Confidence Interval, t-Distribution, Hypothesis Testing, t-Tests

Review

- Gaussian Distribution
  - Areas under the curve
  - Reading the normal distribution table
  - Approximating B(n,p) with Normal distribution
  - Continuity correction
- Central Limit Theorem

Gaussian Distribution

Gaussian Distribution is another name for a normal distribution.

- In statistics, the normal distribution is called the normal curve.
- In the social sciences, it's called the bell curve (because of it's shape).
- In physics, it's called the Gaussian distribution.

As a normal distribution is used for the analysis of astronomical data on positions, hence the term Gaussian distribution.

Gaussian Distribution

- The larger the standard deviation, the flatter the curve. The smaller the standard deviation, the higher the peak of the curve.

- The Gaussian distribution approximates the binomial and has a mean of n*p, where n is the number of events and p is the probability of any integer x.

Example1

According to Census, about 80% of the population in India eat rice as the main course. Suppose 200 Indian people are randomly sampled. Use continuity Correction.

- What is the probability that fewer than 130 people eat rice as their main course?
- What is the probability that between 135 and 145 (inclusive) eat rice as their main course?
- What is the probability that more than 120people eat rice as their main course?

Given,

X = 200

P = probability of population eats rice = 0.8

Example1

- Expected Mean = 0.8*200=160
- Variance = n*p*q= 200*0.8*0.2

1. Area under the curve from –Infinity to 129.5 (continuity correction)

```
> pnorm(129.5, 160, sqrt(200*0.8*0.2)
[1] 3.43397337
```

2. Area under the curve between 145.5 and 134.5

➢ pnorm(145.5,200*0.8, sqrt(200*0.8*0.2)) – pnorm (134.5,200*0.8, sqrt(200*0.8*0.2))

```
[1] 0.0026373080
```

3. Area under the curve above 120.5

```
> 1-pnorm(120.5,200*0.8, sqrt(200*0.8*0.2))
[1] 0.999999999981773
```

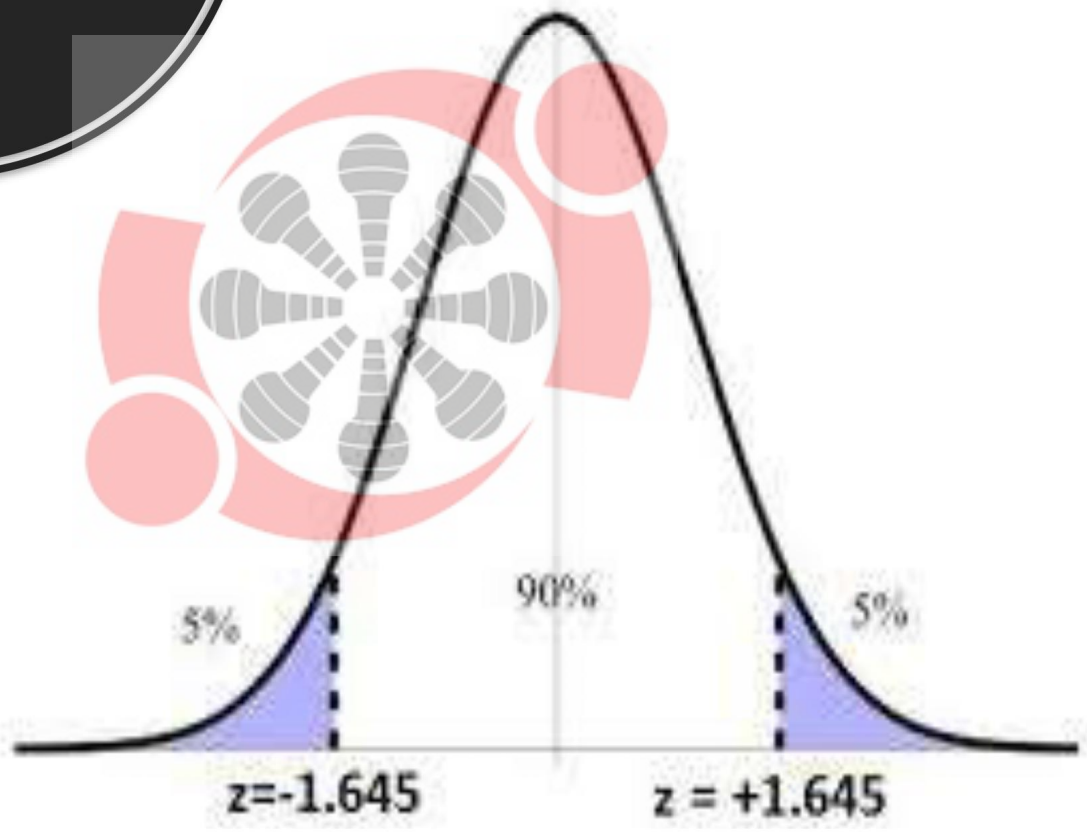

Example2

According to Annual Health Survey of India the Average effective literacy rate for Bihar has a mean of 68.7 as per 2011-13 with a standard deviation of 9.3

- What is the probability that the literacy rate is >74.8?

```
> 1 - pnorm(74.8, mean=68.7, std = 9.3)
[1]  0.9654056362281945
```


CONFIDENCE INTERVAL

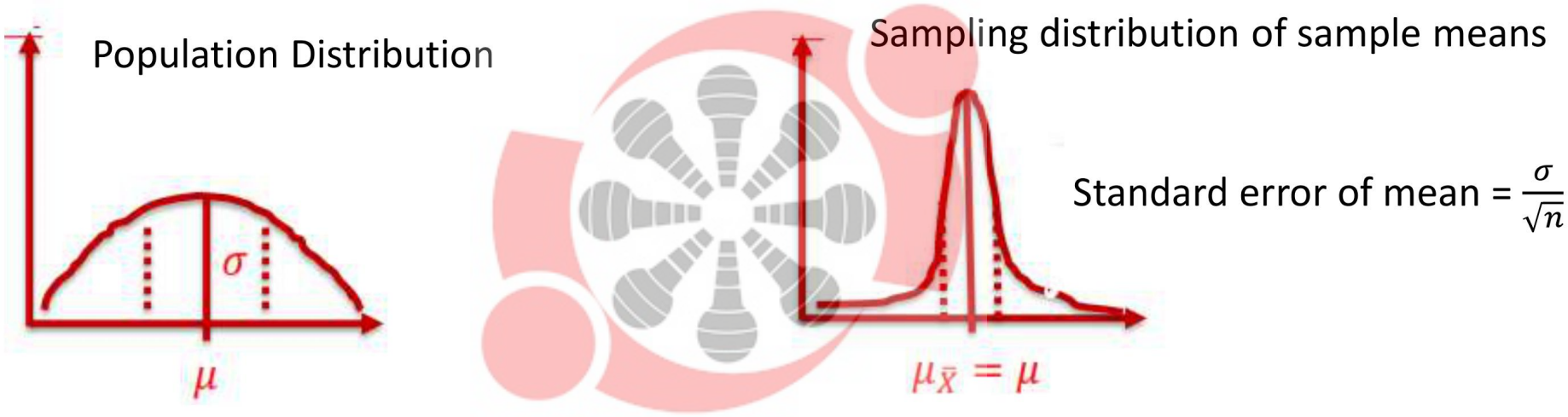


CONFIDENCE INTERVAL

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.

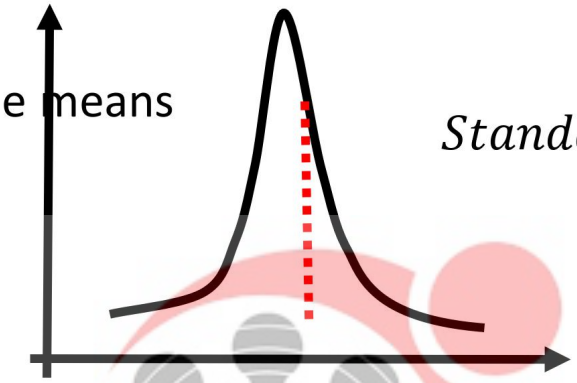
Year	Number of seats - Vidhan Sabhas	Number of electors	Number of votes polled	Number of valid votes polled	Percentage of votes polled
2000	621	96848465	60863266	60086040	62.84
2001	824	132981673	91682025	91503795	68.94
2002	959	162610378	91923473	91742832	56.53
2003	878	102629569	69236882	66806794	67.46
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.



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.

CI



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

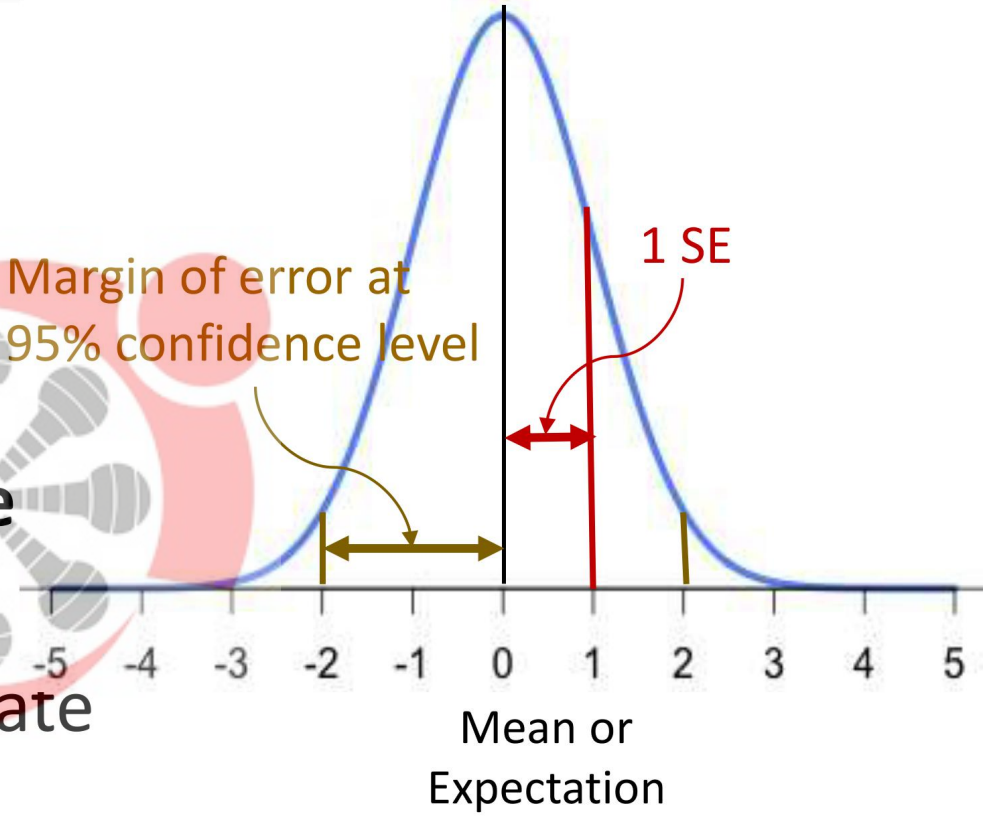
- We have seen that ~ 95% of the samples will have a mean value within the interval +/- 2 SE of the population mean *(recall the Empirical Rule for Normal Distribution).*

- 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

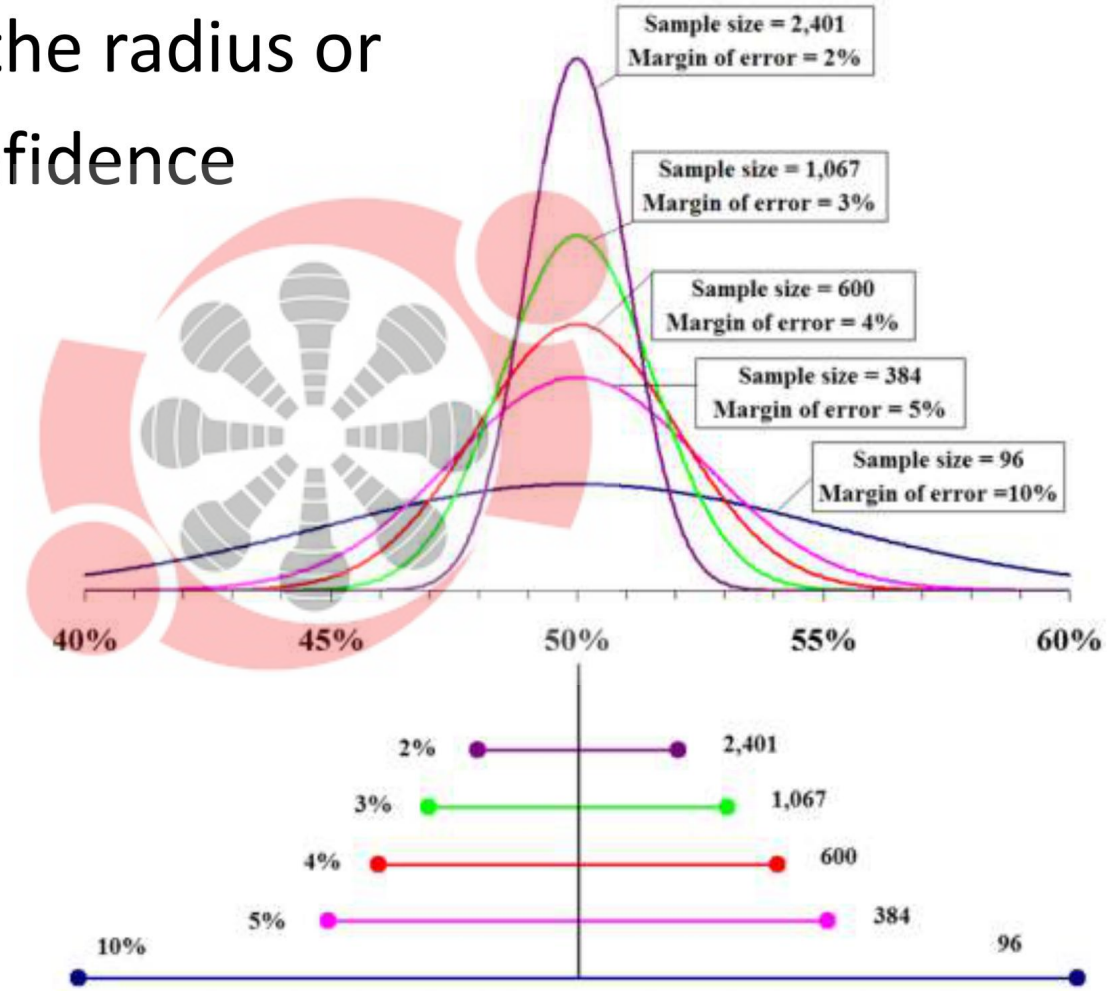
Just like Mean, Proportion is another common parameter of interest in many problems.

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:

$$p0 \cong s = 0$$

SE, Margin of Error, Confidence Interval and Sample Size

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

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

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

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

SE, Margin of Error, Confidence Interval and Sample Size

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

m-1.645d and m+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.

Confidence Intervals

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

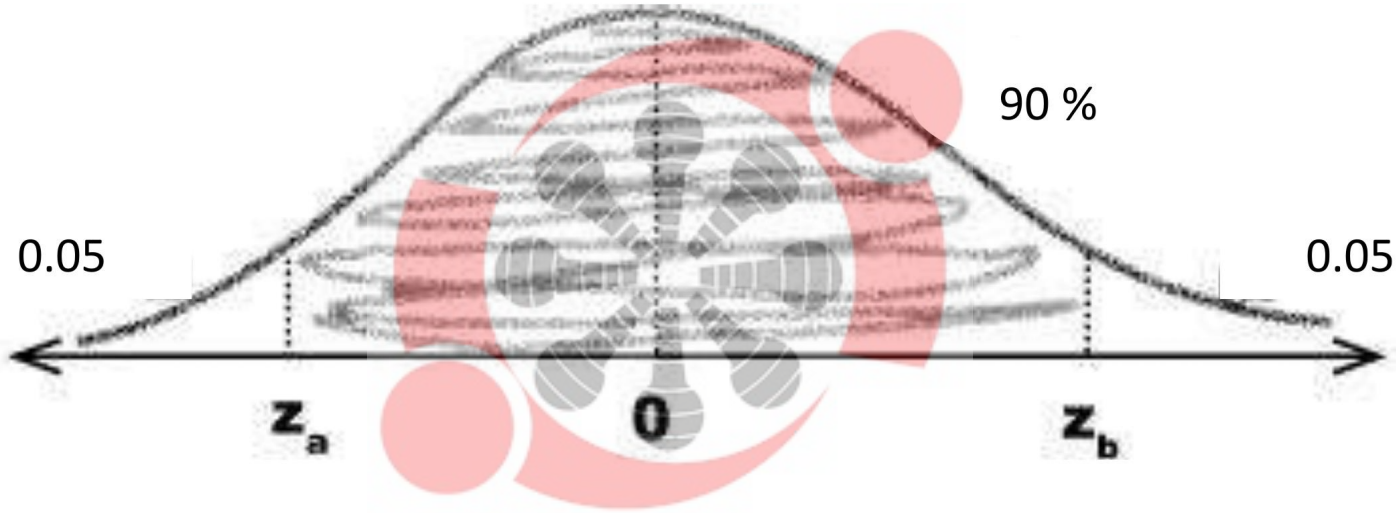
$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

Confidence Interval for the population Mean is

Sample Mean $\pm$ Margin of Error

Confidence Intervals

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



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

Confidence Intervals

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).

Confidence Interval

$Margin\ of\ error\ at$ 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

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

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.

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

$$0.13 < p < 0.46$$

Shortcuts for Calculating Confidence Interval

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

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.

Water consumption values of each individual

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

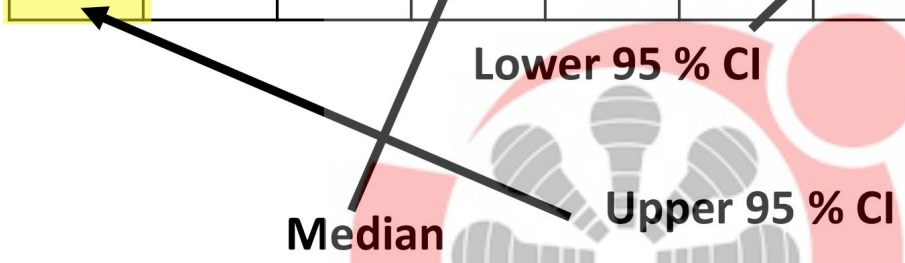
Confidence limits are given by actual values in the sample using the following formulae:

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



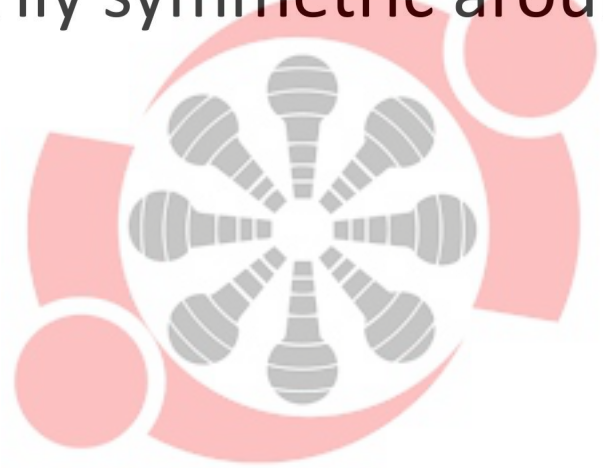
Lower 95 % confidence Limit: $\frac{27}{2} - 1.96 * \frac{\sqrt{27}}{2} = 8.40$ ranked value. 8th ranked value in 3.20

Upper 95 % confidence Limit: $1 + \frac{27}{2} + 1.96 * \frac{\sqrt{27}}{2} = 19.59$ ranked values. 19th ranked value is 3.70.

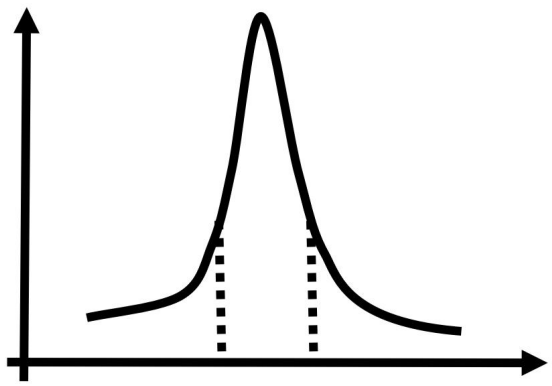
95 %CI: (3.20,3.70)

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

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

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

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

A Short detour – Variance Formula Differences

Population Parameter

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

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

Sample Statistic

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

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