INFERENTIAL STATISTICS

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.

  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 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.

  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

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 – **x**
- Variance – s^2
- Standard Deviation - **s**

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

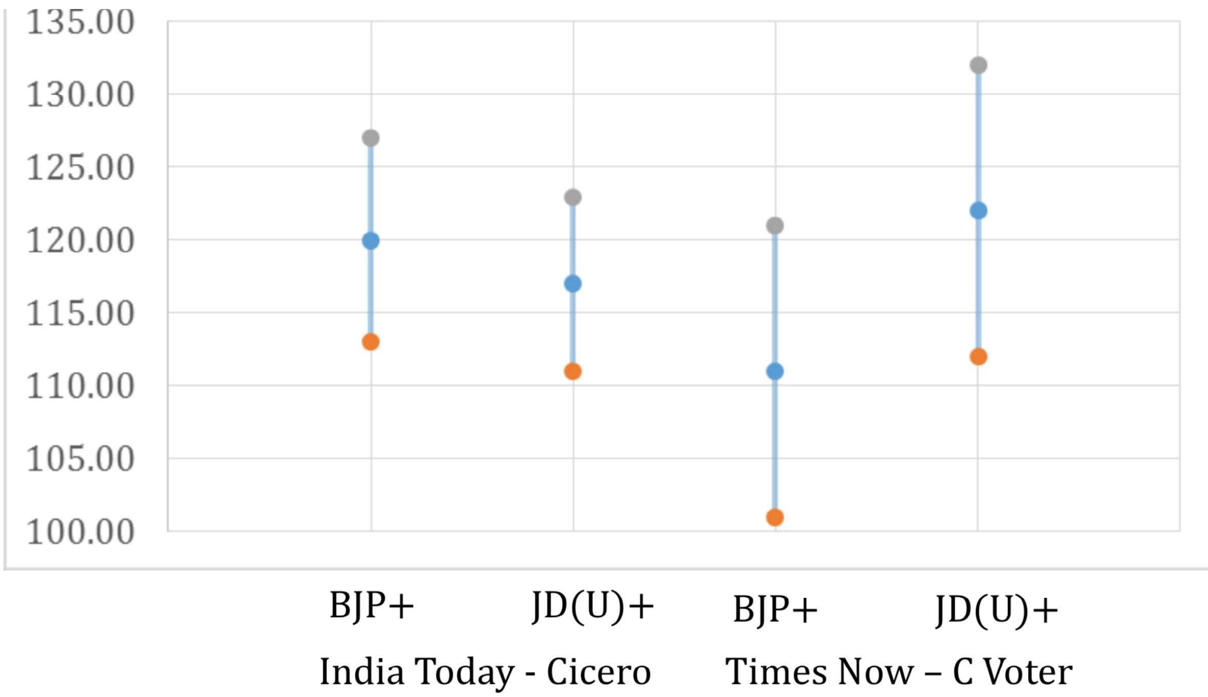
The Bihar Assembly elections appear to be living up to its billing as something of a cliffhanger, with a series of exit polls offering varied verdicts. At the end of the fifth and final round of voting on Thursday, the pollsters released their predictions, with four of them favouring the Grand Secular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

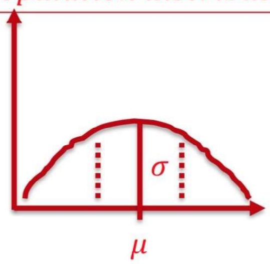
In the 2010 Bihar elections, the NDA, comprising the JD(U) and BJP, had swept the polls, winning 115 and 91 seats respectively, with the RJD getting 22 seats and the Congress just four.


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

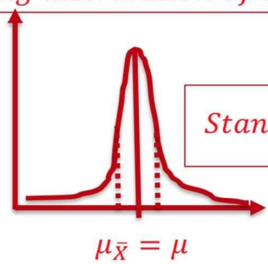
When we use samples to provide population estimates, we cannot be CERTAIN that they will be accurate. There is an amount of uncertainty, which needs to be calculated.

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

$Population\ distribution$

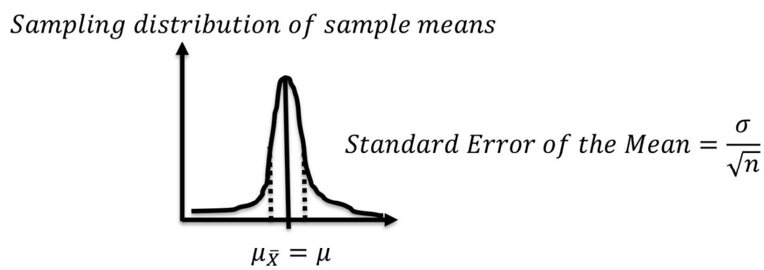


$Sampling\ distribution\ of\ sample\ means$



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

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.

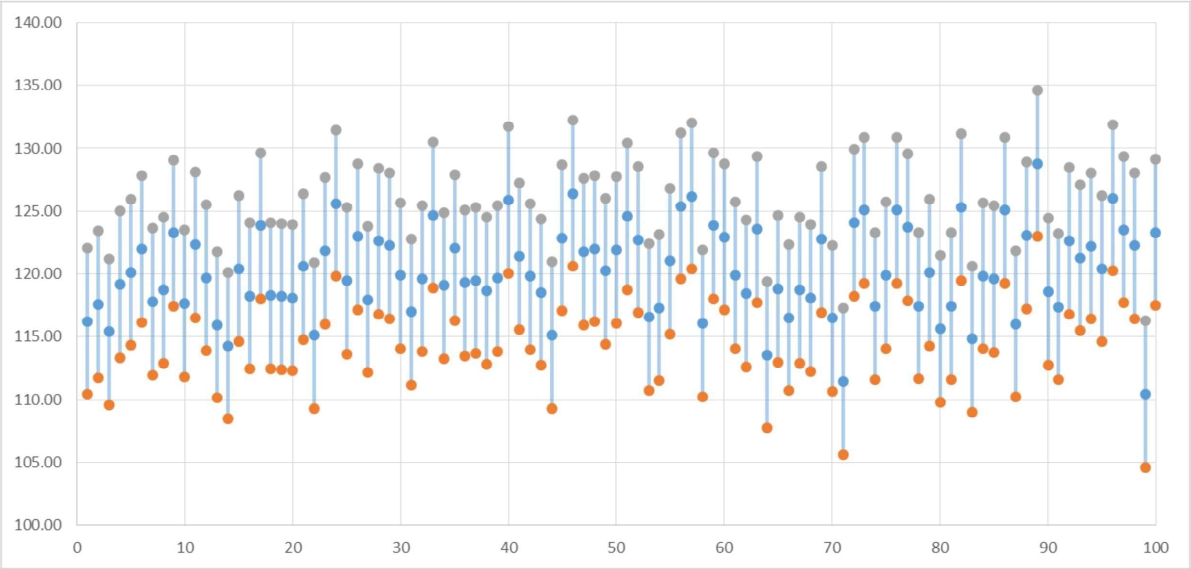


We have seen that ~ 95% of the samples will have a mean value within the interval +/- 2 SE of the population mean *(recall the Empirical Rule for Normal Distribution)*.

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

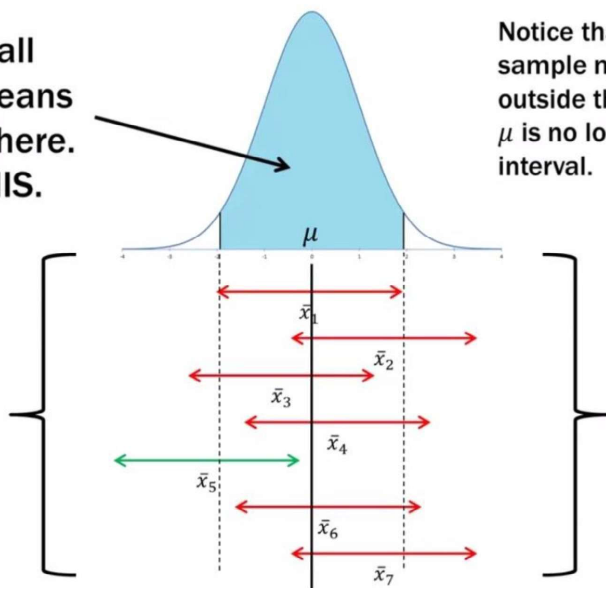


94 of the 100 intervals contain the population mean.

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

← 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]

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.

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

1) 2016 National Republican Primary

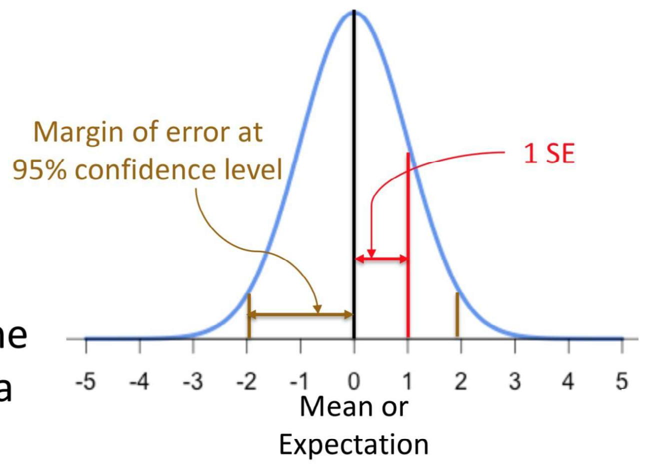
Asked of 476 Republican registered voters

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

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 share... 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 poll... 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.

...And how dumb ...
'Pakistan is breeding ...

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.

SE, Margin of Error, Confidence Interval and Sample Size

If the desired margin of error at 95% confidence level is 1%, what should be the sample size?

$$0.01 = 1.96 * \sqrt{\frac{0.49 * 0.51}{n}}$$

$$\therefore n = \left(\frac{1.96}{0.01} * \sqrt{0.49 * 0.51}\right)^2 = 9600$$

Confidence Intervals

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.

Confidence Intervals

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

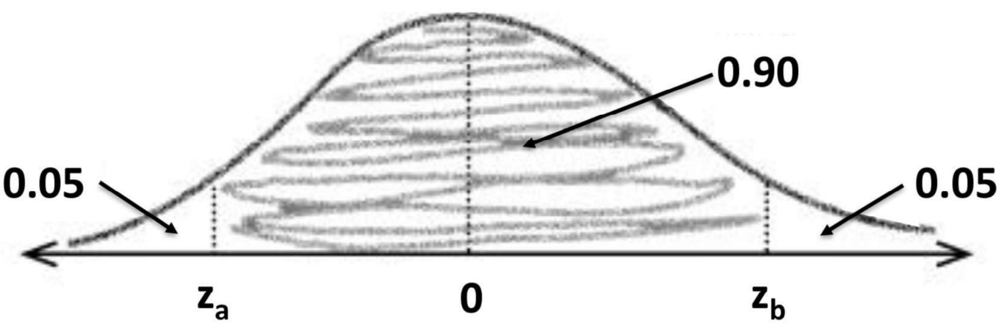
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

$\therefore$ 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

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

Recall Confidence Interval for the Population Mean is Sample Mean $\pm$ Margin of Error

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

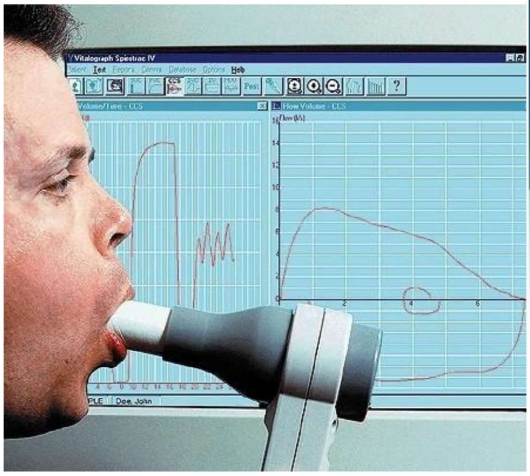
Since the sample mean is 10.455 years, we get the confidence interval for 90% as 8.545 < μ < 12.365.

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.



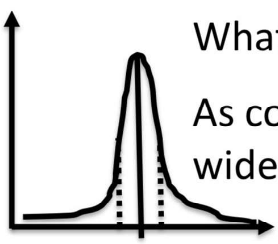
FEV1 values of 57 male medical students

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?

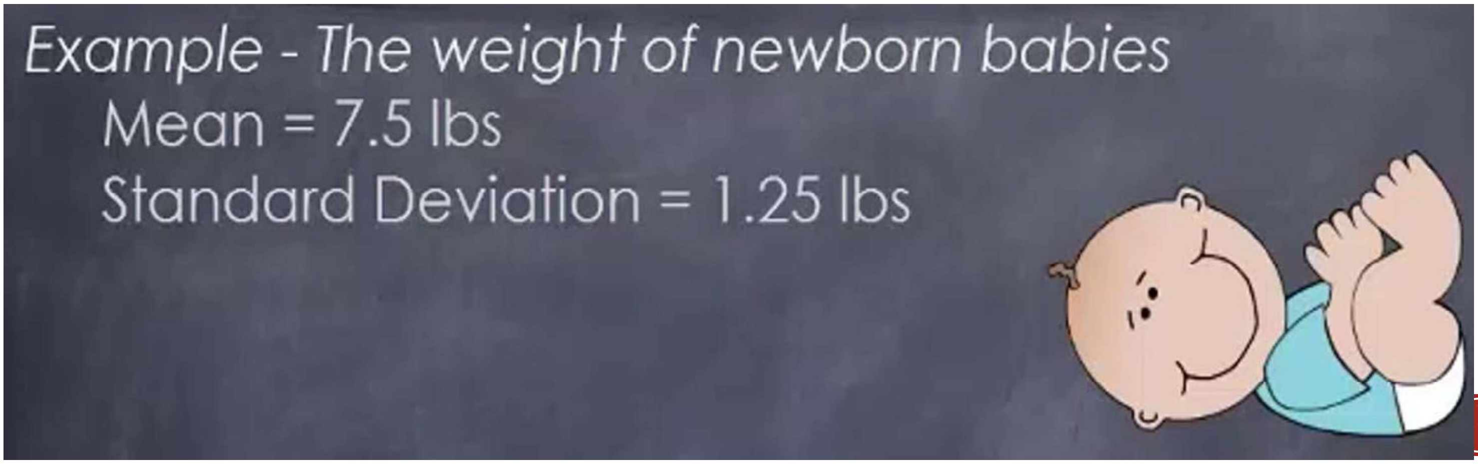
As confidence level increases, the confidence interval becomes wider and *vice-versa*.

What happens to the confidence interval as sample size changes?

As sample size increases, the confidence interval becomes narrower.
Remember $(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$.

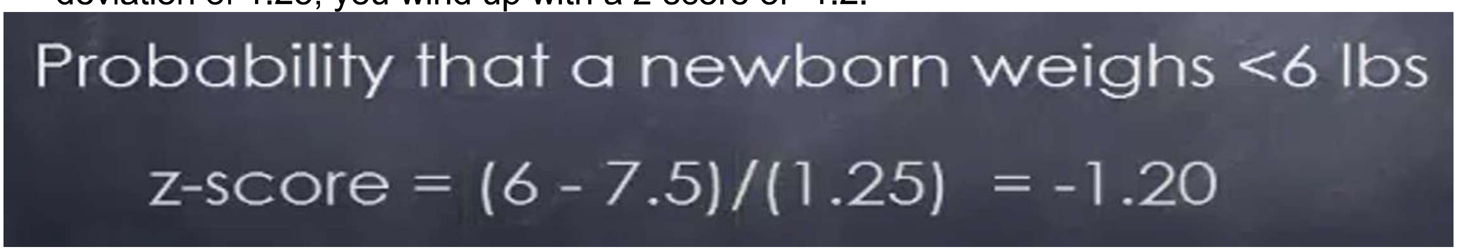
Application of Z-Score

1. Take a look at the weight of new-born babies. Suppose that the mean weight of new-borns is 7.5 pounds and the standard deviation is 1.25 pounds. Say you're interested in determining the probability that a new-born weighs less than 6 pounds. How do you do that?



Application of Z-Score

- The first thing you do is calculate the z-score. To figure out the z-score, take the difference between 6 and 7.5 to arrive at -1.5. When you divide -1.5 by the standard deviation of 1.25, you wind up with a z-score of -1.2.



- When you're looking at a z-table, the first step is to determine which row you want to look at. That's going to be represented by the spot to the left of the decimal point and the first spot to the right of the decimal point, which is -1.2. The second step is to look at the digit that is two spots to the right of the decimal point, which is a 0. This determines the column you look at.

Application of Z-Score

- The point at which both the row and the column intersect one another indicates the probability.

- The intersection in this scenario is 0.1151. That tells you that the probability associated with the baby weighing less than 6 pounds is 0.1151.

z	0.00	0.01	0.02	0.0
-3.4	0.0003	0.0003	0.0003	0.00
-3.3	0.0005	0.0005	0.0005	0.00
-3.2	0.0007	0.0007	0.0006	0.00
-3.1	0.0010	0.0009	0.0009	0.00
-3.0	0.0013	0.0013	0.0013	0.00
-2.9	0.0019	0.0018	0.0018	0.00
-2.8	0.0026	0.0025	0.0024	0.00
-2.7	0.0035	0.0034	0.0033	0.00
-2.6	0.0047	0.0045	0.0044	0.00
-2.5	0.0062	0.0060	0.0059	0.00
-2.4	0.0082	0.0080	0.0078	0.00
-2.3	0.0107	0.0104	0.0102	0.00
-2.2	0.0139	0.0136	0.0132	0.01
-2.1	0.0179	0.0174	0.0170	0.01
-2.0	0.0228	0.0222	0.0217	0.02
-1.9	0.0287	0.0281	0.0274	0.02
-1.8	0.0359	0.0351	0.0344	0.03
-1.7	0.0446	0.0436	0.0427	0.04
-1.6	0.0548	0.0537	0.0526	0.05
-1.5	0.0668	0.0655	0.0643	0.06
-1.4	0.0808	0.0793	0.0778	0.07
-1.3	0.0968	0.0951	0.0934	0.09
-1.2	0.1151	0.1131	0.1112	0.10
-1.1	0.1357	0.1335	0.1314	0.12
-1.0	0.1587	0.1562	0.1539	0.15
-0.9	0.1841	0.1814	0.1788	0.17

Application of Z-Score

2. What if you're interested in looking at the probability that a new-born might weight more than 10 pounds? You could calculate the z-score much the same way you did with the previous calculation:

Probability that a newborn weighs >10 lbs

$$\text{z-score} = (10 - 7.5)/(1.25) = 2.00$$

The same process is involved in looking at the z-table. Look at 2.0 and which row that entails. Since there are no numbers other than 0 to the right of the decimal point, you look at the very first column, which is the 0.00 column. **Note, this table looks different from the one above.** The z scores in the left column are negative instead of positive. For a full PDF on z-score tables,

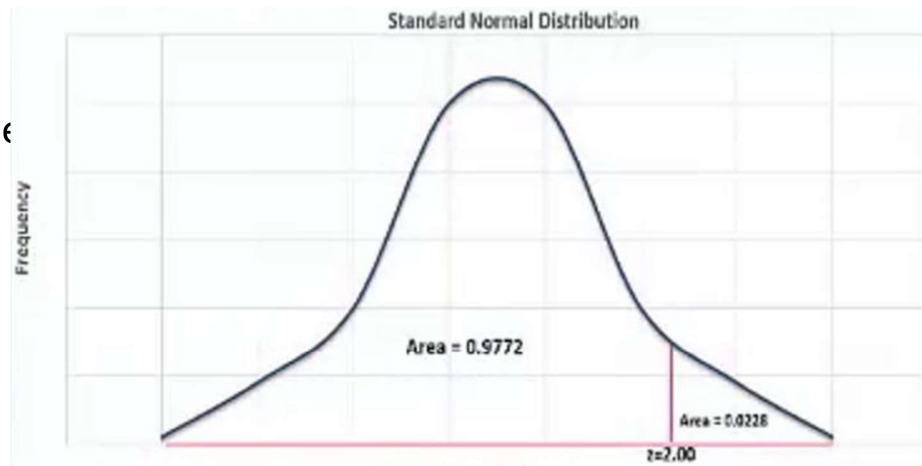
Application of Z-Score

•You wind up with a probability of 0.9772. Remember, you are trying to determine the probability that a new-born weighs more than 10 pounds. The value given to you here is telling you the area underneath the curve to the left of +2 standard deviations, as you see on the graph right here:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.52
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.56
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.60
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.64
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.67
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.71
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.74
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.77
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.80
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.83
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.85
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.87
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.89
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.91
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.92
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.94
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.95
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.96
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.96
1.9	[illegible]	0.9719	0.9726	0.9732	0.9738	0.9744	0.97
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.98
2.1	[illegible]	0.9826	0.9830	0.9834	0.9838	0.9842	0.98
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.98
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.99

Application of Z-Score

- You will need to subtract that value from 1 to get the area that's on the right-hand side of that line. The probability that a new-born would be greater than 10 pounds would be 1 minus 0.9772. The probability is equal to 0.0228.



HYPOTHESIS TESTS

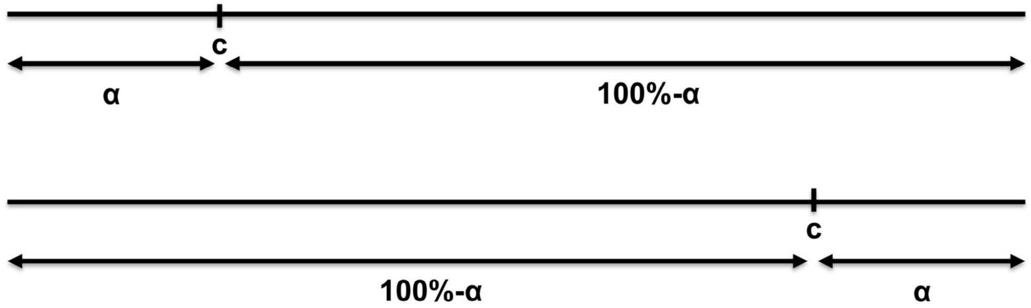
Examples of Hypotheses

- Two hypotheses in competition:
  - H_0: The NULL hypothesis, usually the most conservative.
  - H_1 or H_A: The ALTERNATIVE hypothesis, the one we are actually interested in.
- Examples of NULL Hypothesis:
  - The coin is fair
  - The new drug is no better (or worse) than the placebo
- Examples of ALTERNATIVE hypothesis:
  - The coin is biased (either towards heads or tails)
  - The coin is biased towards heads
  - The coin has a probability 0.6 of landing on tails
  - The drug is better than the placebo

Critical Region Up Close

One-tailed tests

The position of the tail is dependent on H_1.
If H_1 includes a < sign, then the **lower tail** is used.
If H_1 includes a > sign, then the **upper tail** is used.

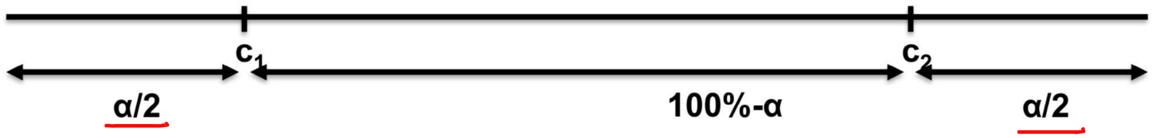


Critical Region Up Close

Two-tailed tests

Critical region is split over both ends. Both ends contain $\alpha/2$, making a total of α.

If H_1 includes a $\neq$ sign, then the two-tailed test is used as we then look for a change in parameter, rather than an increase or a decrease.



Common Test Statistics for Inferential Techniques

Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistics:

- z
- t

Closely related to Sampling Distribution of **Means**

- $\boxed{\chi^2}$ (Chi-squared)
- F

- Closely related to Sampling Distribution of **Variances**
- Derived from Normal Distribution

HYPOTHESIS TESTING STEPS

1. State the Null Hypothesis (H0) and Alternate Hypothesis (H1)
2. Choose the Level of Significance
3. Find Critical Values
4. Find test Statistic
5. Draw your conclusion

Application of Z-Test

- According to a recent survey, the daily one-way commuting distance of U.S. workers averages 13 miles with a standard deviation of 13 miles. An investigator wishes to determine whether the national average describes the mean commuting distance for all workers in the Chicago area. Commuting distances are obtained for a random sample of 169 workers form this area, and the mean distance is found to be 15.5 miles. Test the null hypothesis at the 0.05 level of significance.

Application of Z-Test

- **Review of the assumptions for z test:**

1. The sample size is large enough (>25) to satisfy the requirement of the central limit theorem.
2. Population standard deviation is known as 13 miles.
3. Scale of measurement of variable "commuting distance" is interval-ratio.

Application of Z-Test

- **Research Question:** Does the national average commuting distance describe the mean commuting distance for all workers in the Chicago area?
- Population of interest: All workers in the Chicago area
- Sample: Randomly selected 169 workers from Chicago area

- **Statistical hypothesis:**
- Null hypothesis (H_0): $\mu = 13$
- Alternative hypothesis (H_1): $\mu \neq 13$

  Where: μ is mean commuting distance for all Chicago workers.

Application of Z-Test

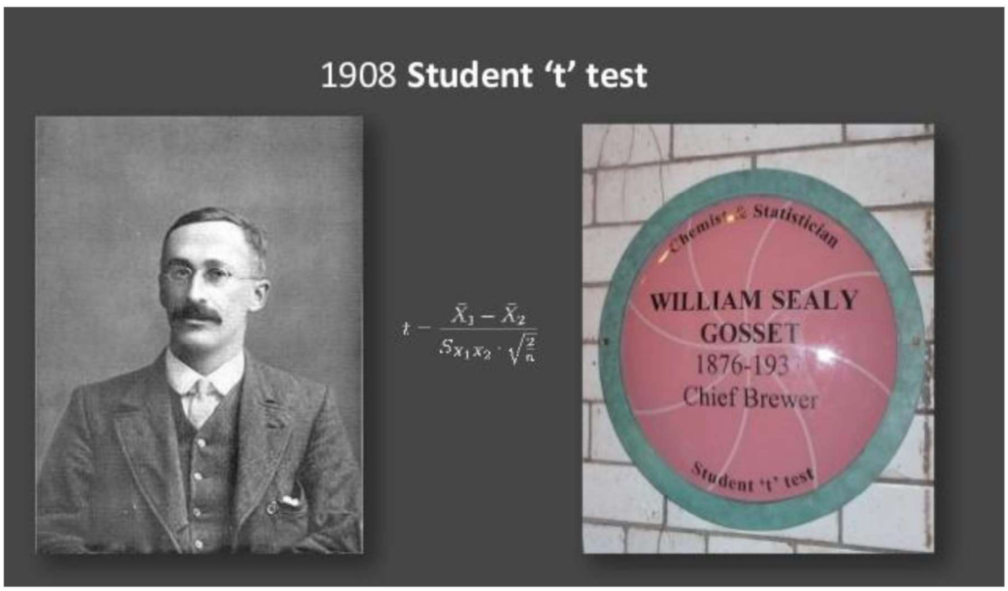
- **Decision Rule:** H_0 should be rejected if observed z equals to or is more positive than the upper critical z (1.96) or if observed z equals to or is more negative than the lower critical z (-1.96) at level of significance (α) of 0.05.
- **Calculations of test statistics (Z):**
- **Given**:
- x (sample mean) = 15.5;
- μ_0 (hypothetical population mean) = 13;
- n (sample size) = 169;
- σ (population standard deviation) = 13

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{15.5 - 13}{13 / \sqrt{169}} = 2.5 > 1.96$$

Application of Z-Test

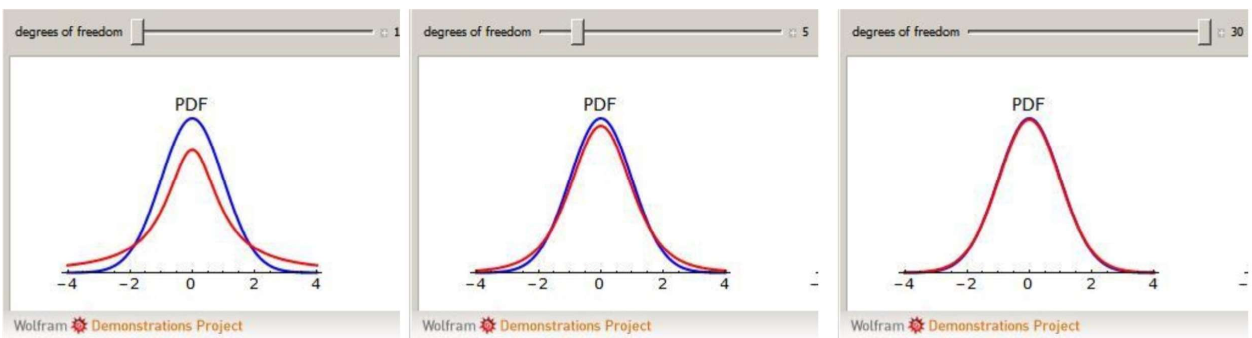
- **Decision:** Reject H_0
- **Interpretation:** The national average commuting distance does not describe the mean commuting distance for all workers in the Chicago area

t-Distribution



t-Distribution

If the sample size is small (<30), the variance of the population is not adequately captured by the variance of the sample. Instead of z-distribution, t-distribution is used. It is also the appropriate distribution to be used when population variance is **not known**, irrespective of sample size.



t-Distribution

$$t\ statistic\ (or\ t\ score), t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$

Degrees of freedom, ν: # of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.*

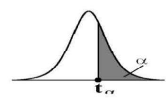
When estimating mean or proportion from a single sample, the # of independent observations is equal to *n-1*.

Confidence Interval to Estimate μ

- Population standard deviation UNKNOWN and the population normally distributed.
- $\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
  - Sample mean, standard deviation and size can be calculated from the data; *t* value can be read from the table or obtained from software.
  - α is the area in the tail of the distribution. For 90% Confidence Level, α=0.10. In a Confidence Interval, this area is symmetrically distributed between the 2 tails ($\alpha/2$ in each tail).

t-table

Percentage Points of the *t* Distribution; $t_{\nu,\alpha}$
$P(T > t_{\nu,\alpha}) = \alpha$



ν	α = 0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.170	2.326	2.432	2.576	2.807	3.291

t-Distribution - Example

The labeled potency of a tablet dosage form is 100 mg. As per the quality control specifications, 10 tablets are randomly assayed.

A researcher wants to estimate the interval for the true mean of the batch of tablets with 95% confidence. Assume the potency is normally distributed.

Data are as follows (in mg):

98.6	102.1	100.7	102.0	97.0
103.4	98.9	101.6	102.9	105.2

t-Distribution - Example

Mean, $\bar{x}$ = 101.24 mg

Standard deviation, s = 2.48

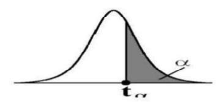
n = 10

$\nu = 10 - 1 = 9$

At 95% level, $\alpha = 0.05$, and $\therefore, \frac{\alpha}{2} = 0.025$

t-table

Percentage Points of the *t* Distribution; $t_{\nu,\alpha}$
$P(T > t_{\nu,\alpha}) = \alpha$



$t_{9,0.025} = 2.262$

	α												
ν	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.170	2.326	2.432	2.576	2.807	3.291

t-Distribution - Example

Mean, $\bar{x}$ = 101.24 mg, Standard deviation, s = 2.48

n = 10, $\nu = 10 - 1 = 9$

$$\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$101.24 - 2.262 * \frac{2.48}{\sqrt{10}} \leq \mu \leq 101.24 + 2.262 * \frac{2.48}{\sqrt{10}}$$

$99.47 \leq \mu \leq 103.01$

The batch mean is 101.24 mg with an error of +/-1.77 mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.47 mg and 103.01 mg.

t-Distribution – Example – R Output

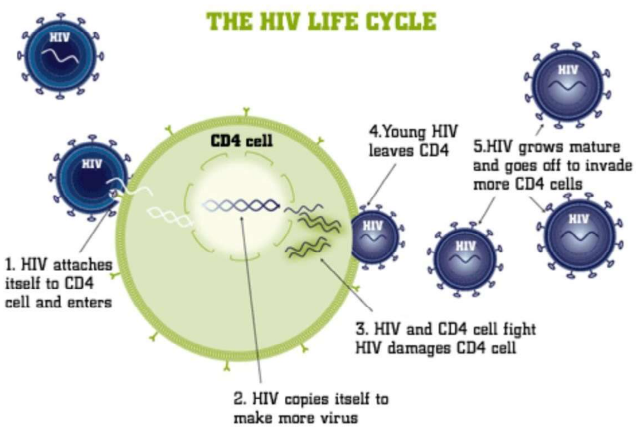
```
        One Sample t-test

data:  potency
t = 1.5824, df = 9, p-value = 0.148
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  99.46735 103.01265
sample estimates:
mean of x
   101.24
```

```
90 percent confidence interval:
  99.80355 102.67645
```

```
99 percent confidence interval:
  98.6934 103.7866
```

t-Distribution - Example

A researcher wants to examine CD4 counts for HIV+ patients at her clinic. She randomly selects a sample of 25 HIV+ patients and measures their CD4 levels. Calculate a 95% CI for population mean given the following sample results:



Variable	n	$\bar{x}$	SE of Mean	s	Min	Q1	Median	Q3	Max
CD4 (cells/μl)	25	321.4	14.8	73.8	208.0	261.5	325.0	394.0	449.0