# Statistical Inference: Sampling and Confidence Interval Estimation

Prof. Abhinava Tripathi

# Introduction: Statistical Inference

# Introduction

- Since one does not have the luxury of working with populations, one has to make only inferences and exact solutions or estimates are not available
- Let us start with a simple example from manufacturing industry
- You are working as a part of food regulator to examine the quality of food
- You can not go to all factory outlets and check each packet
- A feasible way is to take small sample that is representative of the population to make appropriate inferences

# Introduction

- Assume that the company has 30000 packets out of which you select 100 samples
- You find that the led content in these packets is 2.2 ppm with a standard deviation of 0.7 ppm
- Can we say that the population mean and standard deviation parameters would be same as the sample
- Is it possible that these sample parameters would be very different from population parameters

# Types of Sampling: Probability Sampling

# Introduction to Sampling

- Good sampling is important to make better inferences about the population parameters
- In the food sample problem, suppose that you pick all the 100 samples from a single factory
- It may be possible that the higher lead content is specific to that this factory
- This requires that sampling procedure is fair and unbiased so that inferences are accurate

# Simple Random Sampling

- Let us discuss some of the ways in which we can select a sample of 100 noodle packets
- One, though very less efficient way, is to collect all the 30000 packets randomly and select 100 out of these
- This is called simple random sampling
- This is like a blindfolded person picking sample units from population: the process is completely random
- Let us discuss this in more detail

# Stratified Sampling

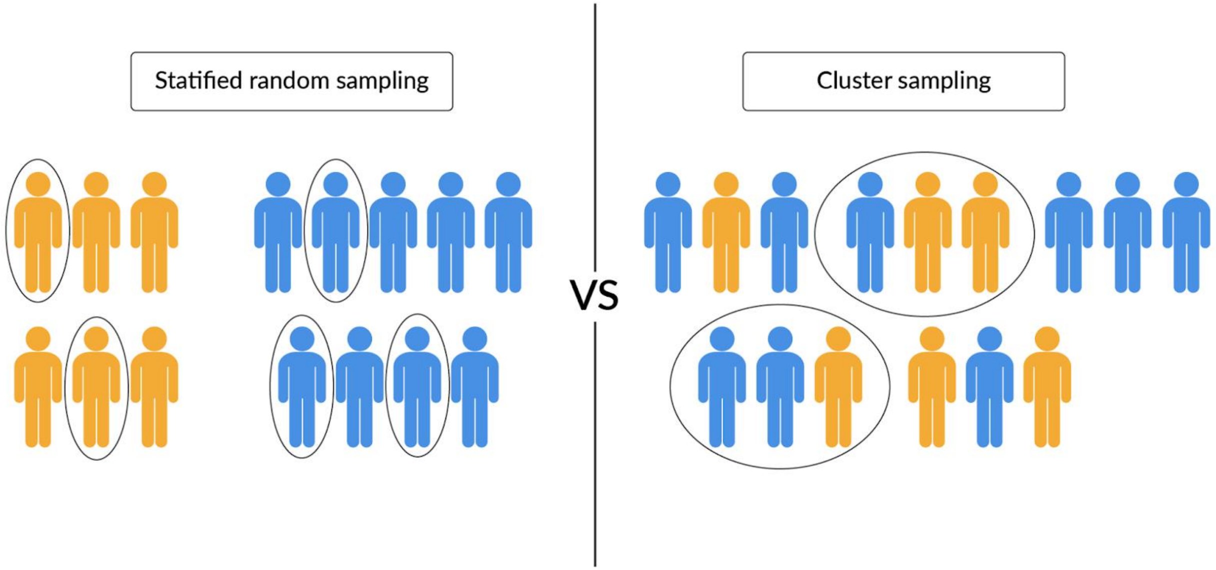
- In the previous example, suppose that 70% of the noodles are from factory A and 30% from factory B
- You sample 70 packages from factory A and 30 packages from factory B randomly
- This kind of sample is expected to be more representative of the population
- It carries proportions of packets from factory A and B, similar to that in the population
- The units are divided into homogeneous stratums (sub-groups) and then samples are taken randomly
- This approach is known as Stratified random sampling

# Cluster Sampling

- In the previous example, suppose that there are 20 warehouses in the country
- You would not like to collect your sample from each warehouse (i.e., consider each warehouse as cluster)
- You can select 3-4 warehouses as clusters (may be through random sampling) and then consider desired number of samples from these clusters
- Cluster sampling is usually used when you see that the population can be divided into different groups or clusters that have different characteristics
- Then you do sampling from dissimilar clusters

# Cluster Sampling vs. Stratified Sampling

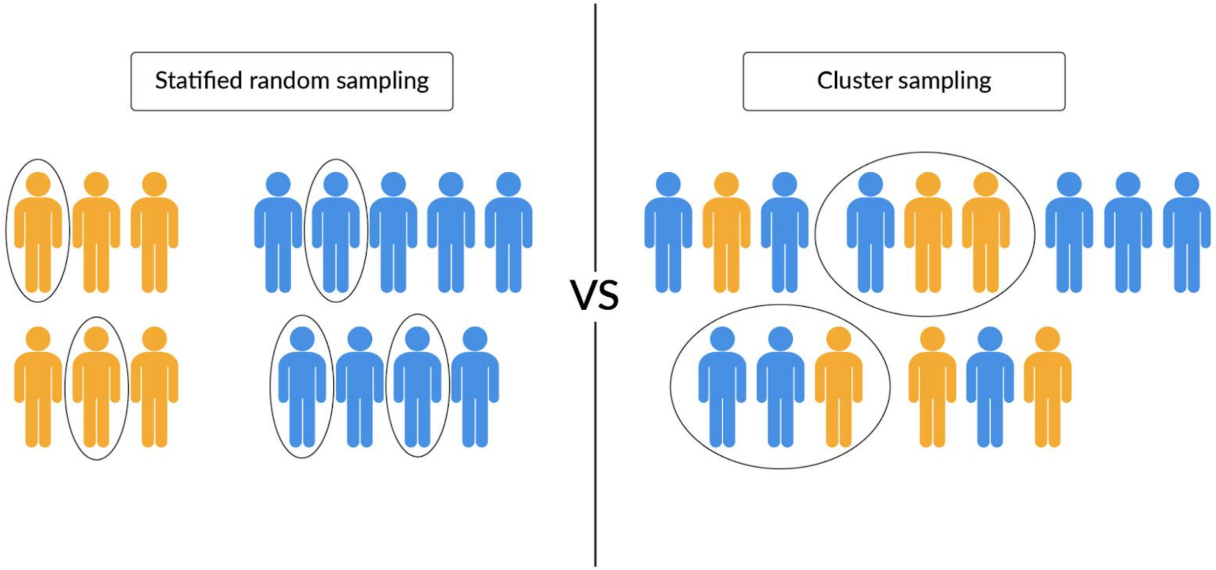
- One may get confused between stratified sampling and cluster sampling
- Since, in both cases we divide populations in sub-populations




- In stratified sampling, we divide the population into sub-populations and then select the sample units in the same proportion as the sub-populations so that the sample is as representative as the parent population

# Cluster Sampling vs. Stratified Sampling

- In cluster sampling also, we divide the population into sub-population
- But here we only study the selected clusters, not all the clusters



# Systematic Sampling

- Let say you label the samples and select every third packet starting from the second packet, as shown in the figure here
- We selected a random starting point and started picking out sample units at some fixed and periodic interval
- This is called systematic sampling

# Sampling methods

- We studied four kinds sampling methods
  - Simple random sampling
  - Stratified random sampling
  - Cluster sampling
  - Systematic sampling
- Which kind of sampling is more suitable for our example
- For many of the food regulators, stratified sampling and simple random sampling is considered as more suitable

# Heterogeneous population

- What could be those cases where population is not heterogeneous in nature
- If all the noodle packets are of the same nature and manufactured in a single factory, then simple random sampling would be the most straightforward
- If the packets came in different flavors and were manufactured in different factories, then ideally stratified sampling would be the recommended method
- All these sampling techniques fall under the category of probability sampling
- In these sampling techniques, every unit of the population has a certain known chance of being included in the sample

# Types of Sampling: Non-Random Sampling

# Non-Random Sampling: Convenient sampling

- There is another sampling method called as non-random sampling
- Here, the odds of a sample unit getting selected can not be calculated
- Convenient sampling: You choose 100 packets that were closer to you and most easily available
- This sampling method is based on the convenience of the person selecting the sample
- This method has a high probability of being biased

# Non-Random Sampling: Judgement sampling

- Judgement sampling: It is done on the basis of the knowledge and judgement of the person who is selecting the person
- Often the survey questions and responses require highly specialized skillset
- In this discussion, we learned about two sampling techniques that fall under non-random sampling
- In these methods, it is often important to understand the implication of sampling techniques on the nature of sample being acquired

# Statistical Inference I: Central Limit Theorem

# Introduction

| Parameter | Population | Sample |
|---|---|---|
| Size | N | n |
| Mean | μ (or $\bar{X}$) | $\bar{x}$ |
| Sigma | σ | s |
| Variance | $\sigma^2 = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{(N)}}$ | $s^2 = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{(n-1)}}$ |

- In the previous example, 30000 packets is called the population, and the small collection to be examined is called sample

- The population size is denoted by capital N, its mean by μ and its standard deviation by σ

- The sample size is denoted by a lowercase n, and the mean by $\bar{x}$

# Introduction

- Remember that in our noodle packet example, we wanted to validate whether our sample mean (2.2 ppm) was a true representation of the population
- It is impossible to exactly find the population mean from sample mean with zero error
- All we can say is that population mean will be within 2.2 plus minus some error
- If we are able to estimate that error, let us say 0.2 ppm; then still we are able to add some value to analysis
- We know that the levels of led will be less than 2.5 ppm and more than 2 ppm

# Introduction

- How to establish if the sample is indeed a true representation of the population
- Consider that you have all the N=30000 packets, that is the population data
- If the mean of this data is $\mu$= 2.199 and the standard deviation is $\sigma$=0.132; these are essentially population parameters
- Let us now consider a sample of size 5 with a mean $\bar{x}$=2.145
- In another sample, the mean comes out to be 2.27 ppm
- So instead of 2, we will choose 100 samples with a sample size of 5

# Central Limit Theorem (CLT)

- If we plot these sample means on a graph, it will look like a normal distribution with a center point around 2.2 which is close to population mean
- If the sample size is increased the distribution keeps getting closer and closer to normal distribution
- Moreover, as the sample size increases to more than 30, the mean of the sample distribution approaches the population mean
- This experiment is the basis for central limit theorem

# Central Limit Theorem (CLT)

- The central limit theorem states that when you take a large number of samples, the mean of the sampling distribution thus formed, will be approximately equal to the population mean

- The second part of the theorem states that the standard deviation of this sampling distribution will be equal to sigma, which is our population standard deviation, divided by the square root of n where n is the sample size

- Finally, the central limit theorem states that if the sample size that you take is greater than 30, the sampling distribution will become normally distributed

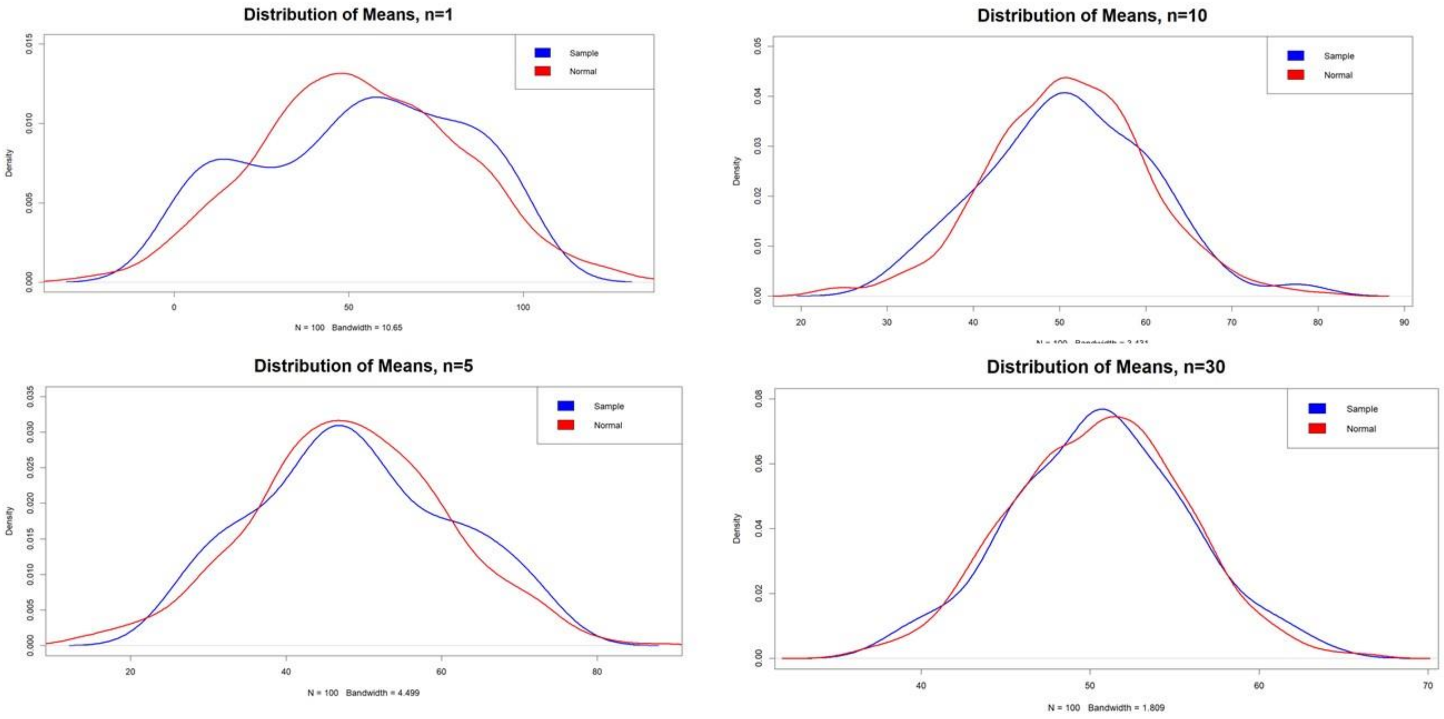
# Central Limit Theorem (CLT)

- The central limit theorem states that when you take a large number of samples, the mean of the sampling distribution thus formed, will be approximately equal to the population mean
- The second part of the theorem states that the standard deviation of this sampling distribution will be equal to sigma, which is our population standard deviation, divided by the square root of n where n is the sample size
- Finally, the central limit theorem states that if the sample size that you take is greater than 30, the sampling distribution will become normally distributed

# Central Limit Theorem (CLT)

- To concretize your understanding of the central limit theorem,  let's try and visualize the central limit theorem.

- If you plot the distribution of sample means, or what is also known as the sampling distribution, then this distribution approaches a normal distribution



# Central Limit Theorem (CLT)

- Notice as the sample size increases how the sample distribution follows the postulations of CLT
- As you increase the size of n, so basically you are bringing the n value closer to the population size value
- The sample mean approaches the population mean as the sample size is increased
- What will happen if you increase the sample size to n=50 or even higher

| Parameter | Population | Sample (n=1) | Sample (n=5) | Sample (n=10) | Sample (n=30) |
|---|---|---|---|---|---|
| Mean | 50.78 | 52.97 | 51.80 | 50.45 | 50.60 |
| SD | 28.80 | 29.72 | 12.91 | 9.32 | 5.30 |
| $28.80/\sqrt{n}$ | | 28.80 | 12.88 | 9.11 | 5.26 |

# Central Limit Theorem (CLT)

- We started with absolutely random population with a mean μ and standard deviation of σ
- We then took a large number of samples of a particular sample size and plotted the means of these samples
- We varied the sample sizes and observed the behavior of resulting sampling distribution vis-à-vis normal distribution
- As the sample size increased, the probability distribution became close and closer to normal distribution, and the mean of sample converged to the mean of population

# Statistical Inference II: Introduction to Confidence Intervals

# Introduction to Confidence Intervals

- Let us go back to the noodle example, and derive conclusions about the population using the sample
- We took a sample of 100 packets and found out that its sample mean was 2.2 ppm and standard deviation was 0.7 ppm
- We will make use of sampling distribution properties
- Sampling distribution is nothing but the distribution of all the possible sample means that can be generated from this population

# Introduction to Confidence Intervals

- With the help of CLT, we have some idea about the properties of this sampling distribution
- If the sample size is greater than 30, then the sampling distribution is normally distributed with a mean equal of population mean and a standard deviation equal to population SD divided by square root of sample size
- We do not know the exact population mean and standard deviation
- There are cases where you have some idea about the population standard deviation, and in some cases you don't, and you employ sample SD for that

# Introduction to Confidence Intervals

- Sample standard deviation (s=0.7) and n=100, we get the SD of sampling distribution as 0.7/10= 0.07
- Here, we are using the sample standard deviation as the substitute for population standard deviation
- Now we will make use of normal distribution properties as elaborated earlier (1-2-3 rule)
- For example, using this rule, we can say that the probability that sample mean lies from $\mu$-2*0.07 to $\mu$+2*0.07 is 95%

# Introduction to Confidence Intervals

- While we do not know the population mean 'μ', but we know the population standard deviation 0.07
- Rearranging this a little bit, we can say that P(2.2-2*0.07 to 2.2+2*0.07 )= 95%
- Or the probability that population mean 'μ' lies in the interval P(2.2-2*0.07 to 2.2+2*0.07) is 95%
- Or you can say with 95% probability that the mean will lie between 2.06 ppm to 2.34 ppm

# Introduction to Confidence Intervals

- The probability associated with this claim is called the confidence level
- Since we are concluding about the population mean with 95% probability, we can say that the confidence level is 95% or alternatively level of significance or alpha value =5% (i.e., 1- confidence level)
- Next, you have the margin of error, which is the maximum error= 2*0.07=0.14
- Final the interval of values or the confidence interval= 2.06 to 2.34
- Since the upper bound of the confidence is less than 2.5, we can conclude with 95% confidence that noodles do not contain led content that is more than the prescribed limit of 2.5 ppm

# Statistical Inference III: Confidence Interval Construction

# Confidence Interval Construction

- Till now we have understood estimation of population mean with the construction of unbiased confidence interval
- Often getting population data is not feasible and you need to rely on inferential statistics
- The objective here is to estimate population mean; the population need not be normal
- To solve the problem we start with the sample, using appropriate sampling technique

# Confidence Interval Construction

- You select a sample of small size=n and calculate the mean of the sample $\bar{x}$ and sample standard deviation ‘s’
- To solve this problem, let us recall central limit theorem (CLT)
- CLT suggests that sampling distribution will behave like a normal distribution as you increase the sample size (>30), with a mean of μ (population mean) and SD of $\sigma/\sqrt{n}$
- Using these values, we will estimate the population mean μ

# Confidence Interval Construction

- If we consider a confidence level of y% and apply the CLT, we can estimate that population mean lies in the range: $\bar{x} - z * \frac{s}{\sqrt{n}}$ to $\bar{x} + z * \frac{s}{\sqrt{n}}$ ; where z is the critical value associated with y% confidence level
- For confidence levels 90%, 95%, 99%, the values are 1.65, 1.96, 2.58
- You want to be highly confident in the noodle packet example, and 99% confidence makes more sense, or may be you have higher tolerance levels and ok to go ahead with 90% confidence levels

# Confidence Interval Construction

- Collect a sample of n>=30 from the population
- Compute the mean and standard deviation of the sample
- Based on the CLT, assume that the sampling distribution is normal with a mean same as the population mean and SD which is same as population SD divided by square root of n; population SD is proxied by sample SD
- Select the appropriate confidence level and based on that and decide the appropriate confidence interval : $\bar{x} - z * \frac{s}{\sqrt{n}}$ to $\bar{x} + z * \frac{s}{\sqrt{n}}$

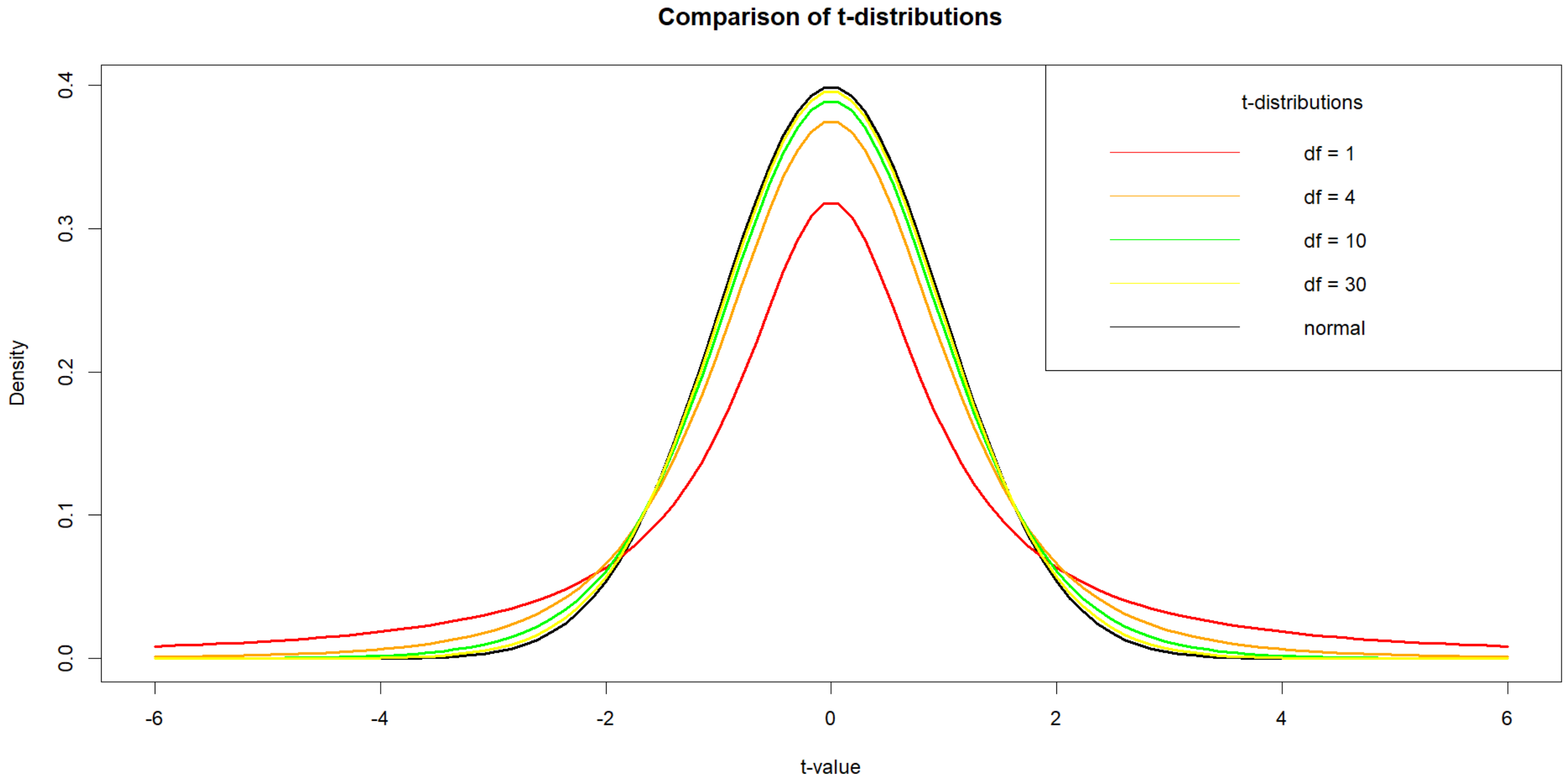
# Statistical Inference IV: Interval Estimation for Small Samples

# Interval Estimation for Small Samples

- Often times, large samples are not available and one has to work with small samples
- For example, you are in pharma company and in a medicine trial you only have 15 volunteers
- In such cases with less than 30 sample size, you work with t-distribution, where population SD is not known and the same is proxied using the sample standard deviation
- A t-distribution is similar to z-distribution only that it has shorter peak and wider tails

# Interval Estimation for Small Samples

- A t-distribution is similar to z-distribution only that it has shorter peak and wider tails



# Interval Estimation for Small Samples

- You work for a pharma company and are testing the effects of a medicine on 15 volunteers
- The medicine increases the presence of a particular hormone XYZ in a patient's blood, by 10.038 micro units
- We estimate the population SD using sample SD= 0.072
- We will use the procedure similar to interval estimation using Z-distribution
- But due to sample size restrictions, we will use t-distribution

# Interval Estimation for Small Samples

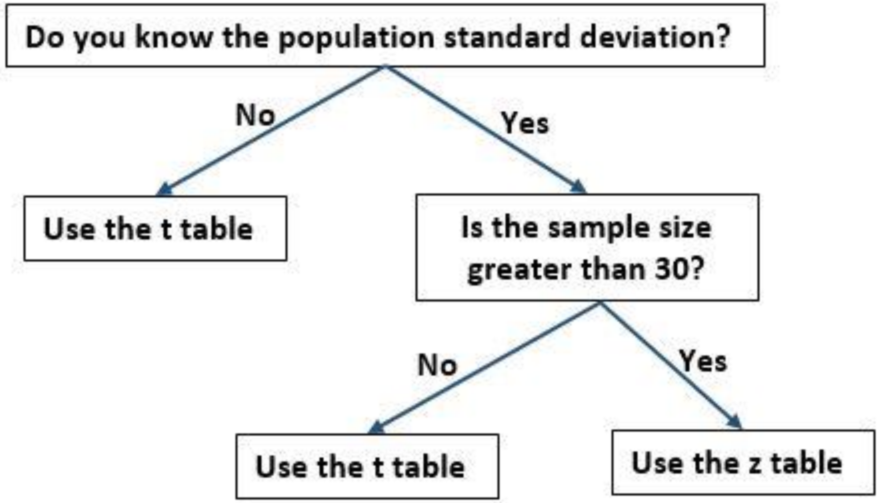
- Population SD is proxied using sample SD; sample mean $\bar{x}$ =10.038 and sample SD=0.072
- In the sampling distribution we are assuming t-distribution; each t-distribution is distinguished by its degrees of freedom
- For a sample size of ‘n’, the corresponding degrees-of-freedom (DOF) would be ‘n-1’
- For smaller sample sizes, t-distributions are flatter than for larger sample sizes
- For large DOF, t-distribution is similar to the standard normal distribution (at sample size n=30)

# Interval Estimation for Small Samples

- For this case (n=15) DOF=14, and we will use t-distribution
- We select a confidence level of 95%; the relevant confidence interval will be: $\bar{x} - t * \frac{s}{\sqrt{n}}$ to $\bar{x} + t * \frac{s}{\sqrt{n}}$; the corresponding t-value is 2.145
- The lower bound is 10.038-2.145*0.072/sqrt(15)=9.998
- The upper bound is 10.038+2.145*0.072/sqrt(15)=10.077
- So the 95% confidence interval lies in the range of 9.998 to 10.077
- t-distribution is preferred when sample size is small and population SD is unknown

# Interval Estimation for Small Samples

- t-distribution depends on degrees-of-freedom (df) or sample size -1
- For a large sample size, both the t-distribution and normal distribution
- The decision rule flowchart to use the t-distribution and z-distribution is provided below
- If the population standard deviation is unknown and the sample size is greater than or equal to 30, then the z distribution is preferred over the t distribution



# Interval Estimation for Small Samples

- If the sample size is less than 30, then even if the population standard deviation is known, it is best to use the t-test as it is ideally suited to dealing with small samples
- The lower and upper bound is given by $\bar{x} - t * \frac{s}{\sqrt{n}}$ to $\bar{x} + t * \frac{s}{\sqrt{n}}$; using this we can estimate the confidence interval

# Statistical Inference V: Interval Estimation for proportions

# Interval Estimation for proportions

- Many times the values are categorical in nature: for example, in an exit poll survey, a sample of people voted one of the two parties
- How to extrapolate this value to the entire population, given that sample mean and standard deviation driven approaches are not valid
- Consider for example, you are working as part of a political science company that specializes in voter polls and designs surveys to keep political office seekers informed of their position in a race

# Interval Estimation for proportions

- Through these surveys you found that 220 registered voters, out of 500 contacted, favor a particular candidate. You want to develop 95% confidence interval estimate for the population of registered voters
- The data is categorical in nature: Voted for a party or not voted
- The proportion of voters who voted is $\bar{p}$ =220/500=0.44; we want the confidence interval around this proportion (e.g., 0.43 to 0.45)
- The approach to estimate the confidence interval remains the same

# Interval Estimation for proportions

- Step 1: is to collect a sample of size n=500
- Step 2: Since data is categorical, we computed the proportions ($\bar{p}$=0.44)
- Step 3: Here we generate the sampling distribution of sample proportions and then find the interval estimate
- For being able to apply the sampling distribution of sampling proportion: n*p >5 and n*(1-p)>5;
- The best estimate of population proportion p here is the sample proportion $\bar{p}$=0.44

# Interval Estimation for proportions

- Since, n=500 here, therefore np=220 and n*(1-p)= 280
- Both of these values are considerably greater than 5, so we can assume that sampling distribution follows normal distribution and go ahead with the formula for 95% confidence interval estimation
- The appropriate confidence interval here is $\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{(1-\overline{p}) * \overline{p}}{n}}$; here SD is taken as $\sqrt{\frac{(1-\overline{p}) * \overline{p}}{n}}$

# Interval Estimation for proportions

- In this case, $\bar{p}$ =0.44, n=500, z=1.96
- Lower limit= 0.44-1.96*sqrt(0.44*(1-0.44)/500)=0.44-0.0435=0.3965
- Upper limit= 0.44+1.96*sqrt(0.44*(1-0.44)/500)=0.44+0.0435= 0.4835
- The aim of the problem is to be able to estimate an interval around the sample proportion $\bar{p}$

# Thanks!