INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 3, 2023

Standard Normal Distribution

- ❑ is a normal distribution with mean = 0, variance = 1
- ❑ Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- ❑ Let $X \sim N(\mu, \sigma^2)$
- ❑ Hence, $X - \mu \sim N(0, \sigma^2)$
- ❑ Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- ❑ Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Basic Ideas of Sampling

1. Population (Sometimes, it is not even observable and only abstract)
2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
3. Subject
4. Parameter(Constant - might be unknown)
5. Statistic (Random Variable)

Basic Ideas of Sampling

1. Population (Sometimes, it is not even observable and only abstract)
2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
3. Subject
4. Parameter(Constant - might be unknown)
5. Statistic (Random Variable)
6. Statistic ceases to be a random variable after it is observed

Types of Sampling

- Simple Random Sample With replacement
- Simple Random Sample without replacement
- Cluster Sampling
- Stratified Sampling

Potential Causes of Bias

- ❑ Convenience Sampling
- ❑ Volunteer Sampling
- ❑ Systematic Sampling
- ❑ Non-response Bias
- ❑ Response Bias

Does that mean we shouldn’t use any of these types of sampling??

Potential Causes of Bias

- ❑ Convenience Sampling
- ❑ Volunteer Sampling
- ❑ Systematic Sampling
- ❑ Non-response Bias
- ❑ Response Bias

Does that mean we shouldn't use any of these types of sampling??
NO, one can use, but with caution. Make sure it is not leading to a systematic error

What after sampling?

- ❑ Ask, why did we do sampling? Objective is to learn about the population
- ❑ Statistical Inference - Learn about parameters from sample statistic
- ❑ Usually, the quantities of interest are mean and proportion in the population (depending on the context)
- ❑ We deal with them separately

Estimating from the statistic

- The rational behind estimating is expectation of statistic should be equal to the population parameter
- Biased and Unbiased Estimator
- Variance of estimate should be minimized to the extent possible

Errors in the Process of Estimation

- **Sampling Error** - Because we are only considering a subset of population, the point estimate is rarely exactly correct. Unavoidable error, but we can estimate the error and hence have some control over it
- **Non-sampling Error** - If there is bias in the observations, or sampling wasn't done properly. Can't be dealt with mathematically. Should be avoided

Estimating Population Mean from Sample Mean

1. Let the true values in the population be $A_1, A_2, A_3, \cdots, A_N$
2. Population mean is denoted by μ and equals $\dfrac{\sum_{i=1}^{N} A_i}{N}$
3. Also, population variance is denoted by σ^2 and equals $\dfrac{\sum_{i=1}^{N} (A_i - \mu)^2}{N}$
4. Let the sample be a SRS of size n
5. Observations are $X_1, X_2, \ldots, X_n$
6. Sample mean is denoted by $\overline{X}$ and defined as follows $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$
7. $E(\overline{X}) = \mu$

$Var(\overline{X}) = \frac{\sigma^2}{n}$

Estimating Population Variance from Sample Observations

- ❑ If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- ❑ If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N-1}{N}\right)\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Check $n-1$ using a dataset

Why is estimating population variance important?

Estimating Population Variance from Sample Observations

- If the sampling scheme is WITH REPLACEMENT

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N-1}{N}\right)\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Check $n-1$ using a dataset

Why is estimating population variance important?

To get an idea about error in estimation of sample mean

Standard Error in Sample Mean

- If the sampling scheme is WITH REPLACEMENT
  $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$
- If the sampling scheme is WITHOUT REPLACEMENT
  $$\mathrm{Var}(\overline{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$
- $\frac{N-n}{N-1}$ is called the finite population correction
- Typically, can be ignored if sampling fraction $\frac{n}{N} \leq 0.05$
- Standard deviation of $\overline{X}$ is called the Standard error of the sample mean
- Do we know σ^2?? What is the remedy??

Central Limit Theorem

If 'n' is large,

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

- ❑ Can we do better about our inference from sample mean??
- ❑ Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Central Limit Theorem

- Can we do better about our inference from sample mean??
- Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?
- Thank you Central Limit Theorem

Central Limit Theorem

Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean $\overline{X}$ is approximately normal with

1. *mean* = μ
2. *variance* = $\frac{\sigma^2}{n}$

What is meant by large n? Typically, $n \geq 30$

Comments about Central Limit Theorem

1. Don’t use CLT when population size is not too large compared with sample size, CLT is approximate result
2. If sample size is large and population size is also much larger as compared to population, one can use CLT even when the sampling scheme is Without Replacement

CLT when population is normally distributed

1. If underlying population is normally distributed, CLT is not required
2. If population is approximately normal, then $\overline{X}$ is approximately normal even for small n

Confidence Interval

$Y \sim$ Normal (μ, σ^2)

$P(\mu - 2\sigma \le Y \le \mu + 2\sigma) =$

Back to Sample Mean $\overline{X}$

1. $\overline{X}$ is a random variable
2. Under certain conditions, large sample size, etc. We use CLT to get better idea about $\overline{X}$
3. Using properties of normal distribution, what can be said about
4. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
5. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.

$\overline{X} \sim$ Normal $\left(\mu, \frac{\sigma^2}{n}\right)$ → Approximate

6. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$

$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 95.44\%$

7. Or, by rearrangement of terms, $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
8. Magic here, we have created an interval for μ
9. This is nothing but the confidence interval

Confidence Interval

$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 9544\%$$

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\underline{\mu - 2\frac{\sigma}{\sqrt{n}}} \le \overline{X} \le \underline{\mu + 2\frac{\sigma}{\sqrt{n}}}\right) = 0.9544$
4. Or, by rearrangement of terms, $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$

$$\left[\mu - \frac{2\sigma}{\sqrt{n}} \le \overline{X} \;\&\; \overline{X} \le \mu + \frac{2\sigma}{\sqrt{n}}\right]$$

5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

$$\left[\mu \le \overline{X} + \frac{2\sigma}{\sqrt{n}} \;\&\; \overline{X} - \frac{2\sigma}{\sqrt{n}} \le \mu\right]$$

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms, $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

$\mu = 100.$

$[95, 99]$

$[99, 110]$

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms, $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

1. Would interval estimate make more sense?
2. How to make intervals??
3. Can we make use of CLT?

Confidence Interval

1. Confidence Interval typically computed by adding and subtracting standard error in the point estimate
2. Point estimate +/- some multiple of Standard error
3. What is confidence level?

Confidence Interval Idea

1. Confidence Interval for any parameter of the population is a random interval that contains the parameter with some probability
2. If sampling is done repeatedly and confidence intervals are constructed, a 95% confidence interval will contain the values about 95% of the times
3. Typically, significance level is used, denoted by α
4. What is confidence level?

Confidence Interval Idea

1. 95% is the confidence level of the interval generated
2. We pick a sample, construct the interval. Can we say that the probability that the interval contains the true value is 0.95??
3. Different schools of thought, most don't agree on the above made statement
4. But, everyone agrees on the fact that confidence is on the procedure used to construct the confidence interval

Example of Confidence Interval

A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.

Sample Proportion

❑ Sometimes, one is interested in estimating population proportion

❑ What is the proportion of IBM312 students who like statistics?

❑ One can attempt the answer to this using sampling

Sample Proportion

❑ Can we make use of results from sample mean?

Sample Proportion

- Can we make use of results from sample mean?
- If the i^{th} respondent says YES, model it as $X_i = 1$
- If the i^{th} respondent says NO, model it as $X_i = 0$
- Denote by n_{YES} and n_{NO} are the responses in the sample of size n
- Denote by N_{YES} and N_{NO} are the actual values in the population of size N

Sampling Proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- $E(\hat{p}) = p$. Do we need to prove this??
- $\mathrm{Var}(\hat{p}) = \dfrac{p(1-p)}{n}$. Why??
- Is p known?
- State CLT for sample proportion
- Additional conditions - $np \geq 10$ and $n(1-p) \geq 10$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- What kind of random variable is n_{YES}??
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \frac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\frac{n_{YES}}{n}\right) = \frac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\frac{n_{YES}}{n}\right) = \frac{Var(n_{YES})}{n^2} = \frac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Confidence Interval Discussions

- Can you also do similar calculations and make a confidence interval for Population proportion? (Hint - Use CLT and our remark that sample proportion can be given a similar treatment as sample mean)
- Khan Academy Video
  `https://www.youtube.com/watch?v=bGALoCckICI`
- Which is bigger - 99% confidence interval or 95% confidence interval?

Summary of results for 100(1-α)% C.I.

n	σ^2	C.I. Type	Symmetric C.I.
Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ, s is sample standard deviation

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

Sample Size Determination

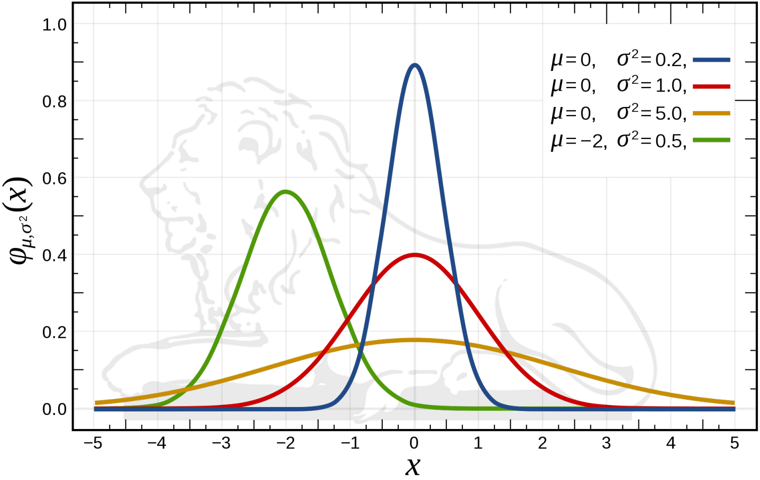
- A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- We want to repeat this study, how many students should you survey so that the 99% confidence interval's width is no more than 2 minutes?

Basics about Random Variable

- A variable whose value depends on outcome of a random phenomenon
- A random variable is characterized by its distribution
- Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- Similarly, any other algebraic operation of two or more random variables also remain a random variable
- If X and Y are random variables, $X + Y, X - Y, XY, \frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



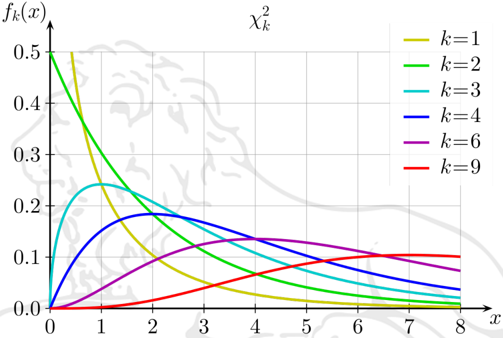
Source - `https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution`

Chi-square distribution

If $Z_1, Z_2, \ldots, Z_n$ are all independent and normally distributed with mean 0 and variance 1, then $U = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

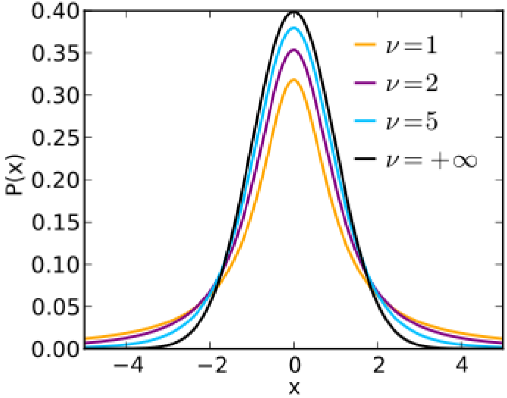
Chi-square distribution



Source - `https://en.wikipedia.org/wiki/Chi-squared_distribution`

t-distribution

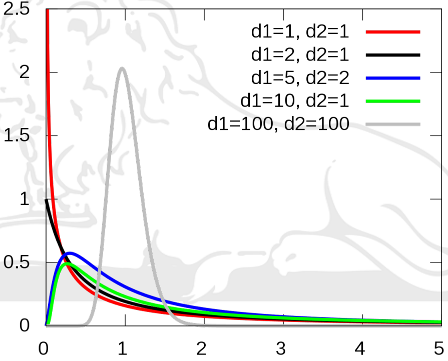
If $Z \sim N(0,1)$ and $U \sim \chi_n^2$ are independent, then the distribution of $\dfrac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom



Source - `https://en.wikipedia.org/wiki/File:Student_t_pdf.svg`

F distribution

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then the distribution of $\frac{\frac{U}{m}}{\frac{V}{n}}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



Source - `https://en.wikipedia.org/wiki/F-distribution`

Application of distributions that we just saw

- Does having the idea of population distribution itself a useful information?
- If yes, how do we make use of it?
- Let us concern ourselves with sample mean
- Assume you have the information that the population distribution is normal.
- How do you use this??
- Is CLT required??

Case of normal population

- ❑ Sample mean is denoted by $\overline{X}$ and defined as follows

  $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- ❑ If the sampling scheme is WITH REPLACEMENT, sample variance equals

  $$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- ❑ It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$

- ❑ Can you guess the distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

- ❑ If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0, 1)$. So what is it then??

- ❑ Distribution of $\frac{\overline{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$

- ❑ Hence, even for small sample size, we can make confidence

Case of normal population

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\dfrac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\dfrac{\overline{X}-\mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Case of normal population

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Moral of the story - results for 100(1-α)% C.I. for Mean

Population distribution	n	σ^2	C.I. Type	Symmetric C.I.
Any	Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Any	Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$
Normal	Any	known	Exact	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Normal	Any	unknown	Exact	$\left(\overline{X} - \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ, s is sample standard deviation

Typically, $t_{n-1,\frac{\alpha}{2}}$ is used only for small n, because for large n, $z_{\frac{\alpha}{2}}$ gives a good approximation

Moral of the story - results for 100(1-α)% C.I. for Proportion

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

For case of proportion, it is advised to use these formulae only when apart from n being large, $n\hat{p} \geq 10$ and also $n(1-\hat{p}) \geq 10$