Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

BU7142 Foundations of Business Analytics

Lecture 4

Sampling

Dr Yufei Huang

Contents

- Sampling
- Central Limit Theorem
- Confidence Interval
- Sampling Techniques

Sampling

The Purpose of Statistics

- Usually, we are interested in properties of the population.
- Examples
  - How many smart phones on average each person has in UK.
    - Population: people
    - Property: average number of cell phones per person
  - Would women like the design of a new dress?
    - Population: women
    - Property: rating of the design of a dress
  - How many cheese sticks children eat a day?
    - Population: children
    - Property: number of cheese sticks consumed per day

The Challenge

- It is usually impossible to measure the property of all members of population.

- Why?
  - Time
  - Money
  - Not all people agree to participate in research
  - Is it really needed?!

The Solution: Sampling

1. Sample people from the required population

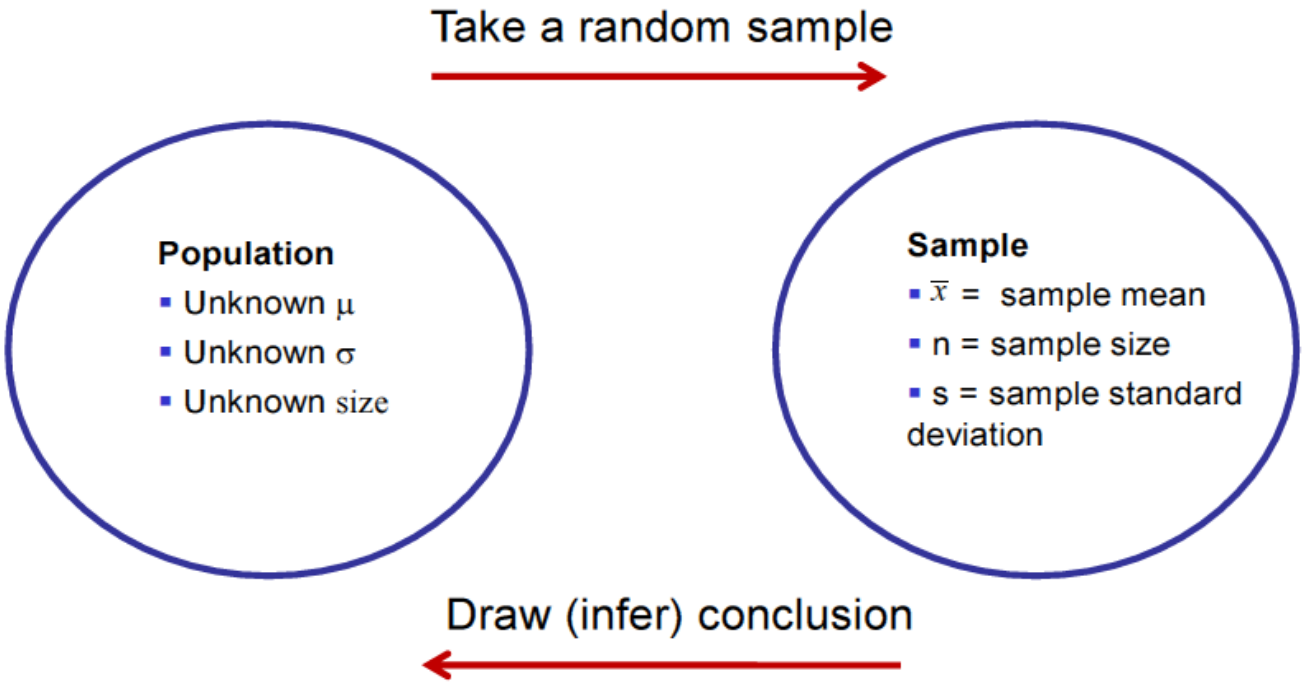
2. Measure the property in the members of the sample

3. Try to draw conclusion on the population from the results of the sample.

A subset of the population.

Point Estimates

- Sample statistic: a numerical measure of the sample.
  Population parameter: a numerical measure of a population.



- We will try to estimate the population parameters using the sample statistics.

Sampling

- Does the way we sample the population matter? Yes
- Example: You are interested in knowing whether people like pizza. You choose the following samples.

- Choosing different samples, may give you different results.
- So how to do sampling?

Random Sampling

- We would like to have samples which resemble the population well
- To avoid biased sampling, we sample randomly from the population
- How?
  E.g. programming a computer to randomly pick numbers from a phone book.

Sampling Distribution

- Even if you attempt to sample randomly, each sample could give a different mean



- Definition. The sampling distribution of $\overline{x}$ is the probability distribution of all possible values the random variable $\overline{x}$ may take when a sample of size n is taken from a specified population.

Central Limit Theorem

- Assume that the mean of a certain property of a population is μ, and that its standard deviation is ϭ.

- Then the distribution of the sample mean $\overline{X}$ tends to be a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, as the sample size n becomes large.

- That is, for large n:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- Remark: We do not require that the population is normally distributed!

Sample Size

- What sample sizes are large enough for the central limit theorem?

$$n \geq 30$$

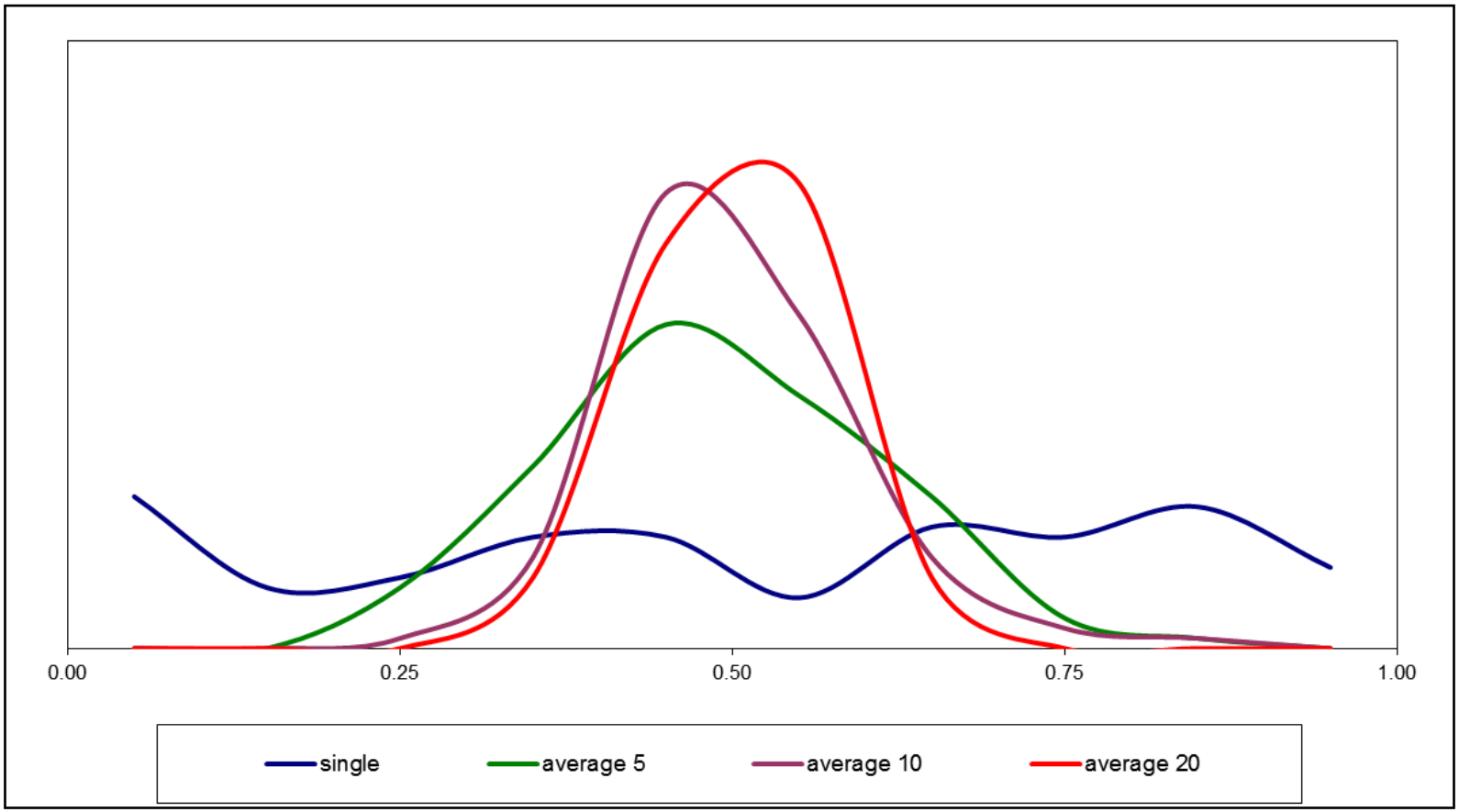
- Remark. Larger sample will give you better estimate of the population, i.e., smaller standard deviation of the sample mean distribution
- Remark. When sample size is small, sample mean follows t-distribution, if the population is normally distributed

Understanding Central Limit Theorem

- If we take many samples of a certain size from a population, each sample might have a different mean.

- However, as the n becomes large, the distribution of the means of these sample tends to a normal distribution.

Understanding Central Limit Theorem



https://www.youtube.com/watch?v=jvoxEYmQHNM

Example

- A company produces calcium-enriched cheese sticks. The company advertises that the mean amount of calcium in each cheese stick is 200 milligram and that the standard deviation is 15 milligram. A scientist of a consumer-right journal samples 100 cheese sticks. What is the probability that the sample mean will be less than 195 milligram?

- Solution. As the sample size is large (>30), by the central limit theorem,

$$\overline{X} \sim N(\mu, \sigma^2/n) = N(200, 15^2/100) = N(200, (15/10)^2).$$

$$P(\overline{X} < 195) = P\left(Z < \frac{195 - 200}{15/10}\right) = P(Z < -3.33) = 0.00043$$

Confidence Interval

- Definition. A confidence interval is a range of numbers, which is believed to include an unknown population parameter with a certain probability.

- Example:

  The price of gold fluctuates every day. We would like to find a price interval [a,b] for gold, such that the mean of the price of gold, μ, is in [a,b] with probability of 90%. If we find such an interval, we say that we are 90% confident that μ lied in [a,b]. And [a,b] is the 90% confidence interval for the price of gold.

Confidence Interval when Population Standard Deviation is Know

Developing a 95% confidence interval for μ – the population mean

By the Central Limit Theorem, the sample mean follows $N(\mu,\sigma^2/n)$, where n is the sample size

$\Rightarrow$ There is a 0.95 probability that the sample mean is in the interval

$$\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

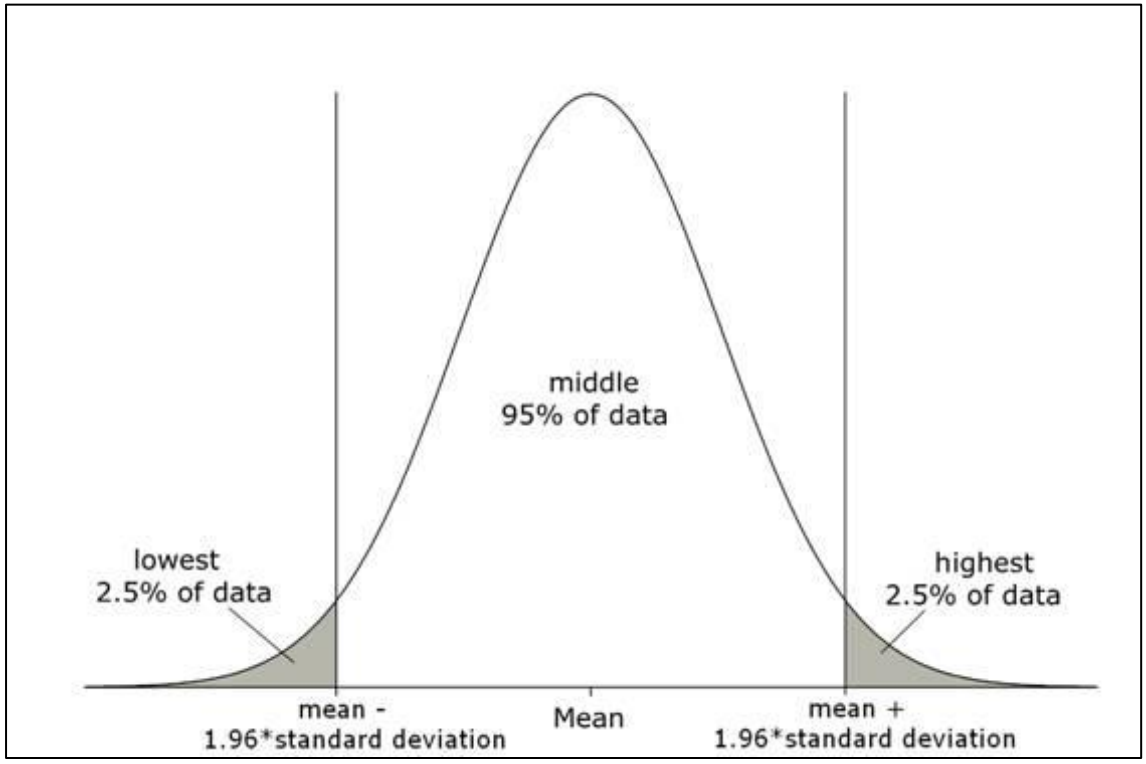
A 95% confidence interval for μ when σ is known and sampling is done from a normal population, or a large sample is used, is

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Linking back to Normal Distribution

- Reminder:

  $P(-1.96<Z<1.96)=2\cdot P(0<Z<1.96)= 2\cdot 0.4750=0.95$



- If we are given a value other than 95% we need to find the relevant z value.

Example

- The mean price of Seaweed, obtained by sampling it in 100 randomly chosen days, was £980. Assume that Seaweed's price follows a normal distribution with standard deviation £280. Construct a 95% confidence interval for the mean price of Seaweed.

- Solution. The required interval is:

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[980 - 1.96\frac{280}{\sqrt{100}}, 980 + 1.96\frac{280}{\sqrt{100}}\right] = [925.12, 1034.88].$$

Therefore a 95% confidence interval for Seaweed's price is [£925.12,£1034.88].

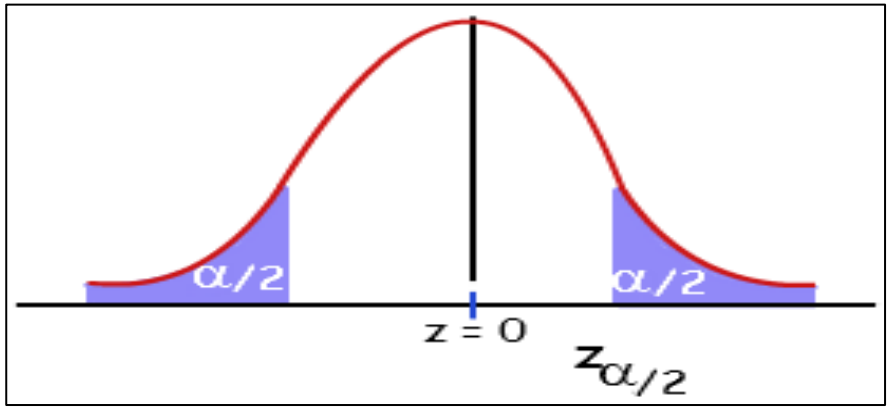
- Does the above confidence interval mean that 95% of all Seaweed's prices should lie in this interval?

  No, it is the interval for the mean.

Confidence Level other than 95%

- (1-α)·100% confidence level: α is the significance level

- Notation. We denote by $z_{\alpha/2}$ the z value that cuts off a right-tail area of α/2 under the standard normal curve.



- A (1-α)·100% confidence interval for μ when σ is known and sampling is done from a normal population, or with a large sample, is

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Example

- The mean price of a product, obtained by sampling it in 100 randomly chosen days, was £980. Assume that this price follows a normal distribution with standard deviation £280. Construct an 80% confidence interval for the mean price of this product.
- Solution. 1- α=0.8 → α=0.2 and α/2=0.1.

  We need to find $z_{\alpha/2} = z_{0.1}$.

  $P(0<Z< z_{0.1})=0.5-0.1=0.4 \rightarrow$

  $z_{\alpha/2}=1.28$

  The required interval is:

$$\left[\bar{x}-1.28\frac{\sigma}{\sqrt{n}},\bar{x}+1.28\frac{\sigma}{\sqrt{n}}\right]$$

$$=\left[980-1.28\frac{280}{\sqrt{100}},980+1.28\frac{280}{\sqrt{100}}\right]=[944.16,1015.84]$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Confidence Interval when Population Standard Deviation is Unknown

- Confidence interval for the population mean when the population standard deviation is not known

- In this case we use $\overline{X}$ and *S*

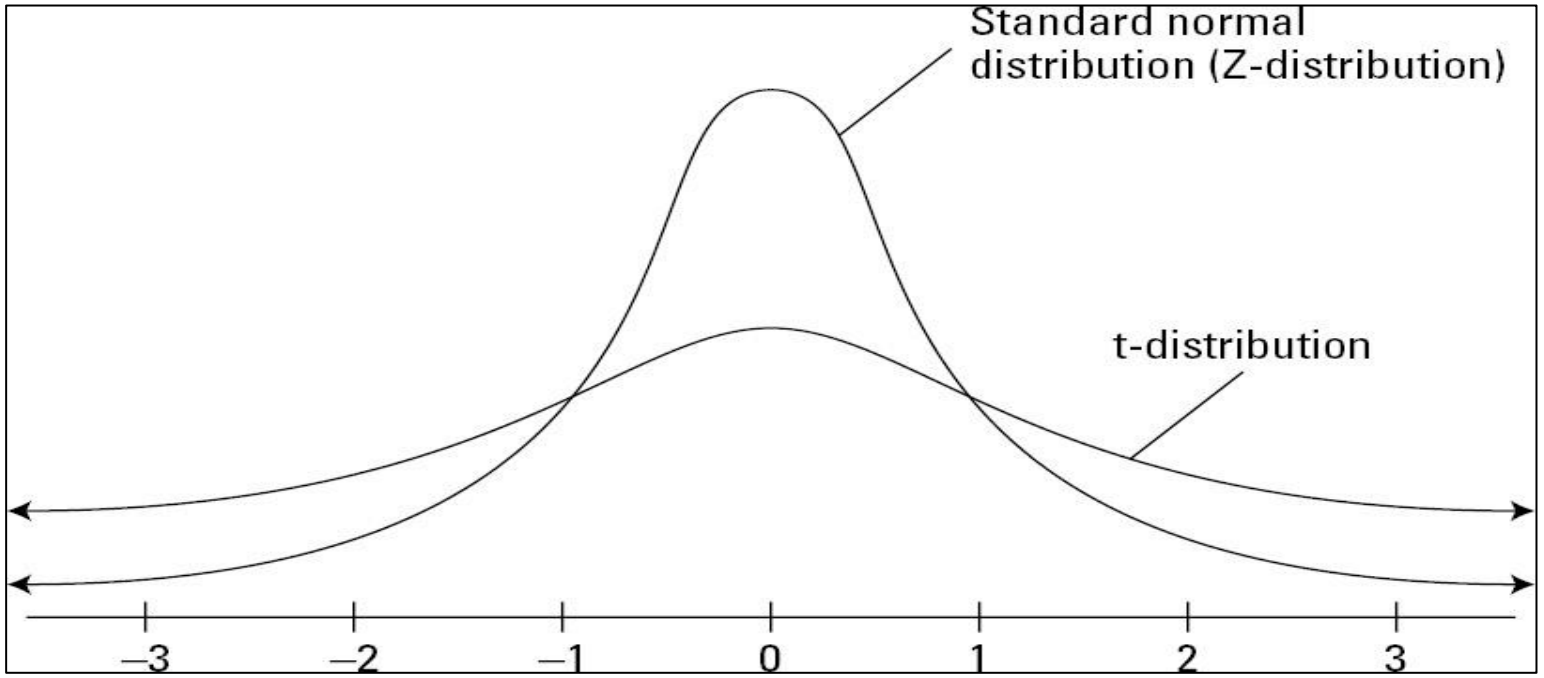
- However, it was shown that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is not normally distributed.

- It has a t-distribution (Student's distribution) , when the population is normally distributed:

$$t \sim \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

The t-distribution

- The mean of t-distribution: 0
- T-distribution depends on a variable called “degrees of freedom” (df).
  df is a measure of how well s estimates σ.
- For df>2, the variance of the distribution is df/(df-2).
- T-distribution approaches the normal distribution as the degrees of freedom (or the sample size) increases.



Confidence Interval when Population Standard Deviation is Unknow

- A (1-α)·100% confidence interval for μ when σ is not known is

$$\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right],$$

- Here $t_{\alpha/2}$ is the value of the t distribution with n-1 degrees of freedom that cuts off a tail area of α/2 to its right.

To calculate $t_{\alpha/2}$?

- T-distribution table.
- The table gives us the probability that a variable exceeds the numbers written in it.

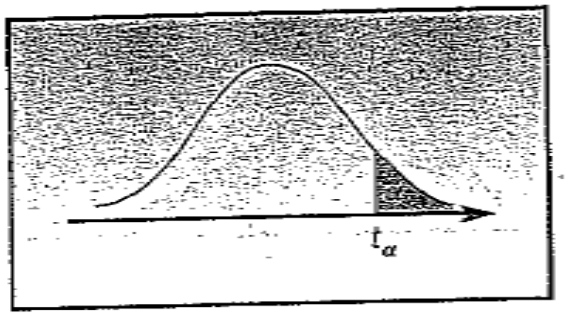


TABLE 6–1 Values and Probabilities of *t* Distributions

Degrees of Freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787

Examples

1. What is $t_{0.05}$ for a sample with 4 degrees of freedom?

   $t_{0.05} = 2.132$.

2. What is $t_{0.005}$ for a sample with 15 degrees of freedom?

   $t_{0.005} = 2.947$

3. A random variable with a t-distribution with 10 degrees of freedom has a 0.1 probability of exceeding 1.372.

TABLE 6–1 Values and Probabilities of t Distributions

Degrees of Freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787

Sampling Distribution of Sample Proportion

p population proportion
$\hat{p}$ sample proportion
n sample size

Expected value of $\hat{p}$ is $\mathrm{E}[\hat{p}] = p$

Standard deviation of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$

Example:

- proportion of people who voted for Brexit in the referendum
- proportion of international students at Trinity

Sample Proportion

- Suppose $np \geq 5$ and $n(1-p) \geq 5$.
- The sampling distribution of p approaches a Normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$ as the sample size becomes large.
- 95% confidence interval for population proportion

$$\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

- If p in unknown, then 95% confidence interval is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sampling Techniques

Sampling Techniques

- Simple Random Sampling (SRS)
- Stratified Sampling
- Quota Sampling
- Systematic Sampling
- Cluster Sampling
- Multi-stage Sampling

Simple Random Sampling (SRS)

- Every member of the population has an equal probability of being selected
- Each individual is chosen randomly and entirely by chance
- Advantages
  - It's a yardstick against which the efficiency of other sampling methods are compared.
  - It's probability based sampling.
- Disadvantages
  - It requires a sampling frame (listing individuals)
  - It's often impractical or expensive for large, dispersed populations.
  - It does not guarantee adequate representation for all sub-groups in the population.

Stratified Sampling

- Designed to ensure that sub-groups are adequately represented in the sample.

- Method

  Step 1. Divide the frame into appropriate sub-groups (= strata).

  Step 2. Randomly sample individuals from each stratum.

- Advantages
  - It's usually more reliable than SRS.
  - It's a probability based sampling.

- Disadvantages
  - Requires a frame within which each sub-group can be identified.
  - Needs prior knowledge of the population to determine the choice of strata.
  - Can be expensive or impractical for large dispersed populations.

Stratified Sampling: Example

Example: Age of employees at a company

Age	% of employees
under 21	10
21 to 40	50
41 and over	40
	100

How would you select a stratified sample of 50 people from this population?

Quota Sampling

- Like stratified sampling –it attempts to reflect sub-groups in the population
- Give interviewers quotas of different types of people to interview

Example: Age of employees at a company

Age	% of employees
under 21	10
21 to 40	50
41 and over	40
	100

Quota Sampling

- Advantages
  - It's relatively cheap and does not require a sampling frame

- Disadvantages
  - You have to be around when the survey is taking place to have a chance of being in the sample.
  - Sample may be biased towards particular types of individual.
  - The interviewer – not chance – decides who is surveyed (non-probability based).

Systematic Sampling

- Every k^{th} member of the frame selected after a random start

  E.g. To obtain a 1/20 sample from an electoral register:

  1. Pick a random no. between 1 & 20 to select first name.
  2. Thereafter, select every 20^{th} name from the list.

- Advantages
  - It's simple to implement.
  - Equivalent to SRS if members appear in frame in random order.

- Disadvantages
  - Requires a frame.
  - Danger of reflecting hidden periodicities in the frame.

Cluster Sampling

- Use this where a frame is unavailable

  E.g. We can't obtain list of bank employees in London. But a list of *bank branches* can be easily obtained

  1. Select bank branches at random
  2. Interview ALL employees at selected branches.

- Advantages
  - cheap as a large number of interviews can be carried out at one location
  - It has the advantages of being a probability sample

- Disadvantages
  - Less reliable than SRS particularly where members of clusters are homogenous.

    - E.g. Employees at a particular bank branch may hold very similar opinions. "Wealthier" vs. "poorer" branches.

  - the sample does not reflect the full diversity of opinion of the bank employee population.

Multi-Stage Sampling

- Use where cluster contains too many individuals to interview them all

  E.g. A sample of students in England

  Stage 1: Select a random sample of universities

  Stage 2: For each university selected, take a random sample of courses

  Stage 3: For each course selected take a random sample of students

- Different sample designs can be used at each stage

  Stage 1: Select a SRS of universities

  Stage 2: For each university selected, take a stratified sample of courses to ensure all main subjects are represented

  Stage 3: For each course selected take a systematic sample of students from class lists

Pilot Testing

Pilot testing refers to testing the questionnaire on a small sample of respondents to identify and eliminate potential problems

Why:

1. Ensures validity and reliability
2. Sense check
3. Practical issues
4. Highlights key issues

How:

1. Focus groups discussions
2. Cognitive interviews
3. Field pre-testing
4. Academic testing

Excel-Central Limit Theorem

Seminar 5- Sampling

Exercise 1

- The amount of time a bank teller spends with each customer has a population mean, μ, of 3.10 minutes and standard deviation, ϭ, of 0.40 minute. If you select a random sample of 36 customers, what is the probability that the mean time spent per customer is at least 3 minutes?

Exercise 2

- A paper manufacturer has a production process that operates continuously throughout an entire production shift. The paper is expected to have a mean length of 11 inches, and the standard deviation of the length is 0.02 inch. At periodic intervals, a sample is selected to determine whether the mean paper length is still equal to 11 inches or whether something has gone wrong in the production process to change the length of the paper produced. You select a random sample of 100 sheets, and the mean paper length is 10.998 inches. Construct a 80% confidence interval estimate for the population mean paper length.

Thank You!

Any Questions?