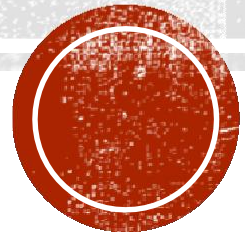
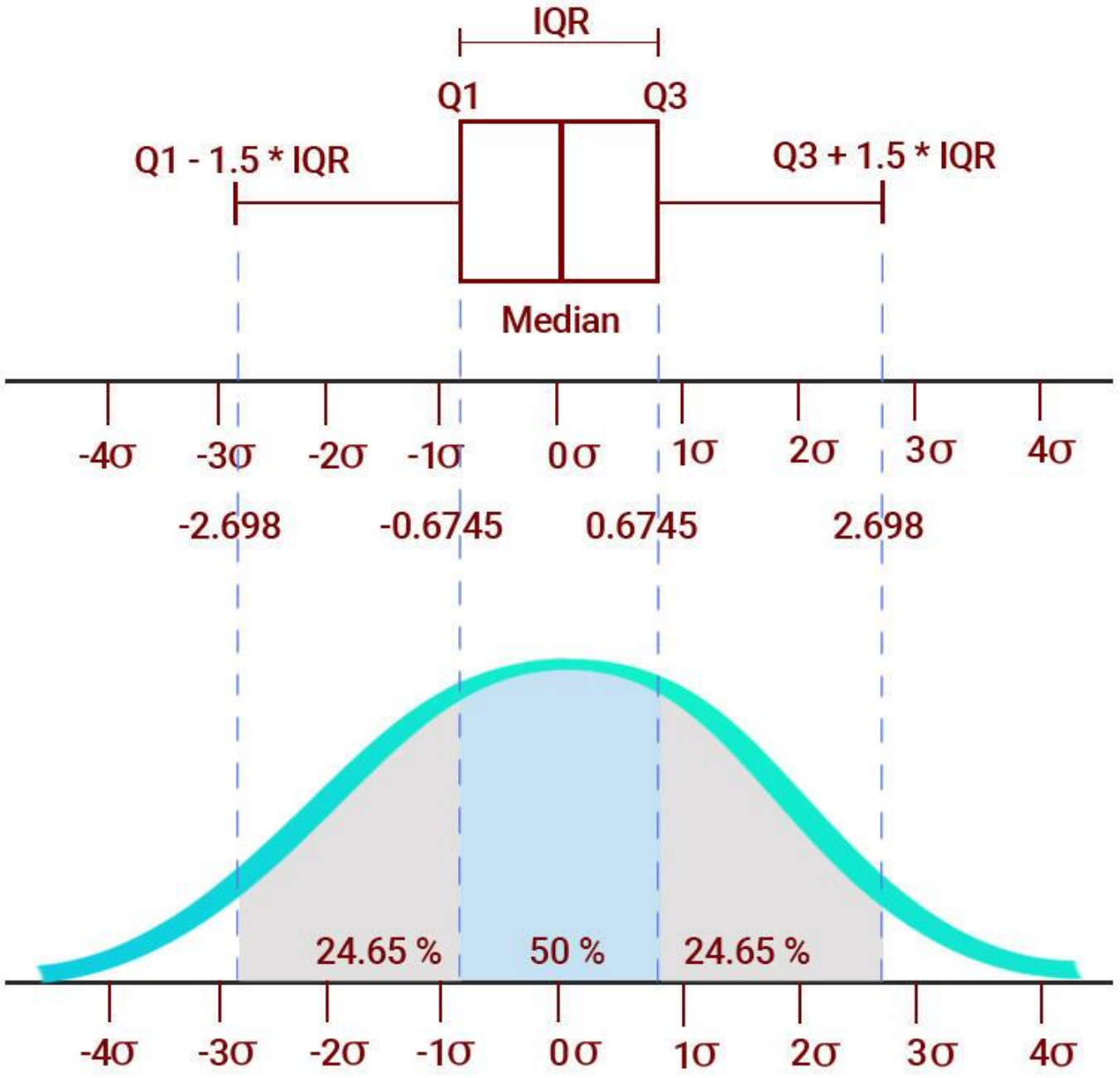
INFERENTIAL STATISTICS




IQR
Q1
Q3
Q1 - 1.5 * IQR
Q3 + 1.5 * IQR
Median
-4σ
-3σ
-2σ
-1σ
0σ
1σ
2σ
3σ
4σ
-2.698
-0.6745
0.6745
2.698
24.65 %
50 %
24.65 %

Census and Survey

Census:

Gathering data from the whole population of interest. For example, elections, 10-year census, etc.

Survey:

Gathering data from the sample in order to make conclusions about the population.

For example, opinion polls, quality control checks in manufacturing units, etc

Parameter and Statistic

Parameter: A descriptive measure of the population.

For example, population mean, population variance, population standard deviation, etc.

Statistic: A descriptive measure of the sample.

For example, sample mean, sample variance, sample standard deviation, etc.

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

  a) Population b) Sample

-

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

  a) Population b) Sample

- You want to know what sports teens prefer so you send out a survey to all the students in your high school.

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

  a) Population b) Sample

- You want to know what sports teens prefer so you send out a survey to all the students in your high school.

  a) Population b) **Sample**

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

  a) Population b) Sample

- You want to know what sports teens prefer so you send out a survey to all the students in your high school.

  a) Population b) **Sample**

- You want data on the shoe size of all West students, so you interview every student at school.

Identify Population Data or Sample Data?

- The US Government takes a census of its citizens every 10 years to gather information.

  a) Population b) Sample

- You want to know what sports teens prefer so you send out a survey to all the students in your high school.

  a) Population b) **Sample**

- You want data on the shoe size of all West students, so you interview every student at school.

  a) Population b) Sample

Identify as Parameter or Statistics?

- You want to know the mean income of the people who subscribe to People magazine, so you question 100 subscribers.

Identify as Parameter or Statistics?

- You want to know the mean income of the people who subscribe to People magazine, so you question 100 subscribers.

  a) Parameter b) **Statistic**

- You want to know the average height of the students in this math class, so you have everyone in the class write their height on a sheet of paper.

Identify as Parameter or Statistics?

- You want to know the mean income of the people who subscribe to People magazine, so you question 100 subscribers.

  a) Parameter b) **Statistic**

- You want to know the average height of the students in this math class, so you have everyone in the class write their height on a sheet of paper.

  a) Parameter b) Statistic

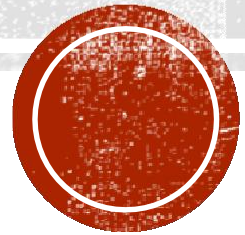
Parameter and Statistics

- Greek – Population Parameter
- Mean – μ
- Variance – σ^2
- Standard Deviation - σ

- Roman – Sample Statistic
- Mean – $\bar{x}$
- Variance – s^2
- Standard Deviation - s


A EXECUTIVE OFFICES

CONFIDENCE LEVELS AND CONFIDENCE INTERVALS



Today's Chanakya predicts landslide victory for BJP, allies

DC CORRESPONDENT
NEW DELHI, NOV. 5

DEAD HEAT

	India Today-Cicero	News X-CNX	ABP-Nielsen	Today's Chanakya	Times Now-C Voter	Average
BJP+	120 ±7	95	108	155	111 ±10	118
JD(U)+	117 ±6	130	130	83	122 ±10	118
Others	6	13	5	5	10	7

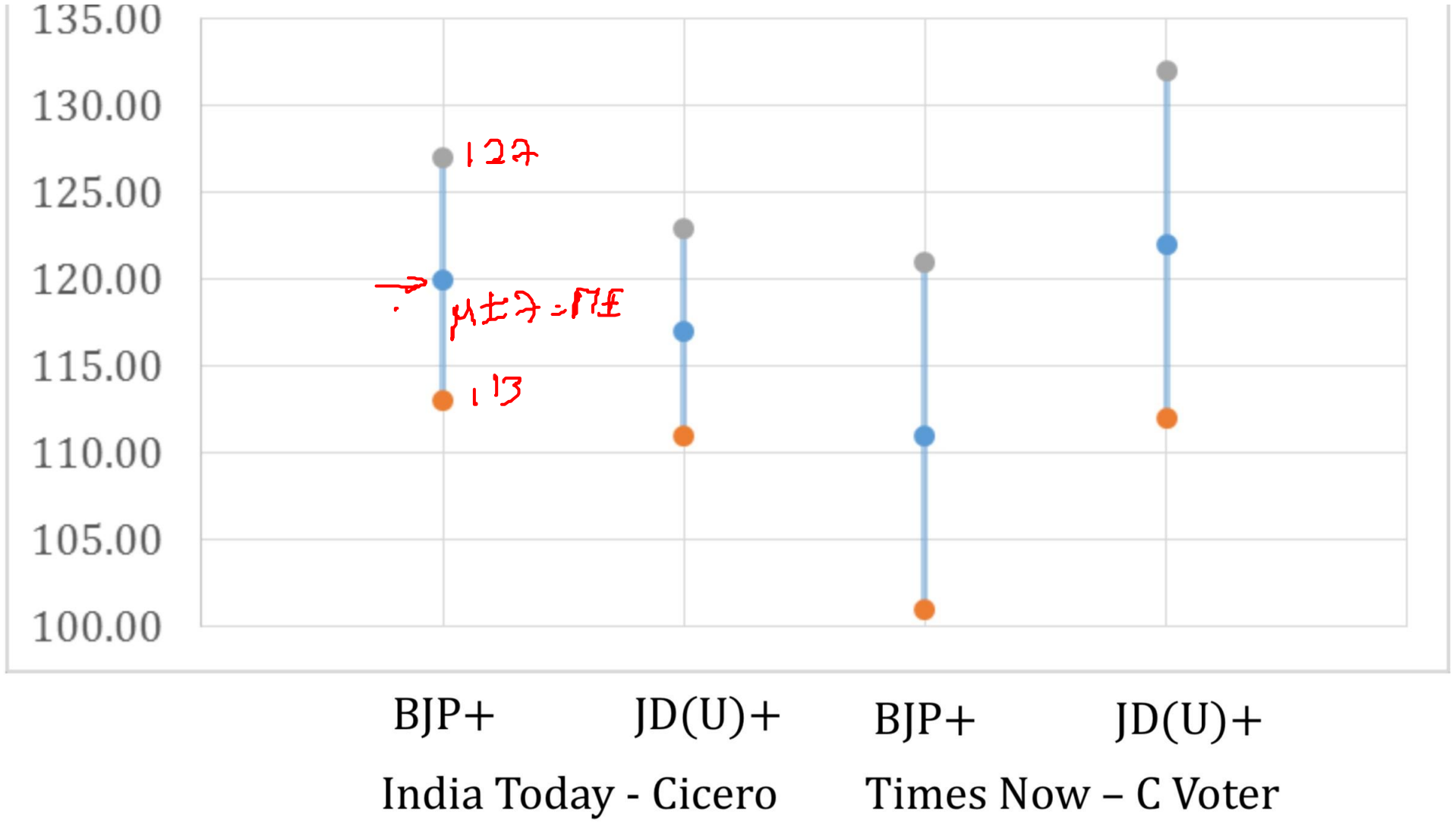
The Bihar Assembly elections appear to be living up to its billing as something of a cliffhanger, with a series of exit polls offering varied verdicts. At the end of the fifth and final round of voting on Thursday, the pollsters released their predictions, with four of them favouring the Grand Secular Alliance and two betting on the NDA.

Today's Chanakya, that hit the bull's eye in predicting the spectacular maiden electoral entry of Arvind Kejriwal's Aam Aadmi Party in 2013, and later followed it up by coming bang on in the 2014 Lok Sabha polls, forecast a near two-thirds majority for the BJP-led NDA, saying it could win 155 seats in an Assembly of 243. The pollster also claimed the Grand Alliance could win 83 seats.

In contrast, C-Voter, Neilsen and NewsX gave clear wins for the Grand Alliance led by Chief Minister Nitish Kumar. However, C-Voter also noted that the BJP would be the single largest party in the Assembly, winning 91 seats. The pollster's verdict is a close call, with the Grand Alliance pegged to win about 122 seats, against 111 for the NDA.

NewsX predicted 130 seats for the Grand Alliance against 90 for the NDA. The India Today-Cicero poll predicted 113 to 127 seats for the NDA and 111 to 123 for the JD(U)-led alliance. India TV forecast 112 to 132 seats for the JDU-led alliance and 101 to 121 for the NDA.

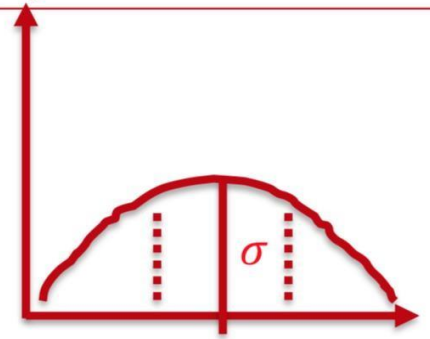
In the 2010 Bihar elections, the NDA, comprising the JD(U) and BJP, had swept the polls, winning 115 and 91 seats respectively, with the RJD getting 22 seats and the Congress just four.


135.00
130.00
125.00
120.00
115.00
110.00
105.00
100.00
127
120
113
BJP+
JD(U)+
BJP+
JD(U)+
India Today - Cicero
Times Now – C Voter

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.

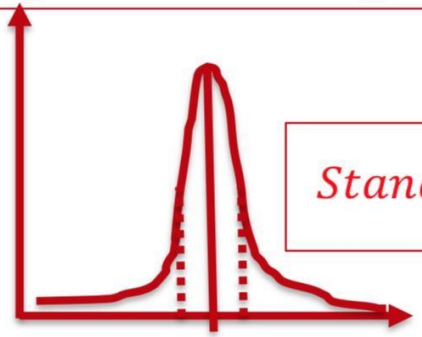
Publish Date	Source	Polling Organisation	NDA	UPA	Other
12 May 2014	[177]	CNN-IBN – CSDS – Lokniti	**276 (±6)**	97 (±5)	148 (±23)
	[177][178]	India Today – Cicero	**272 (±11)**	115 (±5)	156 (±6)
	[177][179]	News 24 – Chanakya	**340 (±14)**	70 (±9)	133 (±11)
	[177]	Times Now – ORG	**249**	148	146
	[177][180]	ABP News – Nielsen	**274**	97	165
	[177]	India TV – CVoter	**289**	101	148
14 May 2014	[181][182]	NDTV – Hansa Research	**279**	103	161
12 May 2014	[177]	Poll of Polls	**283**	105	149
16 May 2014	**Actual Results**[2]		**336**	**58**	**149**

Population distribution



μ

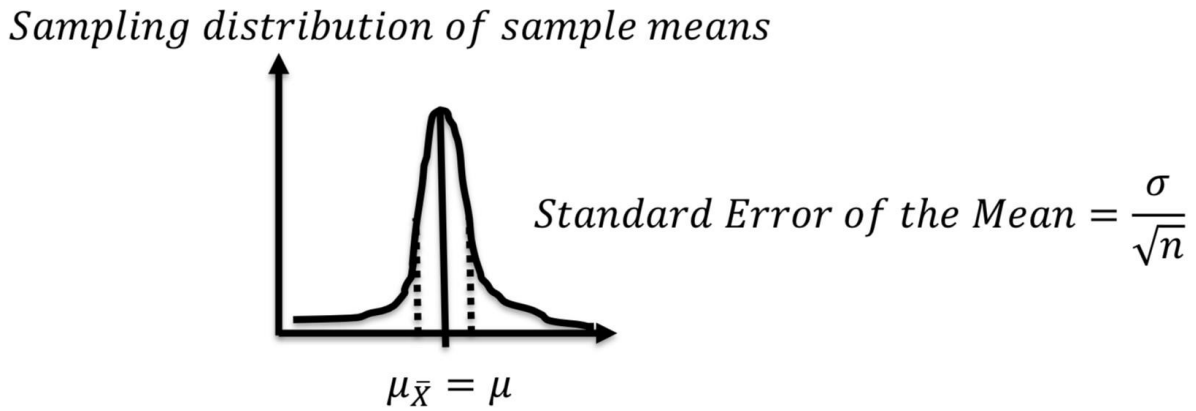
Sampling distribution of sample means



$\mu_{\bar{X}} = \mu$

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

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.



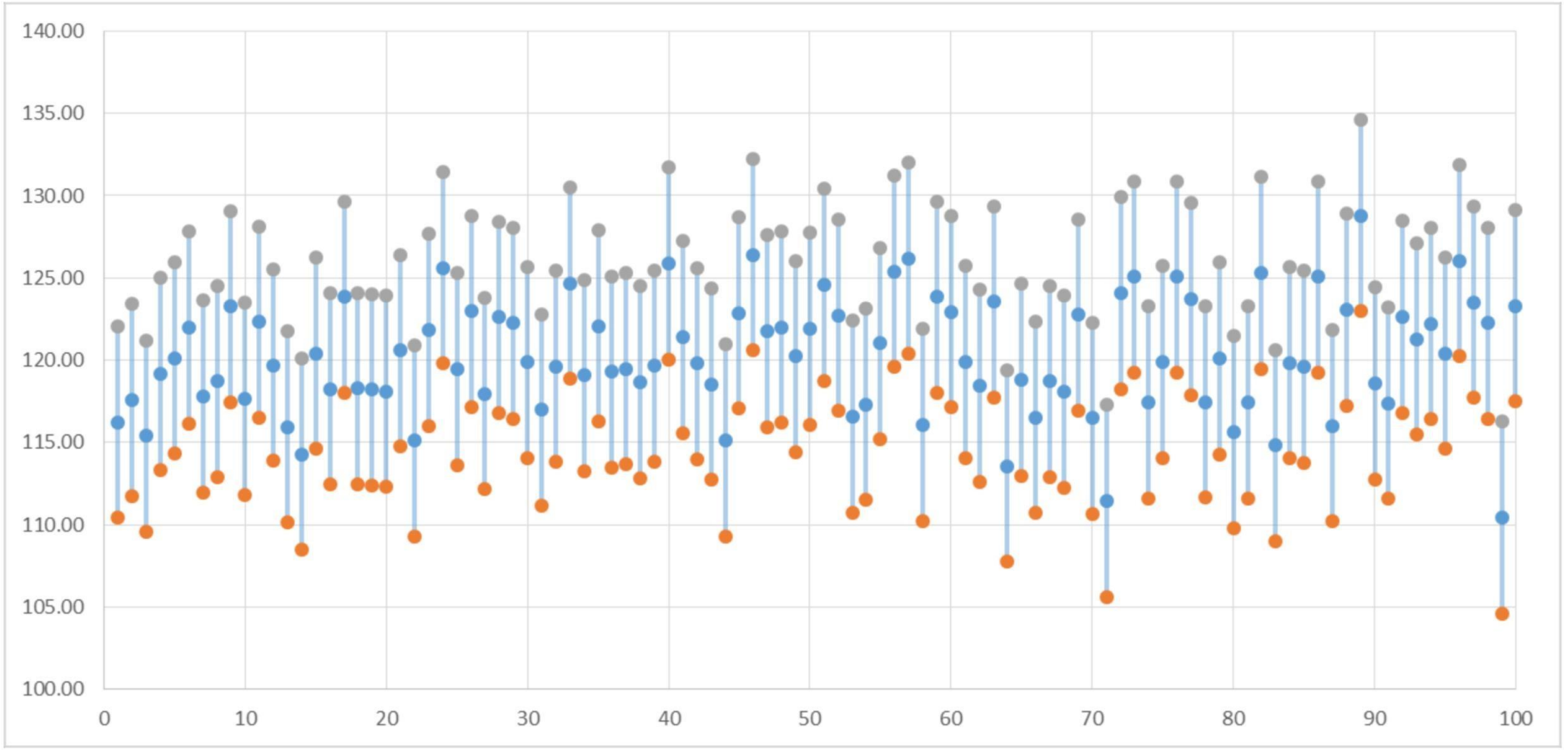
$$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$$

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.

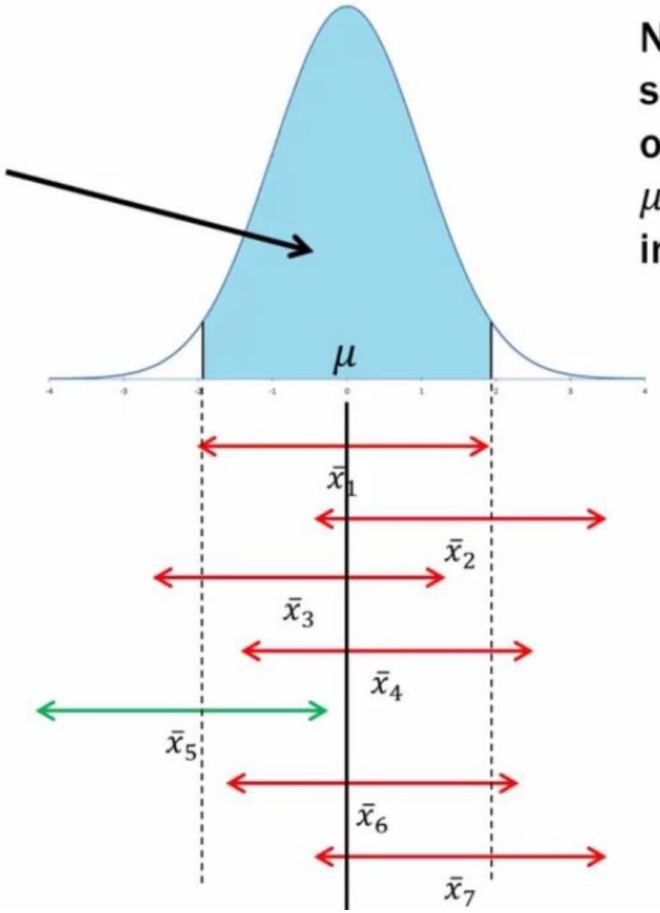
Confidence Level and Interval - Excel



94 of the 100 intervals contain the population mean.

Confidence Level and Interval

95% of all sample means $(\bar{x})$ are in here. THEN THIS.



Notice that as soon as a sample mean steps outside the dotted line, μ is no longer in its interval.

Many samples of the same size. THESE COME FIRST.

Samples of the same size have the same standard error $\sigma_{\bar{x}}$. So the 95% "width" is the same for all samples of that size.

HUFFPOST POLLSTER

FIND A CHART | Try 'New York,' 'Obama' or 'health care'

AdChoices

← PREVIOUS POLL | NEXT POLL →

POLL UPDATE

2016 National Republican Primary - Trump 40%, Cruz 17% (Ipsos/Reuters (Web) 2/13-2/17)

Population	1,473 Adults
Margin of Error	±2.9 percentage points
Polling Method	Internet
Source	Ipsos/Reuters [PDF]

This poll asked respondents 2 questions tracked by HuffPost Pollster. Read our FAQ.

Margin of Error is the range of expected variation for a given survey result or, more specifically, to how confident we can be that, if repeated using the same methodology, the results of a survey would fall within that range of variation.

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

1) 2016 National Republican Primary

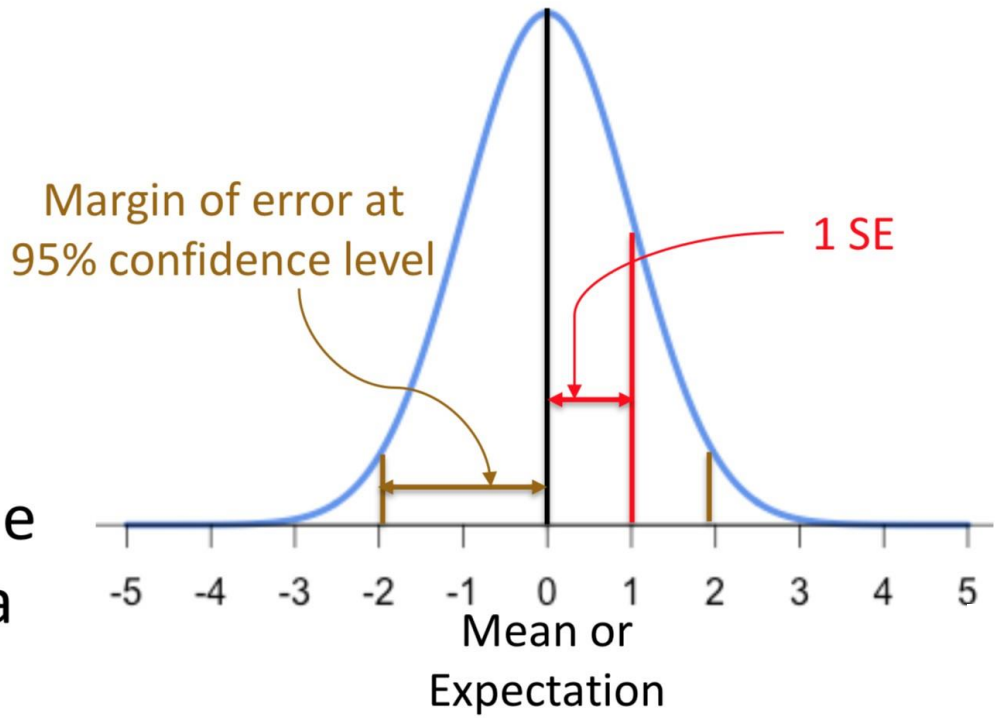
Asked of 476 Republican registered voters

Jeb Bush (R)	9%
Ben Carson (R)	10%
Ted Cruz (R)	17%
Jim Gilmore (R)	1%
John Kasich (R)	9%
Marco Rubio (R)	11%
Donald Trump (R)	40%
Wouldn't vote	4%

SE, Margin of Error, Confidence Interval and Sample Size

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

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

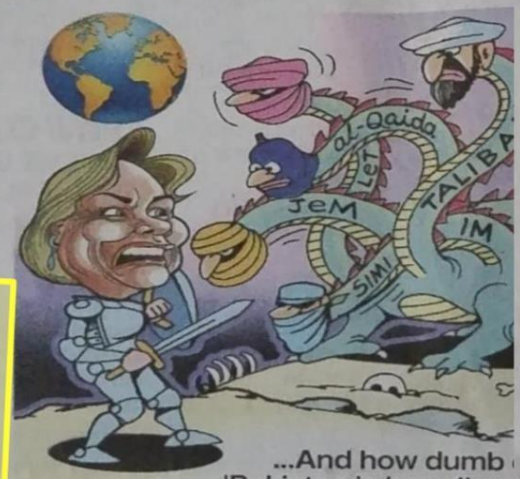


Margin of error is the **maximum expected difference** between the true population parameter and a sample estimate of that parameter.

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

Poll vault: Hillary leads Trump who leads Hillary who leads Trump

Chidanand.Rajghatta
@timesgroup.com



...And how dumb
'Pakistan is breeding

Public opinion polls are rather like children in a garden, digging things up all the time to see how they're growing, a British writer remarked perceptively, and nowhere is it truer than across the pond, where weekly surveys are the norm in the months leading up to the presidential elections.

Depending on the mode, the method, questions, sample size etc, polls conducted at the same time can show a wide range of results, allowing every side to claim imminent victory — although elections are still more than four months away, a week itself is said to be a long time in politics.

This week's crop of polls shows Hillary Clinton leading Donald Trump by double digits in an ABC poll (51-39), by only 5 points in an NBC/WSJ poll (46-41), and a virtual deadlock in a Quinnipaic University poll (42-40). A "Rasmussen Report" survey shows Trump with a four point lead (43-39).

What the polls do not show — or they do in fine print and small footnotes — is large margins of error (+/- 4% in ABC poll), small samples (often less than 1,000), and sketchy methods (Rasmussen is online and by telephone), among other limitations. There is also the business of the missing numbers (the sum of opinion never amounts to 100), suggesting that some 5-15% simply hold back their preferences/views — enough to change the results when they express it.

To top it all, popular votes are not what is going to decide the presidential election; electoral votes based on a winner-take-all system will. Al Gore won more popular votes than George Bush in 2000, but Bush nicked him in the Electoral College. Still, for the record, Clinton leads Trump in most polls, surveys, projections involving both popular vote and electoral college.

"I don't get how they can be deadlocked. This frankly worries me," former labour secretary and Harvard University professor Robert Reich said this week. "Trump hasn't put up a single TV ad, his campaign is in shambles, he has almost no field staff, he's spent almost zilch and his campaign bank is nearly empty, and he's been getting nothing but horrible press. Hillary Clinton has been blanketing swing states with ads, her campaign is being run like clockwork and it's huge, and she's pulling in and spending money like mad.

"More to the point, Trump is a bigoted petulant eg ac with the temperam hyena. She's exper competent, and int What's going on?" Rei dered. What's going o polls never tell the fu much less the end resu servatives claim there reservoir of Trump ters who do not shar tist, coastal view of th lican candidate as a bi fool. Call them Trexite

Besides, there are variables in the electi some pollsters are de new way of forecastin le results — from polls to new algorithms tha to account past resul fic issues, demograp Some even project possible result in percentages to get around the certitude. Nate Silver, a baseball statistician who moved to politics after accurately calling the 2008 and 2012 elections, puts Hillary Clinton's in-chief, imitating something that had worked for Margaret Thatcher. The last pitch misfired badly. Polls did not account for it. It's the kind of pitfall both the camps keep watching for as much as the polls.

Depending on the mode, the method, questions, sample size etc, polls conducted at the same time can show a wide range of results, allowing every side to claim imminent victory — although elections are still more than four months away, a week itself is said to be a long time in politics.

This week's crop of polls shows Hillary Clinton leading Donald Trump by double digits in an ABC poll (51-39), by only 5 points in an NBC/WSJ poll (46-41), and a virtual deadlock in a Quinnipaic University poll (42-40). A "Rasmussen Report" survey shows Trump with a four point lead (43-39).

What the polls do not show — or they do in fine print and small footnotes — is large margins of error (+/- 4% in ABC poll), small samples (often less than 1,000), and sketchy methods (Rasmussen is online and by telephone), among other limitations. There is also the business of the missing numbers (the sum of opinion never amounts to 100), suggesting that some 5-15% simply hold back their preferences/views — enough to change the results when they express it.

Confidence Intervals

$(\bar{x} - M.E, \bar{x} + M.E)$

A survey was taken of US companies that do business with firms in India. One of the survey questions was: Approximately how many years has your company been trading with firms in India? A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of US companies trading with firms in India.

$\bar{x}$

→ confidence level

Confidence Intervals

- $n = 44$
- $\bar{x} = 10.455$
- $\sigma = 7.7$

S.E

$Z \times \sigma/\sqrt{n}$

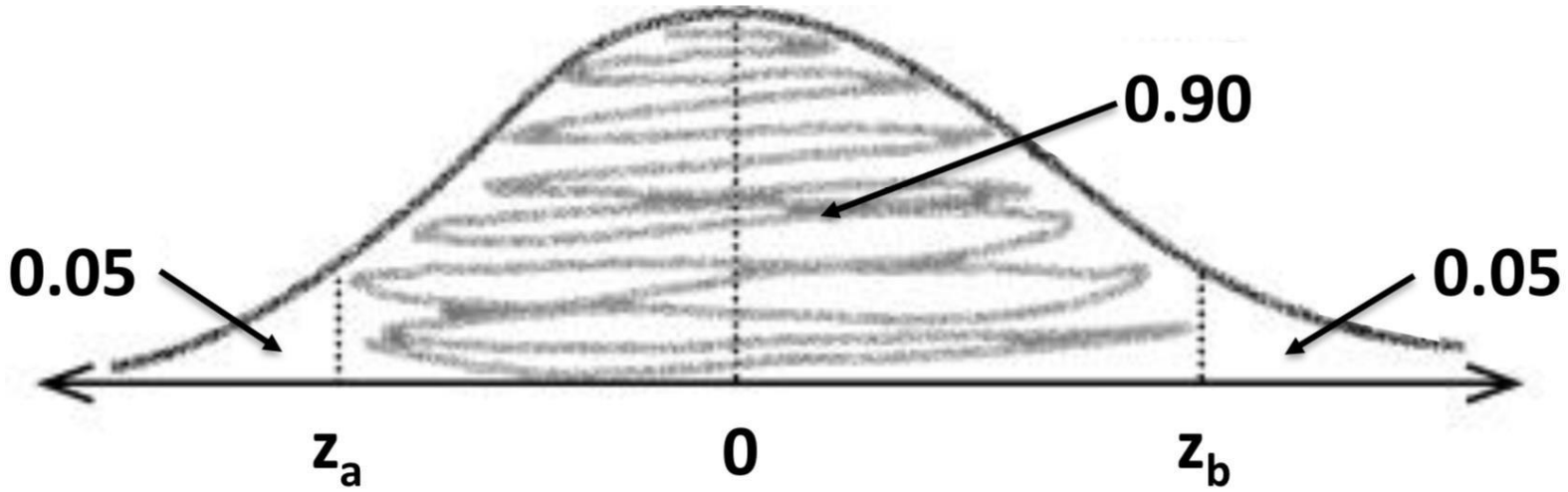
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

∴ 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 - table

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

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) in R.

Confidence Intervals

$$Margin\ of\ error\ at\ 90\%\ Confidence\ Level = 1.645 * \frac{7.7}{\sqrt{44}} = 1.91$$

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

$$\bar{X} - 1.91 < \mu < \bar{X} + 1.91$$

Since the sample mean is 10.455 years, we get the confidence interval for 90% as 8.545 < μ < 12.365.

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 8.545 and 12.365 years.

Shortcuts for Calculating Confidence Intervals

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

The lung function in 57 people is tested using FEV1 (Forced Expiratory Volume in 1 Second) measurements. The mean FEV1 value for this sample is 4.062 litres and standard deviation, *s* is 0.67 litres. Construct the 95% Confidence Interval.

Margin Error = $Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

FEV1 values of 57 male medical students

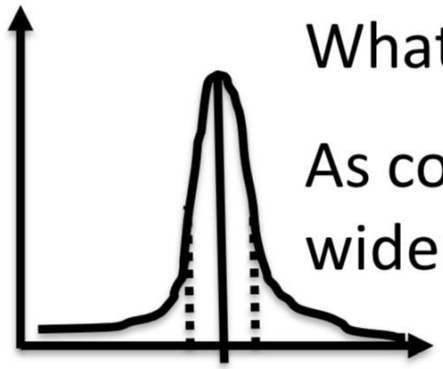
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

2.85	2.85	2.98	3.04	3.10	3.10	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
4.14	4.16	4.20	4.20	4.30	4.30	4.32	4.44	4.47	4.47
4.47	4.50	4.50	4.56	4.68	4.70	4.71	4.78	4.80	4.80
4.90	5.00	5.10	5.10	5.20	5.30	5.43			

$$95\%\ CI: \left(4.062 - 1.96 * \frac{0.67}{\sqrt{57}}, 4.062 + 1.96 * \frac{0.67}{\sqrt{57}}\right)$$

$$= (3.89, 4.23)$$

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 becomes narrower.

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