curriculum**, with applications in multiple domains, where such talent is in demand
- **Highly experienced Subject Matter Experts (SMEs)** from academia, IT and Data Science industry
- **Enhanced learning experience** through the digital LMS - **EduNxt**
- State-of-the-art infrastructure, latest technology and a well-equipped, 77,000 square feet residential campus.

Highlights of the program

- Delivery Models: Online, blended, face to face
- Domain Expertise in Banking, Retail, Healthcare
- Industry Partnerships with Genpact, IBM-BDU, Coursera
- Centre Of Excellence with Deakin University
- Academic Partnership with Manipal University