Business Analytics

Lecture 5

Sampling

Dr. Yufei Huang

UCL
SCHOOL OF
MANAGEMENT

Attendance Code (30 Nov 2023)

398894

Important Information

The week of 06 Nov. :

- Reading week
- No lecture, no seminar, no office hour

Coursework:

- All 5 questions have been published. Check Coursework 1 Assignment Brief under the “Assessment” section on Moodle
- Combine your answers to all 5 questions into one PDF version report.
- Deadline: 2pm,10 Nov. 2023 (Friday), submit via Moodle
- The report should be clearly structured, please briefly summarize and explain your results. You can include figures or tables in your report.
- Do not just print out an Excel sheet as your report
- Do not exceed 10 pages.
- Do not submit multiple files.

Office hour:

- Every Monday 1-2 pm
- Online via Zoom (see bottom of Moodle page for link)

UCL SCHOOL OF MANAGEMENT

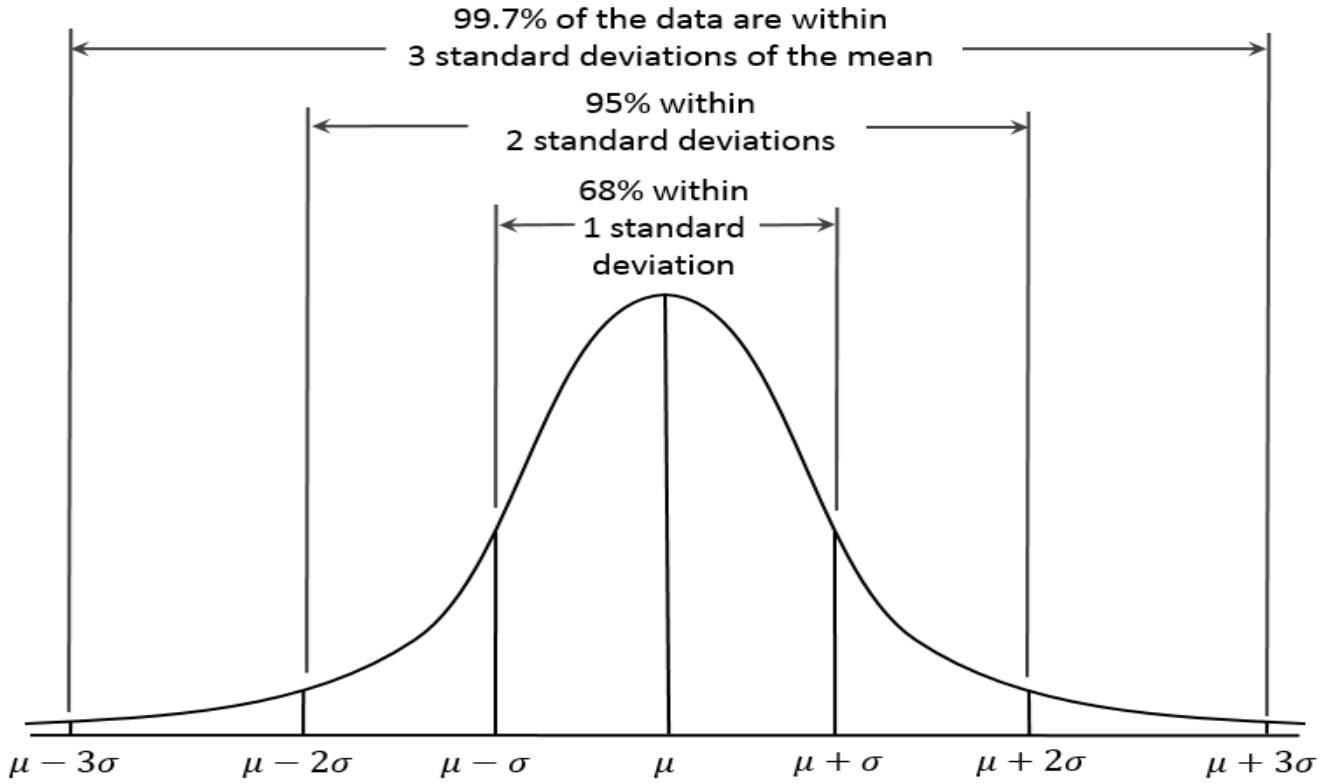
Contents

- Review: Normal Distribution
- Sampling
- Central Limit Theorem
- Confidence Interval

Properties of Normal Distribution

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

- The total area under the density curve = 1
- 68.26% of the area under the curve is between μ-σ, μ+σ,
- 95.44% of the area under the curve is between μ-2σ, μ+2σ,
- 99.72% of the area under the curve is between μ-3σ, μ+3σ.



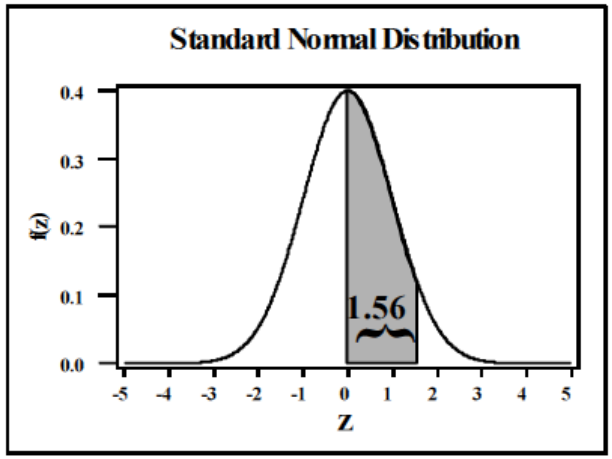
Standard Normal Distribution

- Definition. The standard normal random variable Z is the normal random variable with mean 0 and standard deviation1. That is:

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

- Notation. Z~N(0,1).

Using Standard Normal Distribution Table



Look in row labeled 1.5 and column labeled .06 to find $P(0 \le Z \le 1.56) = 0.4406$

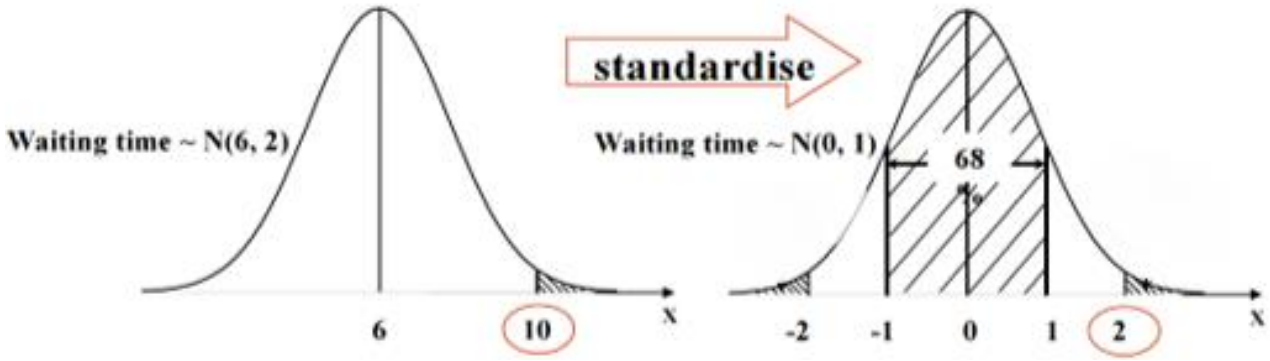
Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Standardization of Normal Distribution

$$X = \mu + \sigma Z, \quad Z = (X - \mu)/\sigma$$

Example. Suppose the waiting time for customer service is normally distributed with a mean of 6 min. and standard deviation of 2 min. What is the probability that a customer will wait more than 10 minutes?



P(Time>10)=P(Z>(10-6/2)=P(Z>2)=2.3%

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

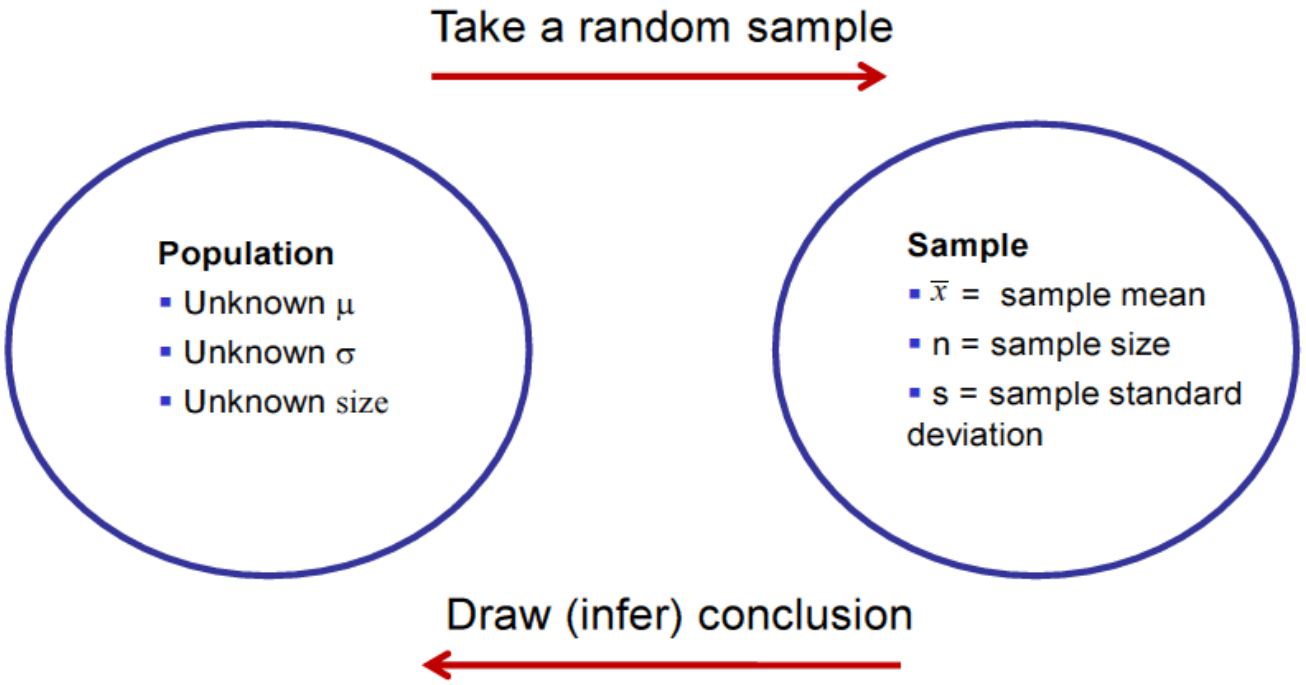
2. Measure the property in the members of the sample

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 $\bar{x}$ is the probability distribution of all possible values the random variable $\bar{x}$ 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 σ.

- 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?

$$n \geq 30$$

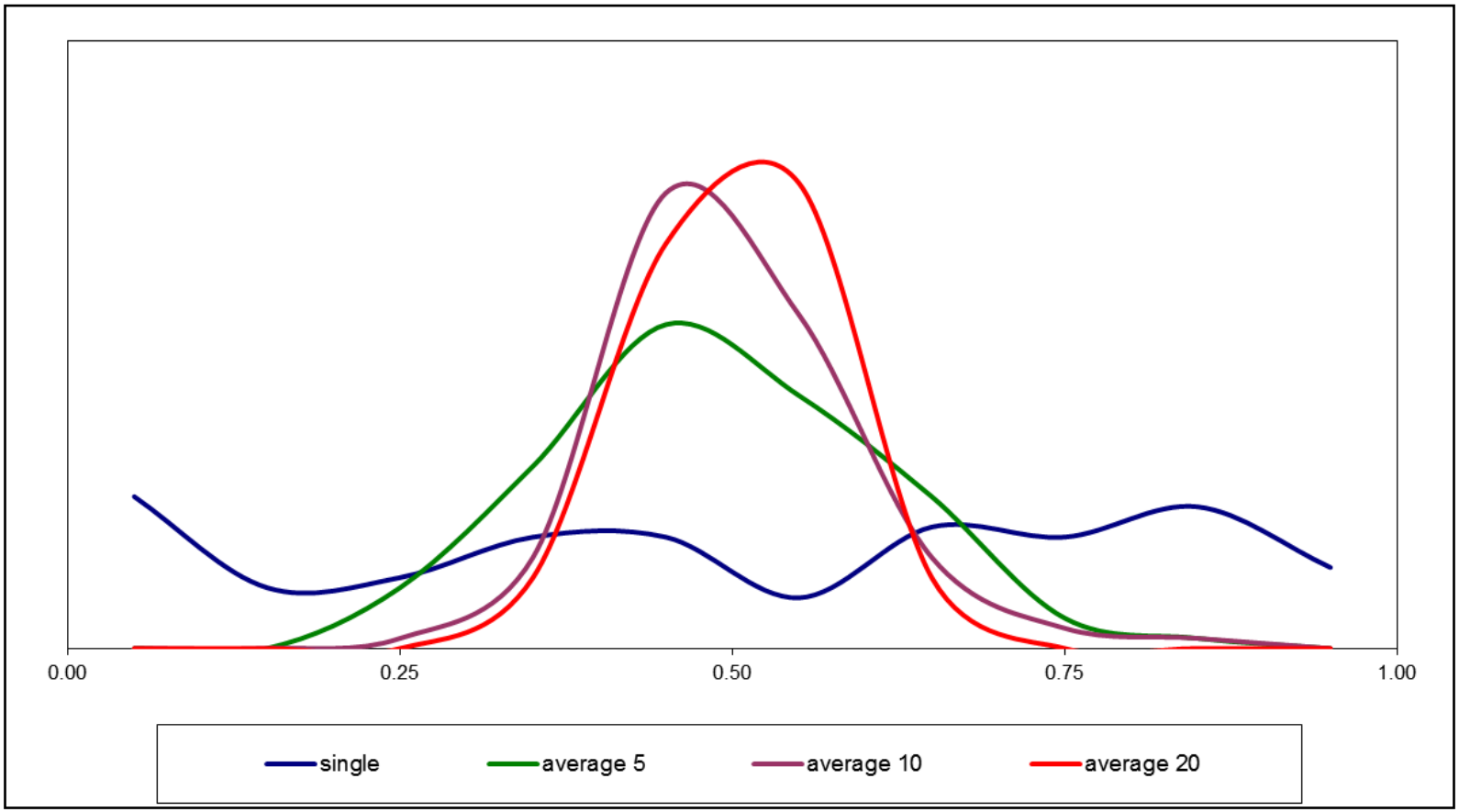
- Remark. Larger sample will give you better estimate of the population, i.e., smaller standard deviation of the sample mean distribution
- Remark. When sample size is small, sample mean stills follow normal distribution if the population is normally distributed

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,

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

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

Exercise

- 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?

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

$$\overline{X} \sim N(\mu, \sigma^2/n) = N(3.1, 0.4^2/36) = N(3.1, (0.4/6)^2)$$

$$= N(3.1, 0.06666^2).$$

$$P(\overline{X} \geq 3) = P\left(Z \geq \frac{3-3.1}{0.06666}\right) = P(Z \geq -1.5) = 1\text{-}P(Z < -1.5) = 0.9332.$$

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

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

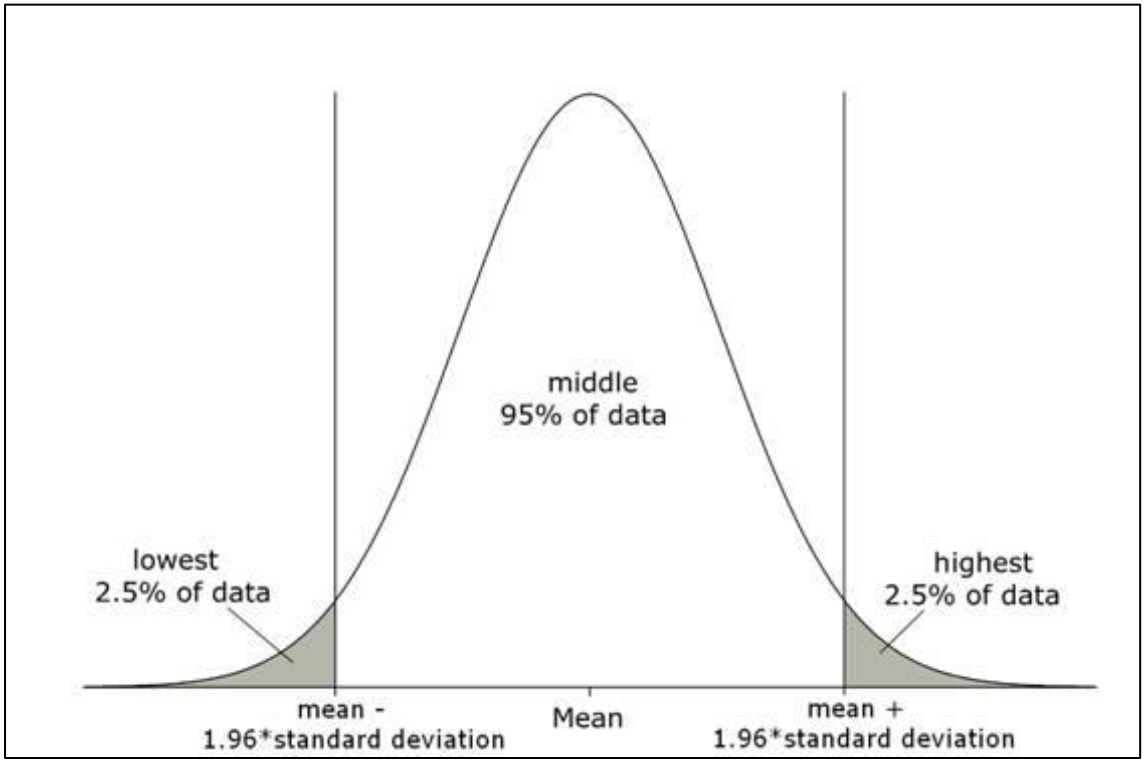
A 95% confidence interval for μ when σ is known and sampling is done from a normal population, or a large sample is used, is

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

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].

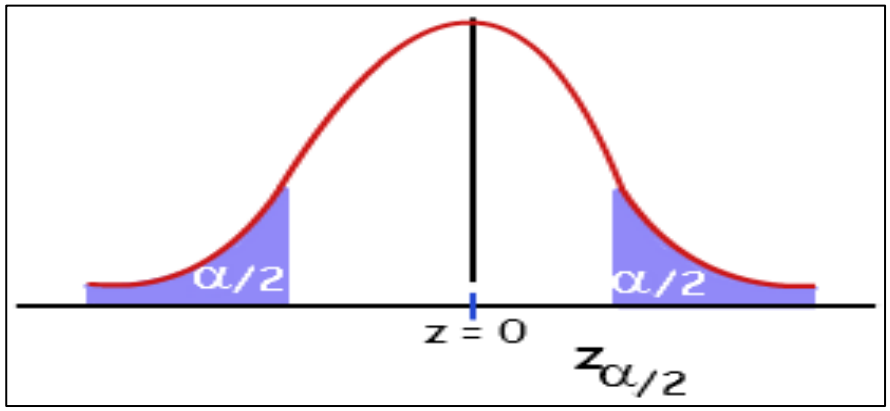
- Does the above confidence interval mean that 95% of all Seaweed's prices should lie in this interval?

  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

Exercise

- 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.
- Solution.

  The required interval is:

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

$$=\left[10.998-1.28\frac{0.02}{\sqrt{100}},10.998+1.28\frac{0.02}{\sqrt{100}}\right]=[10.995,11.00056].$$

  As 11 is included in this interval, the production process is ok.

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*

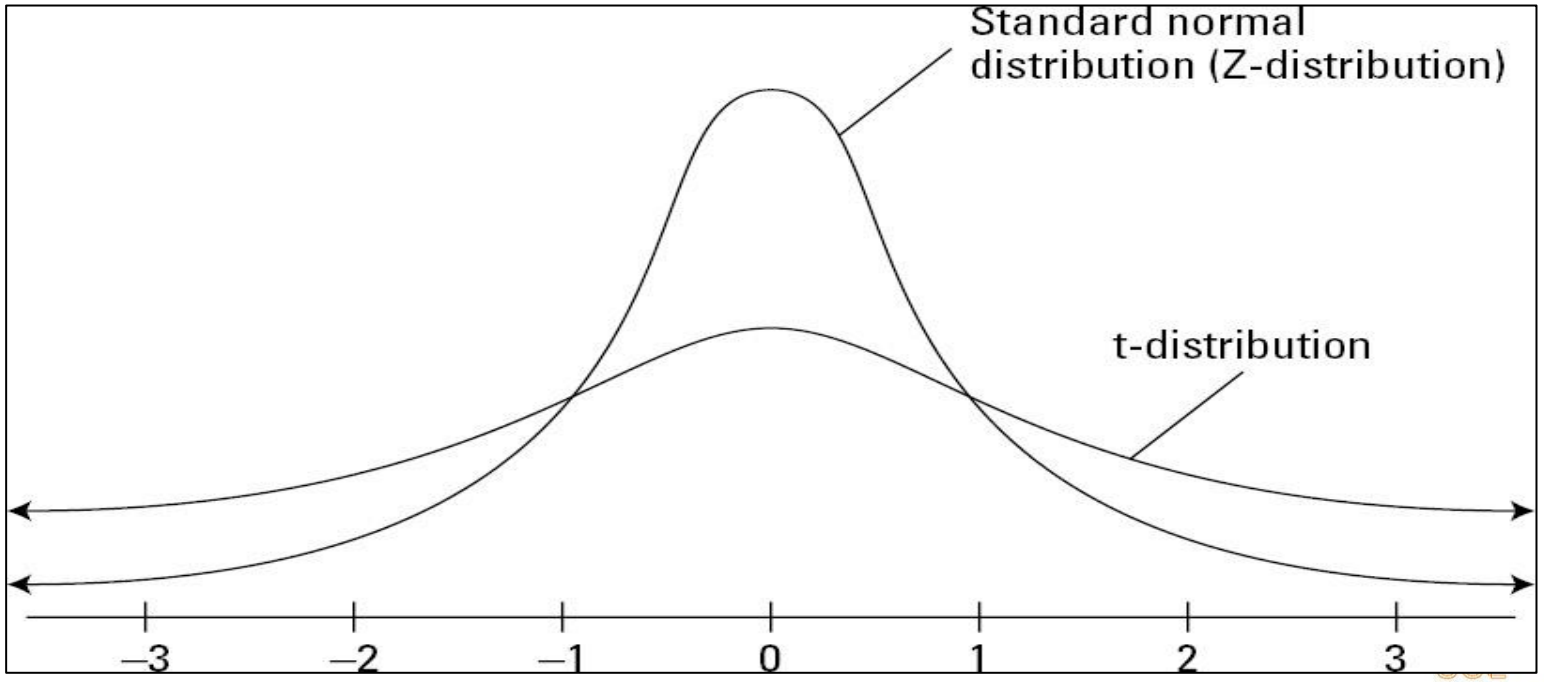
- However, it was shown that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is not normally distributed.

- 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.

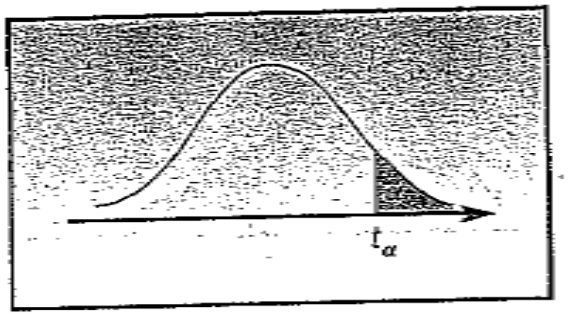


TABLE 6–1 Values and Probabilities of *t* Distributions

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 UCL

Sample Proportion

- Suppose np ≥ 5 and n(1 – p) ≥ 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

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

Excel: Central Limit Theorem and Confidence Interval

UCL
SCHOOL OF
MANAGEMENT

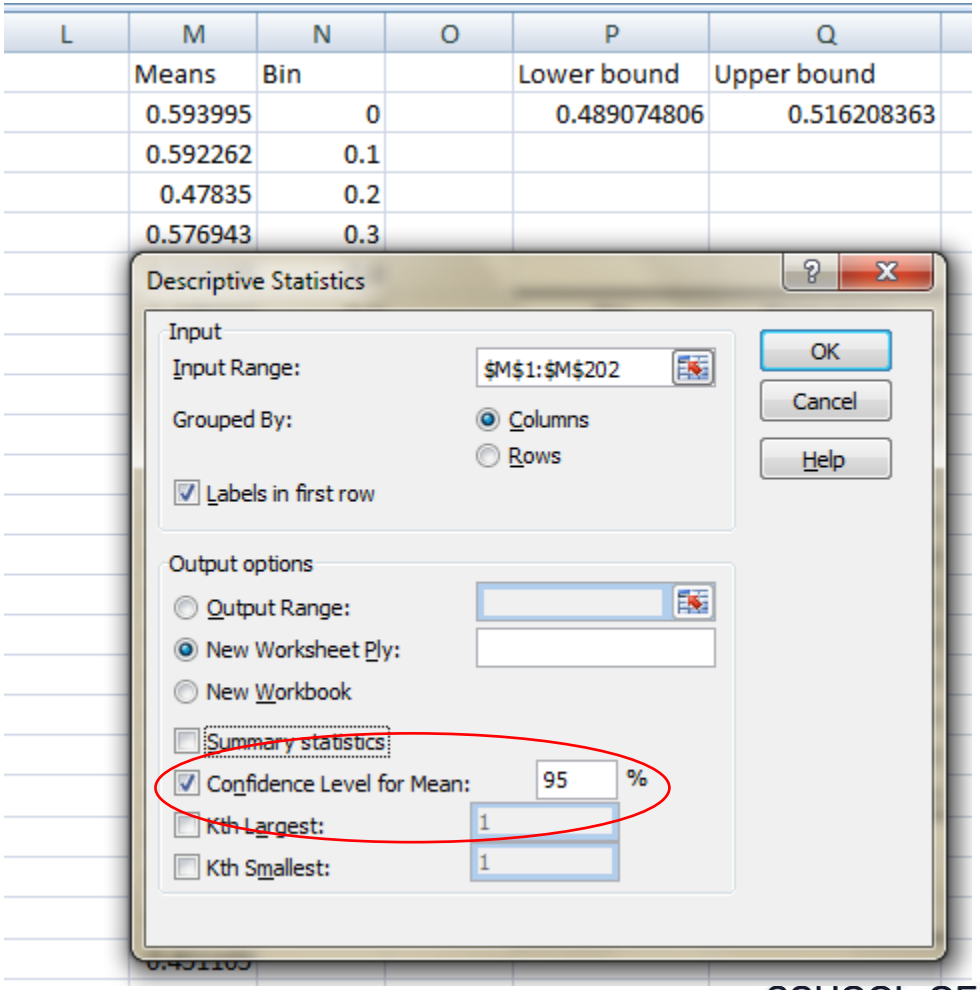
Use Excel for confidence interval

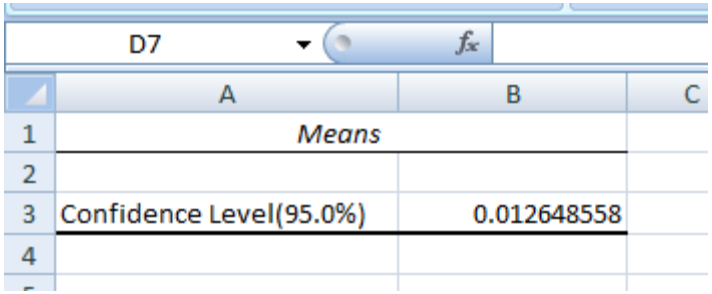
- Choose Data

-> Data Analysis

-> Descriptive statistics

- Tick Confidence Level for Mean: 95%





Reference

Chapters 5 and 6 of:

Aczel, A., & J. Sounderpandian. 2008. Complete Business Statistics. McGraw-Hill/Irwin, Seventh Edition.