


INNOMATICS
RESEARCH LABS



Central Limit Theorem (CLT)

Central Limit Theorem (CLT)

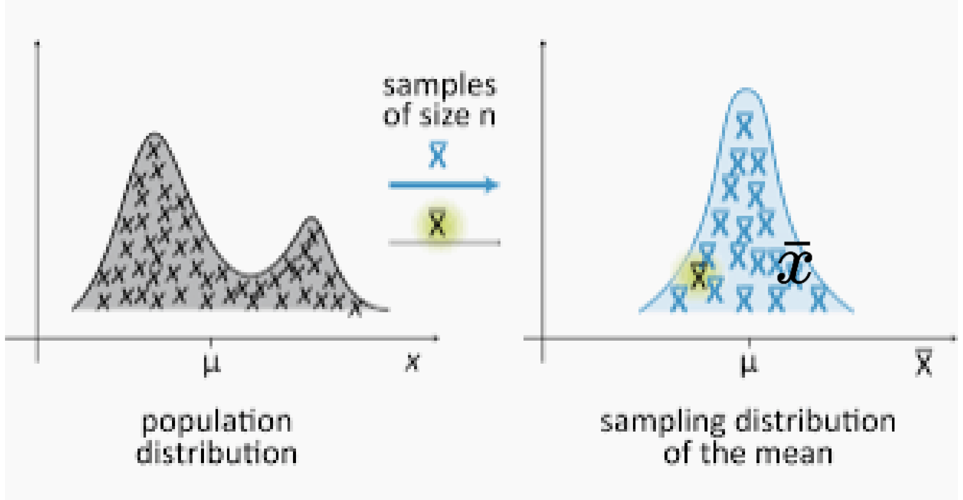
- Central limit theorem implies that no matter what the population distribution is, the sample mean (X bar) is normally distributed with mean (µ) and standard error (σ / $\sqrt{n}$) approximately.The distribution of the sample mean
  - will be normal when the distribution of data in the population is normal
  - will be approximately normal even if the distribution of data in the population is not normal, under some condition

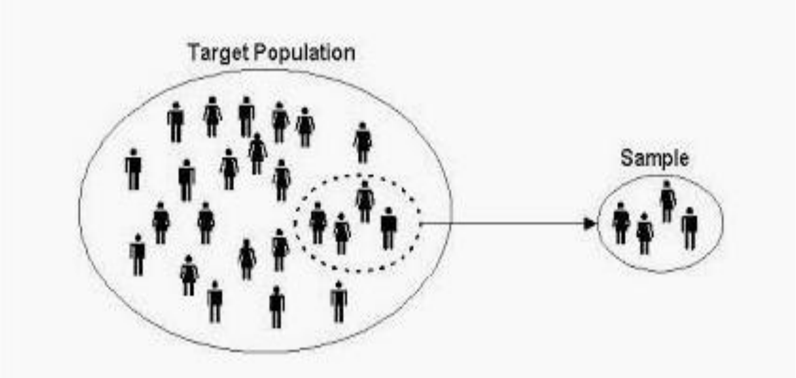
- Mean(X bar) = μ (the same as the population mean of the raw data)
- Standard deviation(X bar) = σ / $\sqrt{n}$
- where σ is the population standard deviation and n is the sample size
- This is also referred to as Standard Error of the Mean and is also denoted by SE(X bar) or σX bar.

Central Limit Theorem

This theorem is the primary reason why the normal distribution appears in so many statistical results. The theorem can be stated as follows.

For any **population distribution** with **mean μ** and **standard deviation σ**, the **sampling distribution** is approximately normal with mean μ and standard deviation σ/sqrt(n), and the approximation improves as n increases.





Sample Standard Deviation

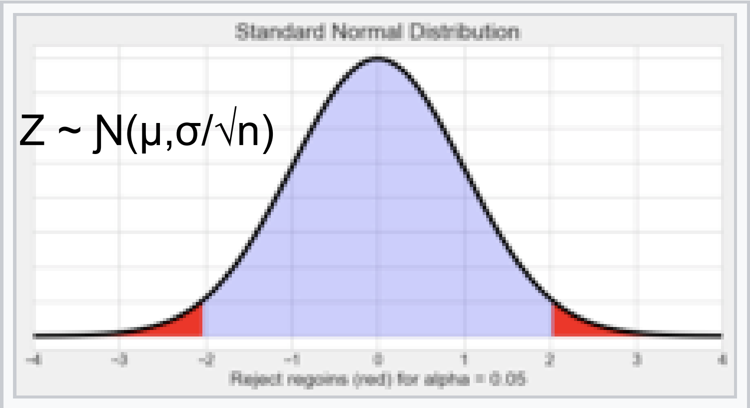
S.E = $\frac{\sigma}{\sqrt{n}}$

The sample standard deviation is also called the "**standard error**" and abbreviated S.E.

Confidence Interval

$$\bar{x} \pm z\text{-}value \times S.E.$$

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability12.html



Normal Distribution

z “statistic” or “z-value” is sample mean converted to the z-axis

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Many textbooks suggest **n ≥ 30** as a rule of thumb. However, this depends heavily on the population distribution.

If the population distribution is very non-normal—extremely skewed or bimodal, for example—the normal approximation might not be accurate unless n is considerably greater than 30. if the population distribution is already approximately symmetric, the normal approximation is quite good for n considerably less than 30.

In the special case where the population distribution is normal then the sampling distribution is normal.

CLT Validity

- Each data point in the sample is independent of the other
  - The sample size is large enough
  - A sample size of 30 is usually considered large enough but there are more precise conditions

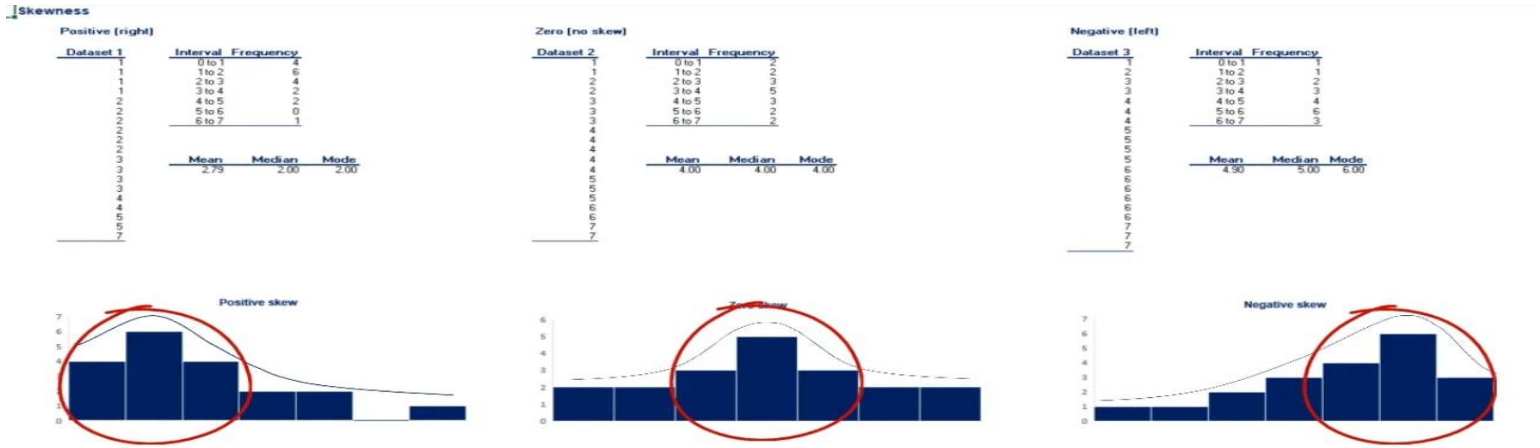
- n > 10 (K3)2, where K3 is sample skewness, and
- n > 10 |K4|, where K4 is sample kurtosis
  - Adequate sample size depends on the distribution of data – primarily its symmetry and presence of outliers
  - If data is quite symmetric and has few outliers, even smaller samples are fine. Otherwise, we need larger samples

Skewness:

- Skewness basically gives the shape of normal distribution of values.
- Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right.
- Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

Skewness:

- Skewness tells us a lot about where the data is situated.



Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

Zero (no skew)

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

Negative (left)

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
6
7
7
7

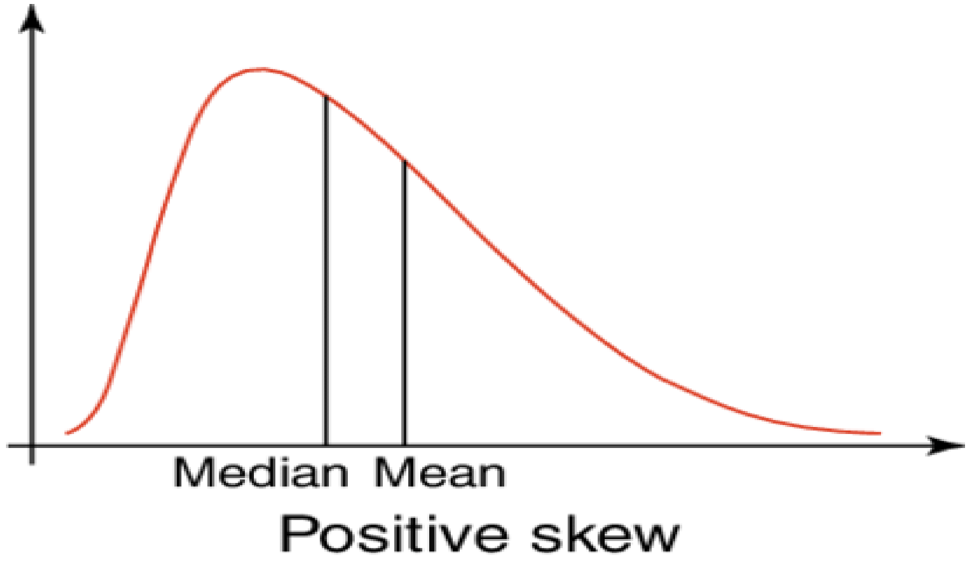
Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00

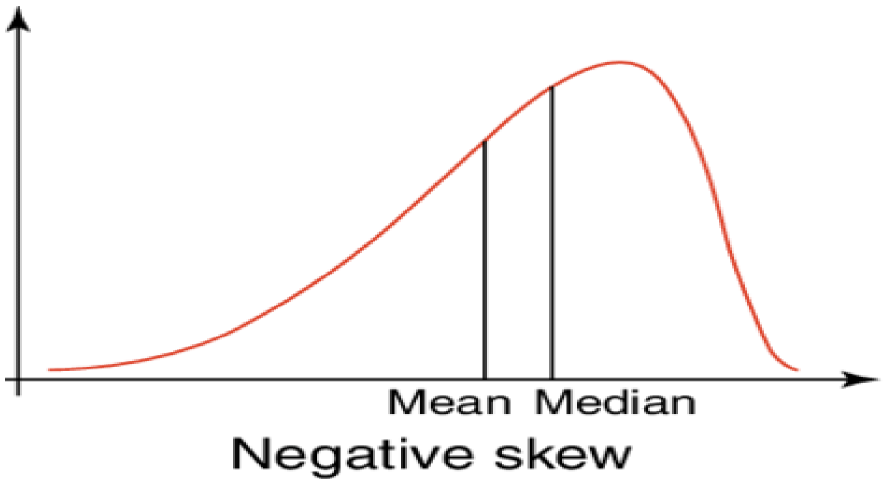
- In fact, the mean, median and mode should be used together to get a good understanding of the dataset.
- Measures of asymmetry like skewness are the link between central tendency measures and probability theory.
- This ultimately allows us to get a more complete understanding of the data we are working with.

Positive and Negative Skewness:

A positively skewed distribution means that the extreme data results are larger. This skews the data in that it brings the mean (average) up. The mean will be larger than the median in a Positively skewed distribution.

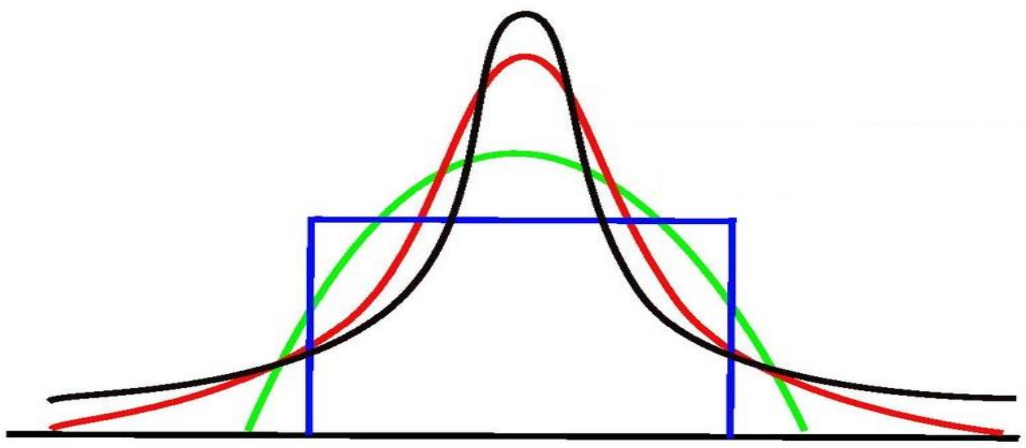


A negatively skewed distribution means the opposite: that the extreme data results are smaller. This means that the mean is brought down, and the median is larger than the mean in a negatively skewed distribution.



Kurtosis:

The exact interpretation of the measure of Kurtosis used to be disputed, but is now settled. It's about existence of outliers. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution.



Example:

- A certain brand of tires has a mean life of 25,000 miles with a standard deviation of 1600 miles. What is the probability that the mean life of 64 tires is less than 24,600 miles?

Example Contd.

- The sampling distribution of the means has a mean of 25,000 miles (the population mean)

  $\mu = 25000$ mi

- And a standard deviation (i.e.. standard error) of:

  $1600/8 = 200$

- Convert 24,600 mi. to a z-score and use the normal table to determine the required probability.

  $z = (24600-25000)/200 = -2$

  $P(z< -2) = 0.0228$

  or 2.28% of the sample means will be less than 24,600 mi.

ESTIMATION OF POPULATION VALUES

Point Estimates:

A single value given as an estimate of a parameter of a population

Interval Estimates:

An interval within which the value of a parameter of a population has a stated probability of occurring.

CONFIDENCE INTERVAL ESTIMATES for LARGE SAMPLES

- The sample has been randomly selected
- The population standard deviation is known or the sample size is at least 30.

Confidence Interval Estimate of the Population Mean

$$\bar{X} - z\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + z\frac{s}{\sqrt{n}}$$

-

X: sample mean

s: sample standard deviation

n: sample size

Example:

Estimate, with 95% confidence, the lifetime of nine volt batteries using a randomly selected sample where:

X = 49 hours

s = 4 hours

n = 36

EXAMPLE continued

- Lower Limit: 49 - (1.96)(4/6)

  49 - (1.3) = 47.7 hrs

- Upper Limit: 49 + (1.96)(4/6)

  49 + (1.3) = 50.3 hrs

We are 95% confident that the mean lifetime of the population of batteries is between 47.7 and 50.3 hours.

CONFIDENCE BOUNDS

- Provides a upper or lower bound for the population mean.
- To find a 90% confidence bound, use the z value for a 80% CI estimate.

Example

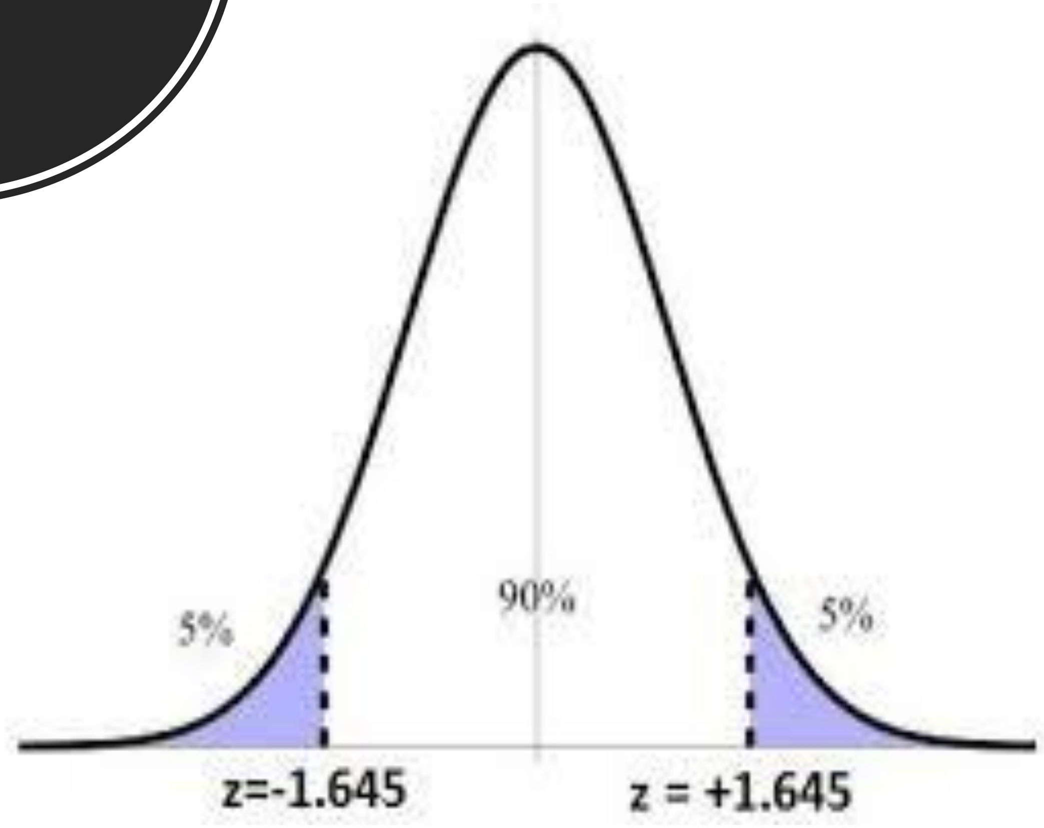
- The specifications for a certain kind of ribbon call for a mean breaking strength of 180 lbs. If five pieces of the ribbon have a mean breaking strength of 169.5 lbs with a standard deviation of 5.7 lbs, test to see if the ribbon meets specifications.
- Find a 95% confidence interval estimate for the mean breaking strength.
- You sample 36 apples from your farms harvest of 200,000 apples. The mean weight of sample is 112g with 40g samples standard deviation. What is the probability that the mean weight of all 200,000 apples is between 100 and 124g?

Questions

- I don't know how population is distributed, mean and sd is not known., and I know sample is coming from normal distribution. How should we solve?
- What should we find out? (is population mean is between 100 or 124?)
- We have to find if mean = x $\pm$ 12, similar is x= mean $\pm$ 12

Confidence Interval (CI)

CONFIDENCE INTERVAL



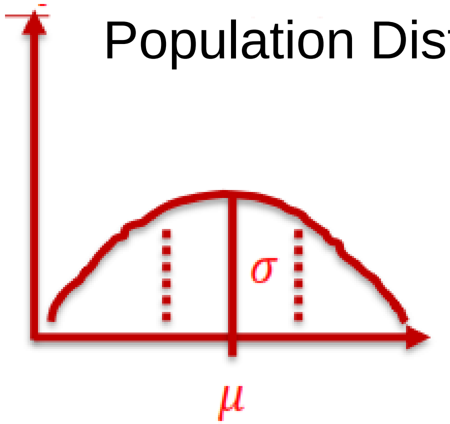
CONFIDENCE INTERVAL

When we use samples to provide population estimates, we cannot be CERTAIN that they will be accurate. There is an amount of uncertainty, which needs to be calculated.

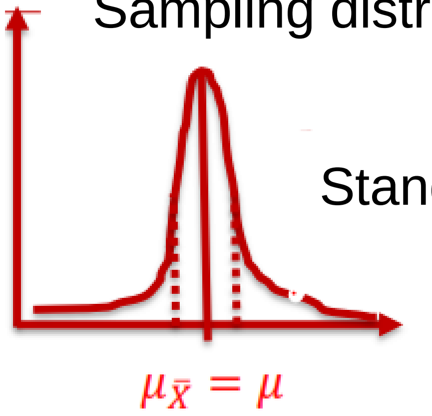
Year	Number of seats - Vidhan Sabhas	Number of electors	Number of votes polled	Number of valid votes polled	Percentage of votes polled
2000	621	96848465	60863266	60086040	62.84
2001	824	132981673	91682025	91503795	68.94
2002	959	162610378	91923473	91742832	56.53
2003	878	102629569	69236882	66806794	67.46
2004	697	115667178	78066138	78026621	67.49
2005	657	134575644	67354156	67342495	50.05
2006	824	134345929	101812769	101755877	75.78
2007	938	180225337	95760846	95926937	53.13
2008	180	4554900	4106644	4100021	90.16
2009	992	193268747	125422697	125733981	64.9
2010	243	55120656	29034705	29058604	52.67
2011	824	145606345	116111638	116463760	79.74
2012	940	197117093	126575548	127027322	65.53
2013	1034	197117093	121437273	120514648	73.27
2014	1079	223262105	154928094	153354084	69.39
2015	70	13313295	8982228	8942372	67.47

Incorrect way to present data as it gives the feeling that the population parameter will lie within these ranges.

Population Distribution



Sampling distribution of sample means

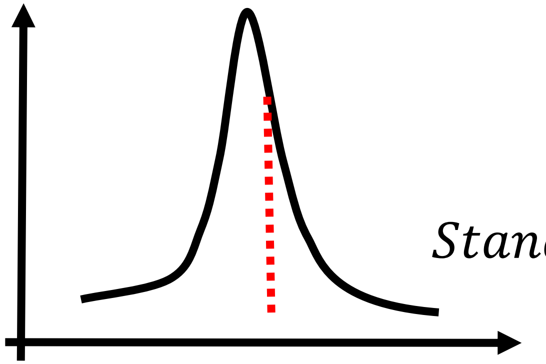


Standard error of mean = $\frac{\sigma}{\sqrt{n}}$

Standard Error (SE) is the same as Standard Deviation of the sampling distribution and a sample with 1 SE may or may not include the population parameter.

Confidence Interval:

Sampling distribution of sample means



$$Standard\ Error\ of\ the\ mean = \frac{\sigma}{\sqrt{n}}$$

- We have seen that ~ 95% of the samples will have a mean value within the interval +/- 2 SE of the population mean *(recall the Empirical Rule for Normal Distribution).*

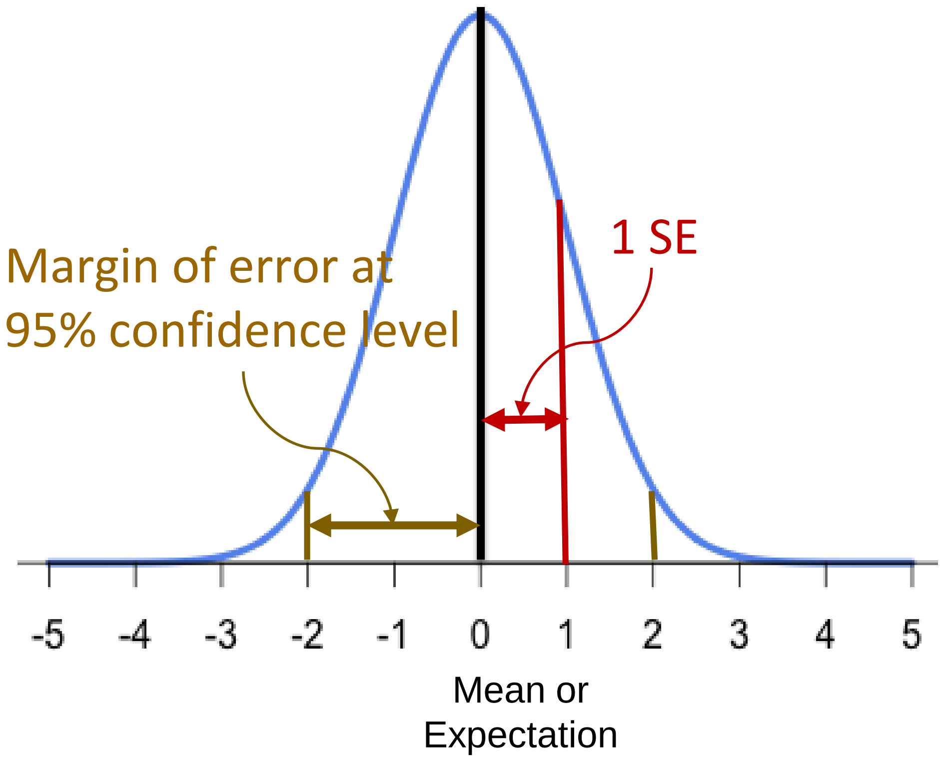
- Alternatively, 95% of such intervals include the population mean. Here, 95% is the Confidence Level and the interval is called the Confidence Interval.

SE, Margin of Error, Confidence Interval and Sample Size

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$Margin\ of\ Error = z * SE$$

Margin of error is the **maximum expected difference between the** true population parameter and a sample estimate of that parameter.



Margin of error is meaningful only when stated in conjunction with a probability (confidence level).

SE, Margin of Error, Confidence Interval and Sample Size

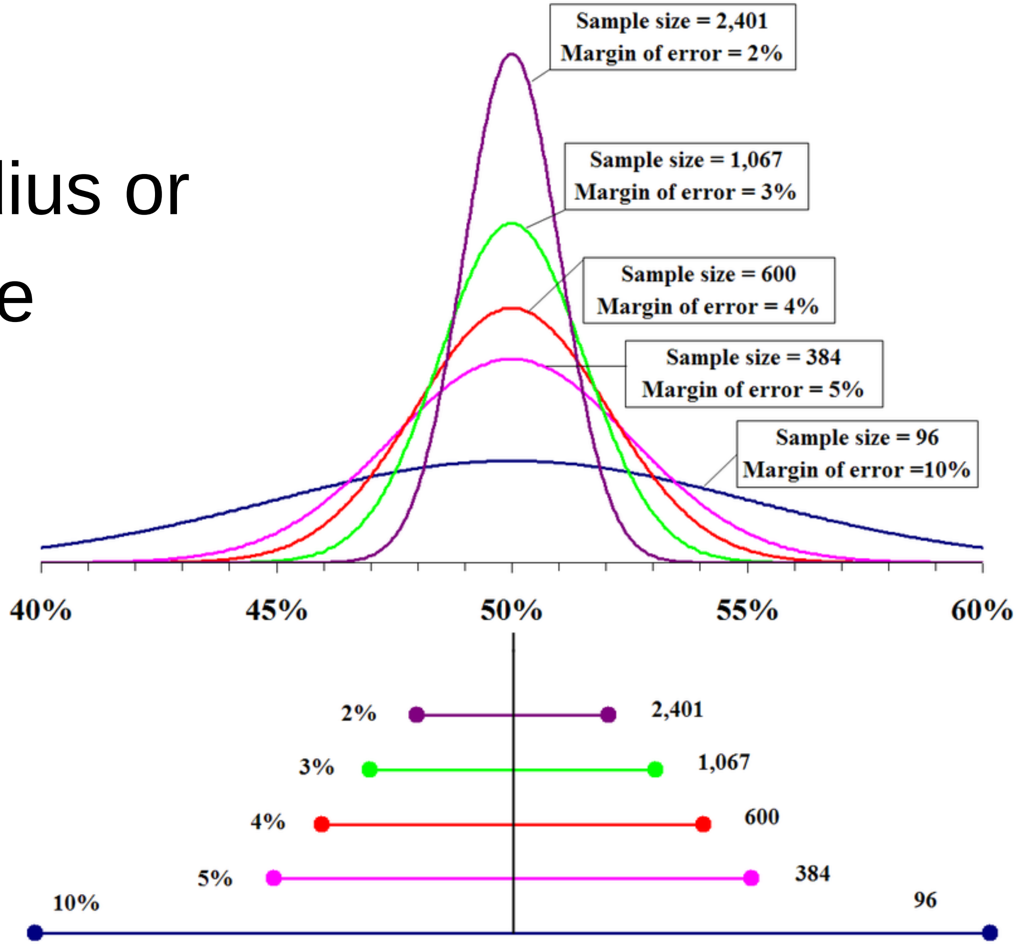
Just like Mean, Proportion is another common parameter of interest in many problems.

Expectation of a sample proportion = p

SE of a sample proportion = $\sqrt{\frac{pq}{n}}$

SE, Margin of Error, Confidence Interval and Sample Size

Margin of error is the radius or half-width of a confidence Interval.



SE, Margin of Error, Confidence Interval and Sample Size

Suppose you would like to estimate the percentage of people in Arizona that say they enjoy the summer heat. You survey 500 people and find that 267 of them say that they do. This translates into a ratio of the whole of 267/500 = 0.534 or a percentage of 53.4%. What is the margin of error in the estimate if you use a confidence level of 90% ?

First, the sample size is n = 500. We represented the number of people who answered yes as a ratio of the whole: s = 0.534. This will approximate the unknown population ratio:

$$p0 \cong s = 0$$

Continued..

When we know the population probability, the formulas for the population parameters are

$\mu = np_0$

$\sigma^2 = np_0(1-p_0)$

$\sigma = sqrt(np0(1-np0))$

We use these and the approximation of p0 to compute a sample mean, a sample variance, and a sample standard deviation

$\mu = nm = ns = 500 * 0.534 = 267$

$d^2 = ns(1-ns) = 500 * 0.534 * 0.466 = 122.82$

$d = sqrt(ns(1-ns)) = sqrt(122.82) = 11.082$

Continued…

So we are 90% confident that the actual mean μ is between,

μ-1.645d and μ+1.645d

That is,

267 - (1.645)*(11.082) <= μ <= 267 + (1.645)*(11.082)

248.77 <= μ <= 28

But $\mu = np_0$ .Thus

$248.77 <= 500p_0 <=285.23$

So we are 90% confident that the actual probability p_0 is in the interval

$0.49754 = 248.77/500 <= p_0 <= 285.23/500 = 0.57046$

Thus the population percentage P0 is in the interval

$49\% <= P_0 <= 57\%$

We round off making sure to widen the interval so that we do not loose any confidence in our estimation interval. The final result we obtain is an estimate of 53.4% within an error of ±4.4% and a confidence of 90%

Confidence Intervals

A survey was taken by the Indian Government to do business with firms in China. A survey questions is: Approximately how many years has your company been trading with firms in China?

A random sample of 54 responses to this question yielded a mean of 11.566 years. Suppose the population standard deviation for this question is 8.4 years.

Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in China for the population of Indian companies trading with firms in China.

Continued……….

$n = 54$,
$\bar{x} = 11.566,$
$\sigma = 8.4$

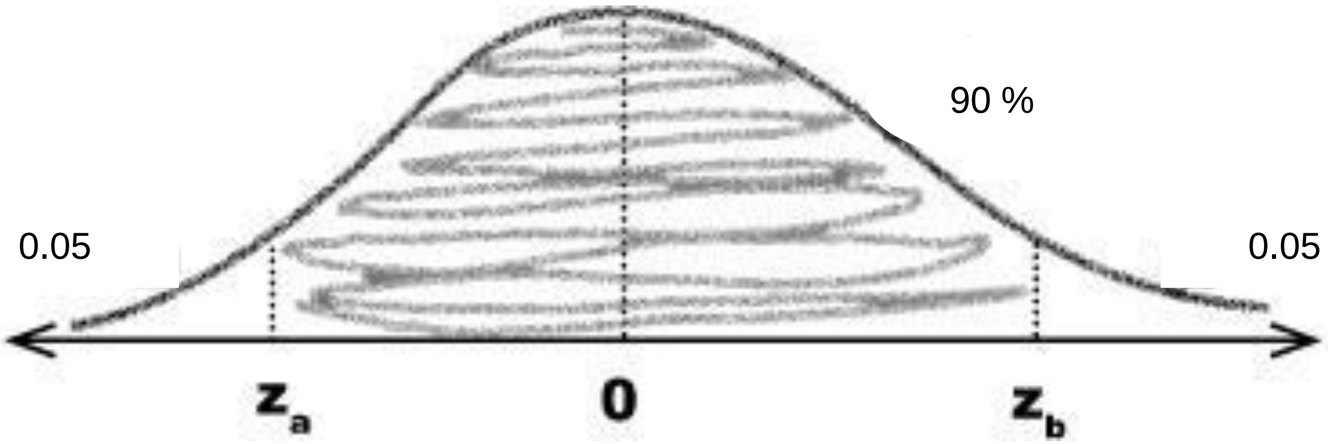
$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z \ * \frac{\sigma}{\sqrt{n}}$

Confidence Interval for the population Mean is

Sample Mean $\pm$ Margin of Error

Continued.....

Find Z_a and Z_b where $P(Z_a < Z < Z_b) = 0.90$



$P(Z < Z_a) = 0.05$ and $P(Z > Z_b) = 0.05$

Continued……

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

From probability tables using interpolation, we get Z_a= -1.645 and Z_b= 1.645.

Check qnorm(0.05, 0,1) and qnorm(0.95, 0, 1).

Continued…..

$Margin\ of\ error\ at$ 90% $Confidence\ Level$ = 1.645 *$\frac{8.4}{\sqrt{54}}$ = 1.88

Recall Confidence Interval for the Population Mean is Sample Mean ± Margin of Error

$\bar{X}$ - 1.88 < μ < $\bar{X}$ + 1.88

Since the sample mean is 11.566 years, we get the confidence interval for 90%

as 9.686 < μ < 13.446

The analyst is 90% confident that if a census of all US companies trading with firms in India were taken at the time of the survey, the actual population mean number of trading years of such firms would be between 9.686 and 13.446 years.

Shortcuts for Calculating Confidence Intervals

Population Parameter	Population Distribution	Conditions	Confidence Interval
μ	Normal	You know σ^2 n is large or small $\bar{X}$ is the sample mean	$\left(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z\frac{\sigma}{\sqrt{n}}\right)$
μ	Non-Normal	You know σ^2 n is large (>30) $\bar{X}$ is the sample mean	$\left(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z\frac{\sigma}{\sqrt{n}}\right)$
μ	Normal or Non-Normal	You don't know σ^2 n is large (>30) $\bar{X}$ is the sample mean s^2 is the sample variance	$\left(\bar{X} - z\frac{s}{\sqrt{n}}, \quad \bar{X} + z\frac{s}{\sqrt{n}}\right)$
p	Binomial	n is large p_s is the sample proportion q_s is 1 - p_s	$\left(\bar{X} - z\frac{s}{\sqrt{n}}, \quad \bar{X} + z\frac{s}{\sqrt{n}}\right)$

Shortcuts for Calculating Confidence Interval

You took a sample of 54 Pens and found that in the sample, the proportion of red Pens is 0.30. Construct a 99% confidence interval for the proportion of red Pens in the population.

Level of Confidence	Value of z
90 %	1.64
95 %	1.96
99 %	2.58

$$0.30 - 2.58 * \sqrt{\frac{0.30 * 0.70}{54}} < p < 0.30 + 2.58 * \sqrt{\frac{(0.30 * 0.70)}{54}}$$

$$0.13 < p < 0.46$$

Water consumption values of each individual

The marriage invitations were send to 27 people. The mean water consumed for this sample is 3.3907 liters and standard deviation, *s* is 0.67 liters. Construct the 95 % confidence Interval.

Level of Confidence	Value of Z
90 %	1.64
95 %	1.96
99 %	2.58

2.85	2.85	2.98	3.04	3.10	3.19	3.20	3.30	3.39
3.42	3.48	3.50	3.54	3.57	3.60	3.60	3.69	3.70
3.70	3.75	3.78	3.83	3.90	3.96	4.05	4.08	4.10

$$95\%\ CI: (3.3907 - 1.96 * \frac{0.67}{\sqrt{27}}, 3.3907 + 1.96 * \frac{0.67}{\sqrt{27}}$$

$$= (3.198, 3.643)$$

Attention Check

What happens to confidence interval as confidence level changes?

As confidence level increases, the confidence interval becomes wider and *vice-versa.*

What happens to the confidence interval as sample size changes ?

As sample size increases, the confidence interval become narrower.

Remember $\left(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}}\right).$

Confidence Intervals for a Sample Median

Confidence limits are given by actual values in the sample using the following formulae:

Lower 95 % Confidence Limit: $\frac{n}{2} - 1.96 * \frac{\sqrt{n}}{2}$ ranked value.

Upper 95 % Confidence Limit: $1 + \frac{n}{2} + 1.96 * \frac{\sqrt{n}}{2}$ ranked values.

Confidence Intervals for a Sample Median

2.85	2.85	2.98	3.04	3.10	3.10	3.19	3.20	3.30	3.39
3.42	3.48	3.50	3.54	3.54	3.57	3.60	3.60	3.69	3.70
3.70	3.75	3.78	3.83	3.90	3.96	4.05	4.08	4.10	4.14

Lower 95 % CI

Upper 95 % CI

Median

Lower 95 % confidence Limit: $\frac{27}{2} - 1.96 \ * \frac{\sqrt{27}}{2} = 8.40$ ranked value.

8th ranked value in 3.20

Upper 95 % confidence Limit: $1 + \frac{27}{2} + 1.96 \ * \frac{\sqrt{27}}{2} = 19.59$ ranked values.

19th ranked value is 3.70.

95 %CI: (3.20,3.70)

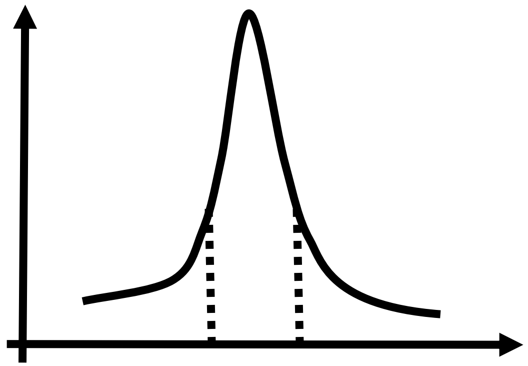
Confidence Intervals for a Sample Median

- Lack of distributional assumptions makes it difficult to obtain an exact CI for the median.
- CI are not necessarily symmetric around the sample estimate.

The Summary of CI

Confidence Interval = Sample statistics $\pm$ Margin of Error

$$\left(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}}\right)$$



Margin of error = z * Standard Error *(Recall the standardization formula)*

Depends on the Confidence level

$$\frac{\sigma}{\sqrt{n}}$$

Probability density.
Area under the curve between the limits.
Probability that a certain % of samples will contain the population mean within this interval.

Standard deviation of the population: Measure of deviation from the mean

A Short detour – Variance Formula Differences

Population Parameter

$$\mu = \frac{\sum x}{N}$$

$$\text{Variance } \sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Sample Statistic

$$\bar{\mathrm{x}} = \frac{\sum x}{N}$$

$$\text{Variance } S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

THANK YOU