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 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**

A EXECUTIVE OFFICES

[redacted]

CONFIDENCE LEVELS AND CONFIDENCE INTERVALS

Today's Chanakya predicts landslide victory for BJP, allies

DC CORRESPONDENT
NEW DELHI, NOV. 5

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.

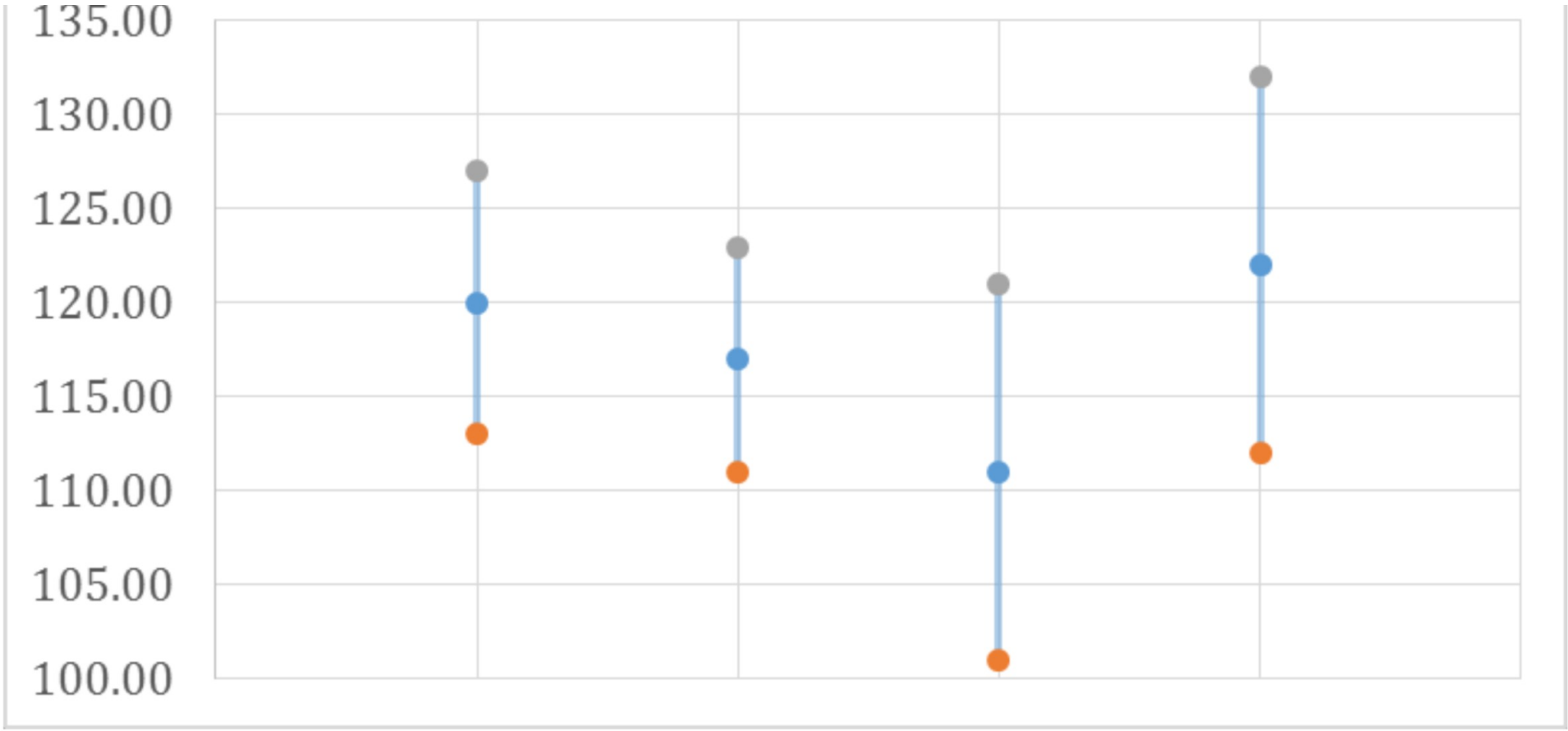
In contrast, C-Voter, Neilsen and NewsX gave clear wins for the Grand Alliance led by Chief Minister Nitish Kumar. However, C-Voter also noted that the BJP would be the single largest party in the Assembly, winning 91 seats. The pollster's verdict is a close call, with the Grand Alliance pegged to win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

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



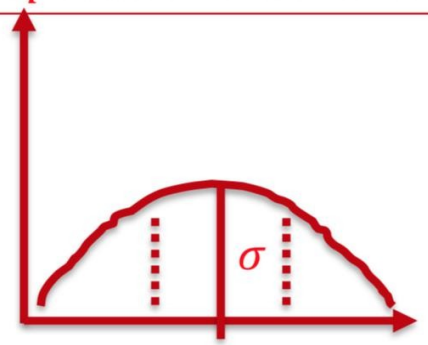
BJP+ JD(U)+ BJP+ JD(U)+

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.

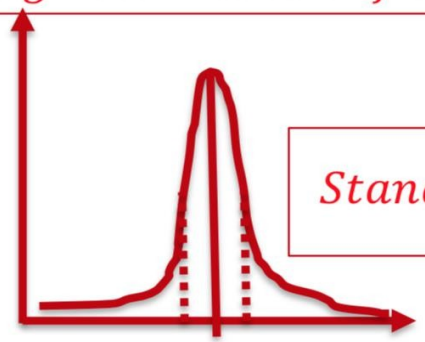
Publish Date	Source	Polling Organisation	NDA	UPA	Other
12 May 2014	[177]	CNN-IBN – CSDS – Lokniti	**276 (±6)**	97 (±5)	148 (±23)
	[177][178]	India Today – Cicero	**272 (±11)**	115 (±5)	156 (±6)
	[177][179]	News 24 – Chanakya	**340 (±14)**	70 (±9)	133 (±11)
	[177]	Times Now – ORG	**249**	148	146
	[177][180]	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

$Sampling\ distribution\ of\ sample\ means$

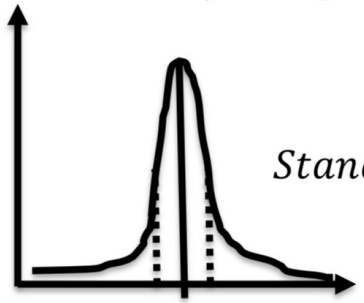


$$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$$

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

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.

$$Sampling\ distribution\ of\ sample\ means$$



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

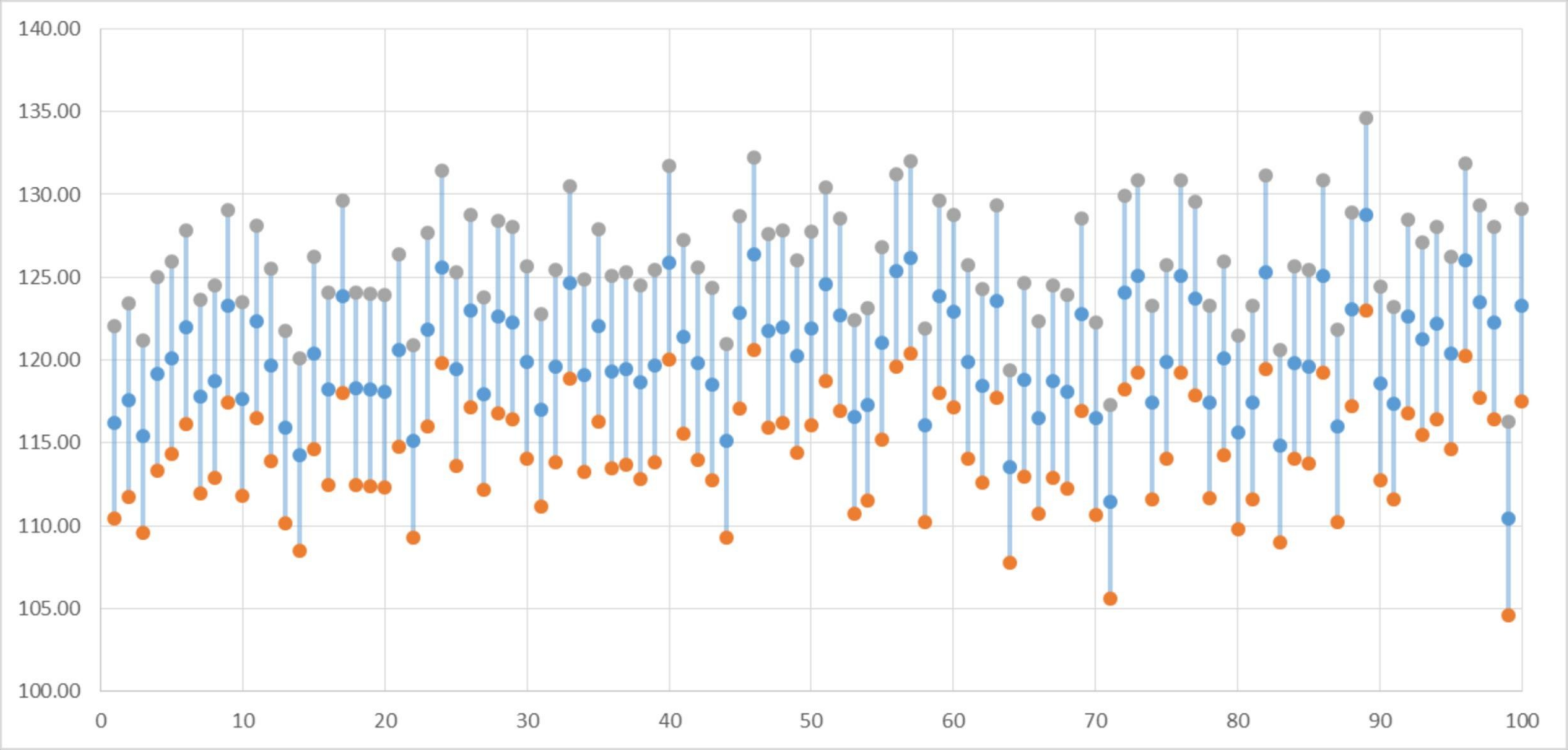
$$\mu_{\bar{X}} = \mu$$

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

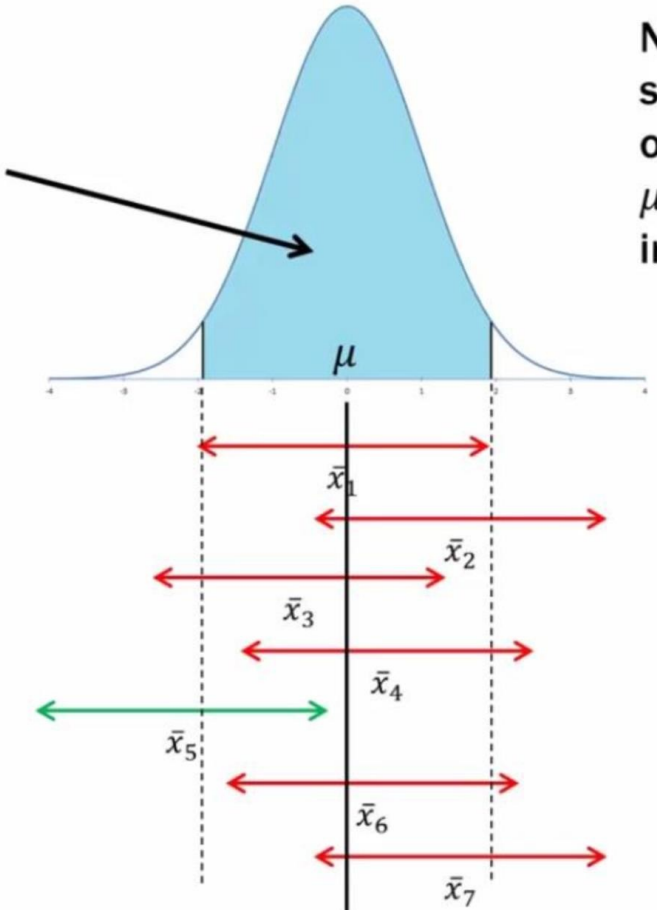


94 of the 100 intervals contain the population mean.

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]

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

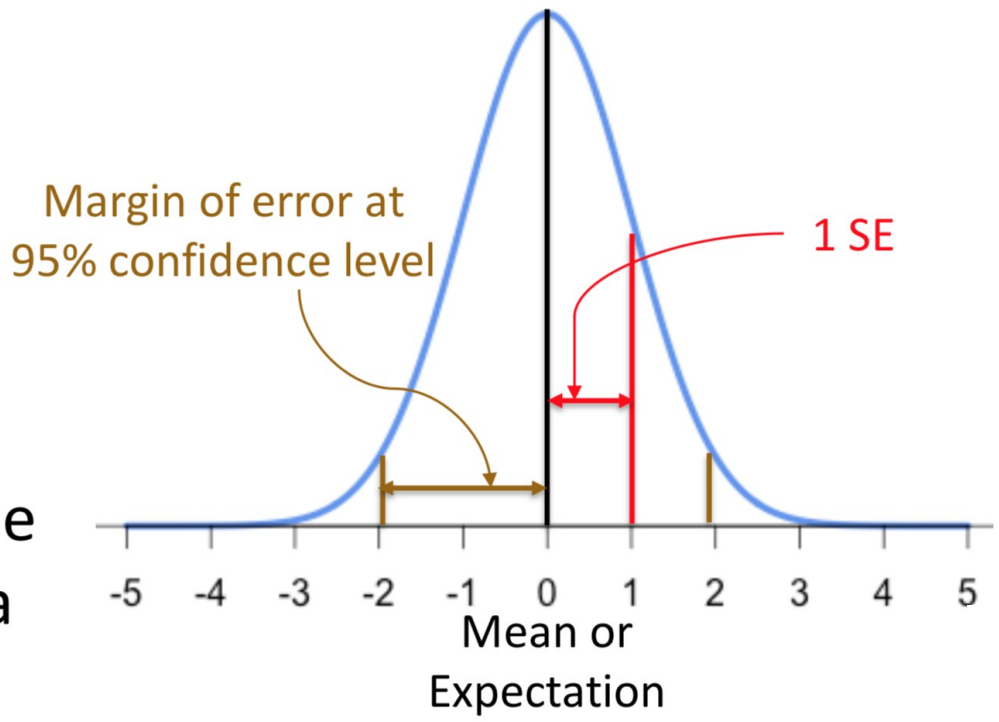
Asked of 476 Republican registered voters

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

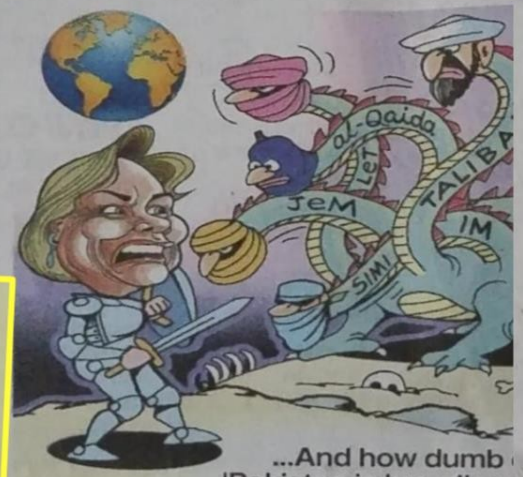
Public opinion polls are rather like children in a garden, digging things up all the time to see how they're growing, a British writer remarked perceptively, and nowhere is it truer than across the pond, where weekly surveys are the norm in the months leading up to the presidential elections.

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.



...And how dumb
'Pakistan is breeding

"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 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.

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

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

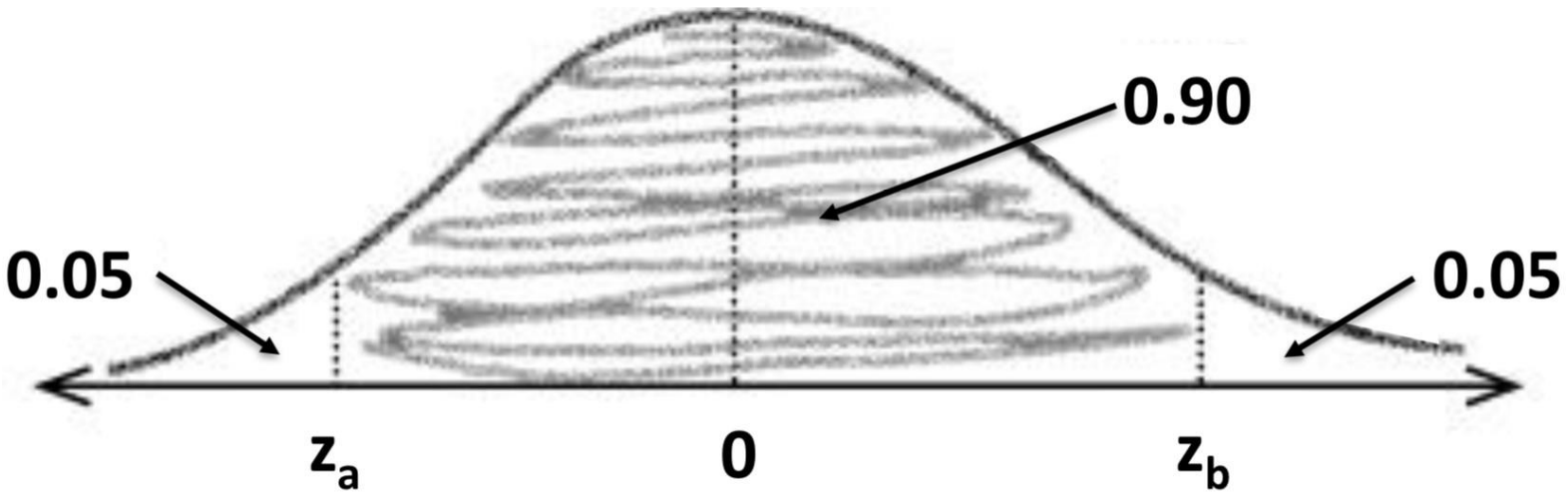
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

$\therefore$ 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

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 $\pm$ 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

Population Parameter	Population Distribution	Conditions	Confidence Interval
μ	Normal	You know σ^2 n is large or small $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
μ	Non-normal	You know σ^2 n is large (> 30) $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
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	$(\bar{X} - z\frac{s}{\sqrt{n}}, \bar{X} + z\frac{s}{\sqrt{n}})$
p	Binomial	n is large p_s is the sample proportion q_s is $1 - p_s$	$(p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}})$

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

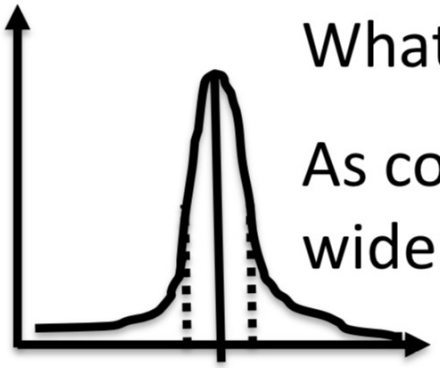
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

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*.

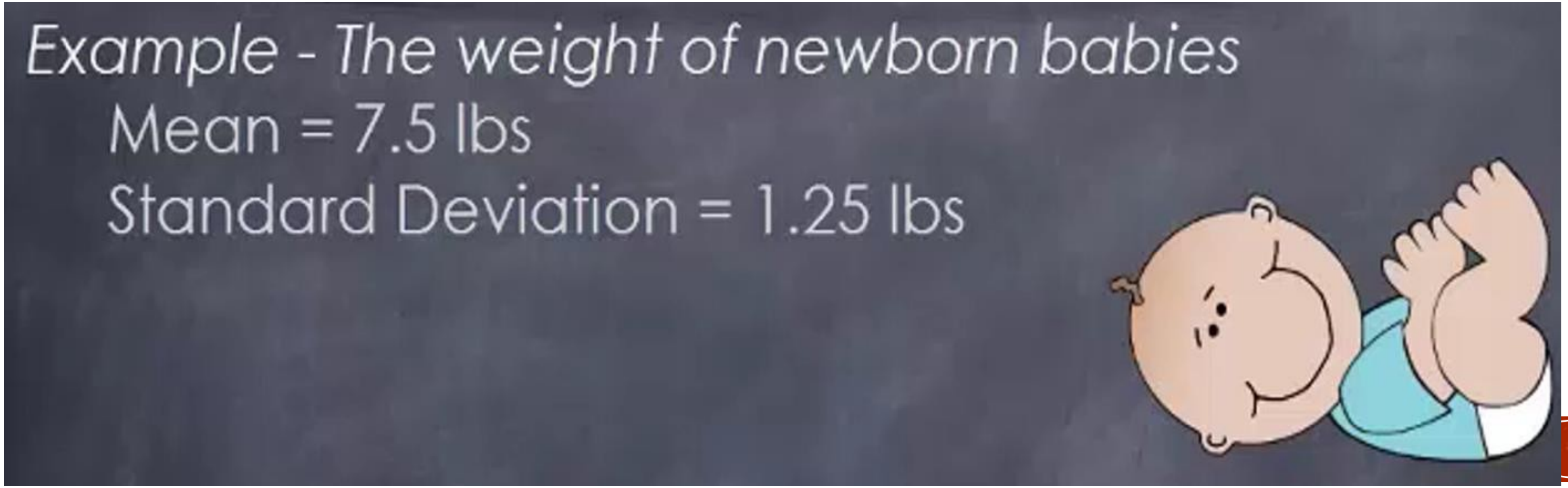
What happens to the confidence interval as sample size changes?

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.

Probability that a newborn weighs <6 lbs

z-score = (6 - 7.5)/(1.25) = -1.20

- 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

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,

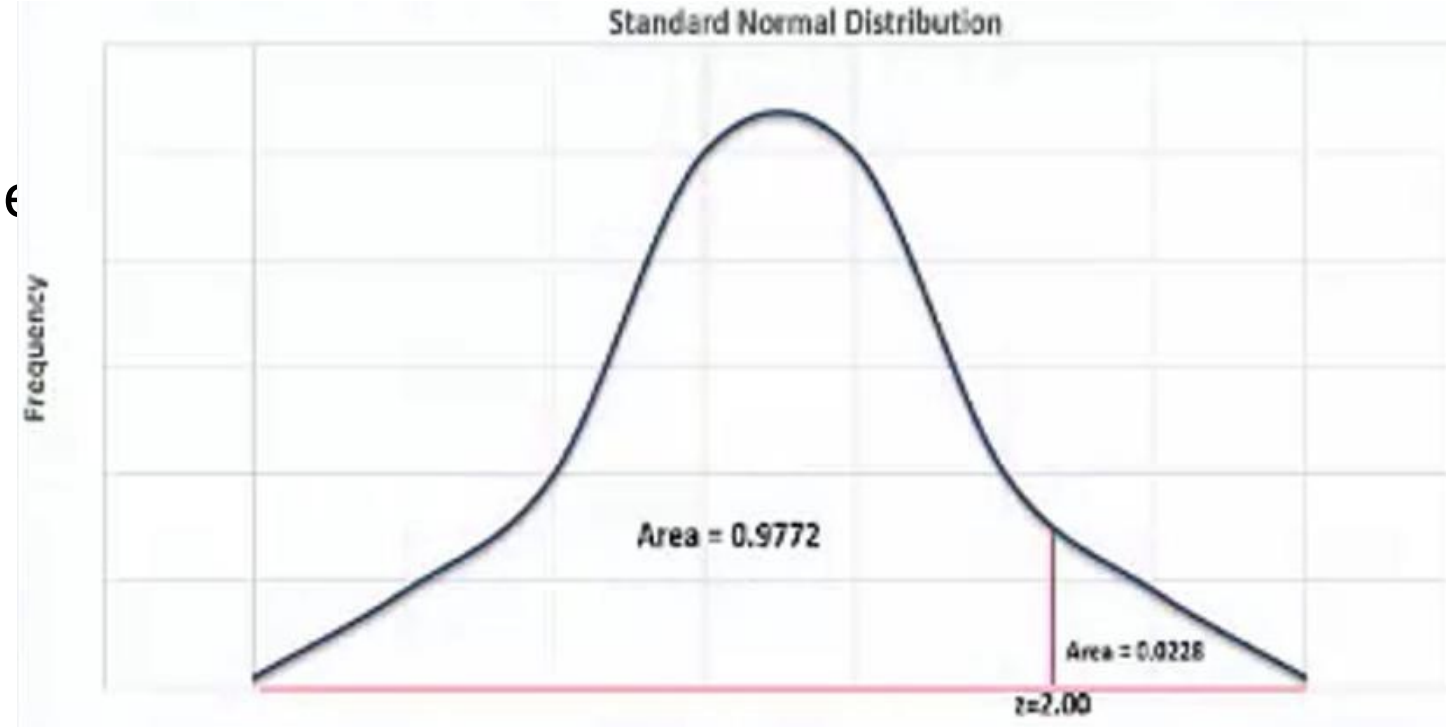
Application of Z-Score

- 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

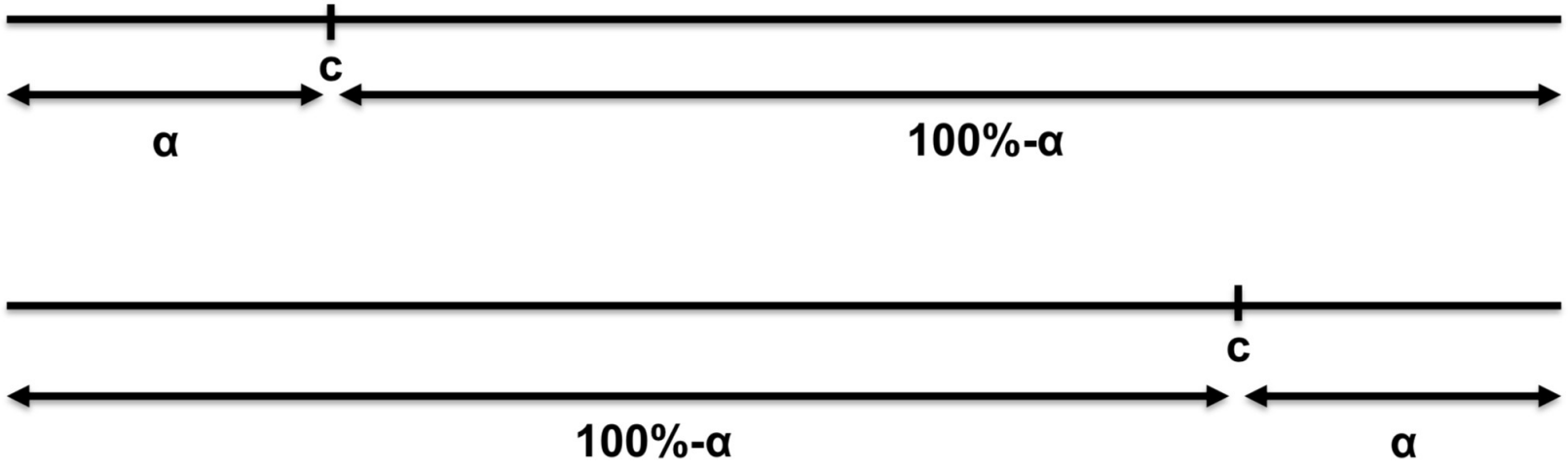
Critical Region Up Close

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.

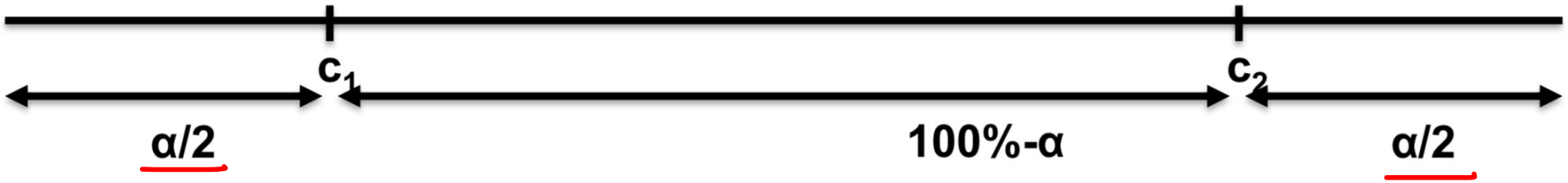


Critical Region Up Close

Two-tailed tests

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

If H_1 includes a ≠ 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
- $\boxed{\chi^2}$ (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

- 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 <u>0.05 level of significance</u>.

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 <u>rejected</u> 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 <u>lower critical z</u> (-1.96) at <u>level of significance</u> (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

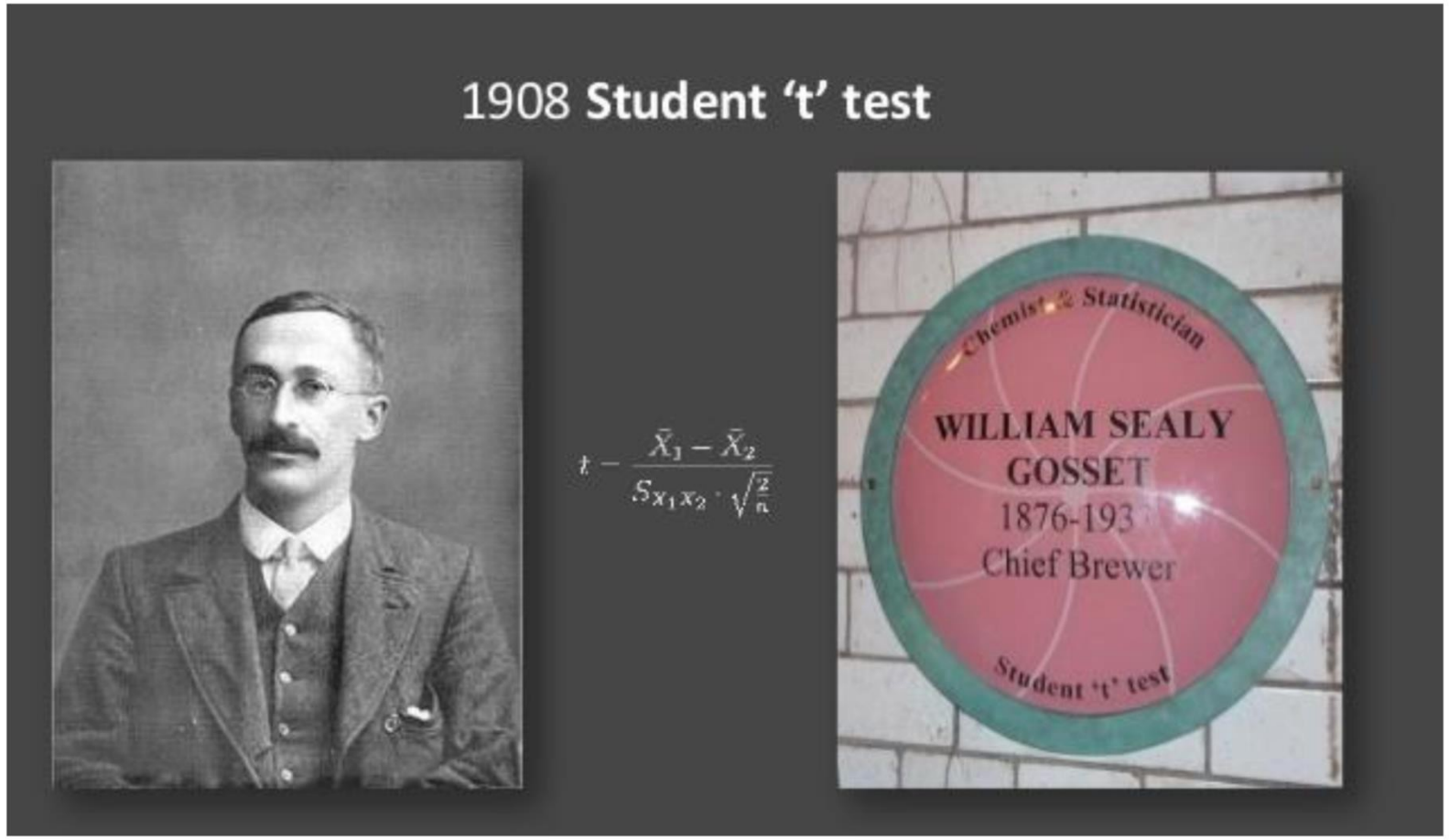
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{15.5 - 13}{13 / \sqrt{169}} = 2.5 > 1.96$$

Application of Z-Test

- **Decision:** Reject H_0

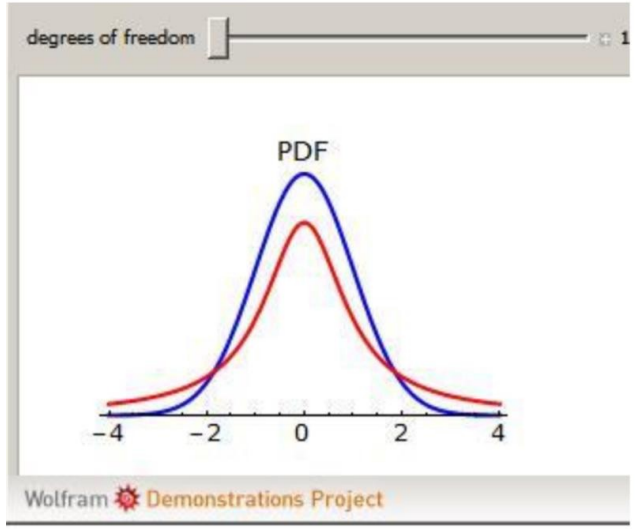
- **Interpretation:** The national average commuting distance does not describe the mean commuting distance for all workers in the Chicago area

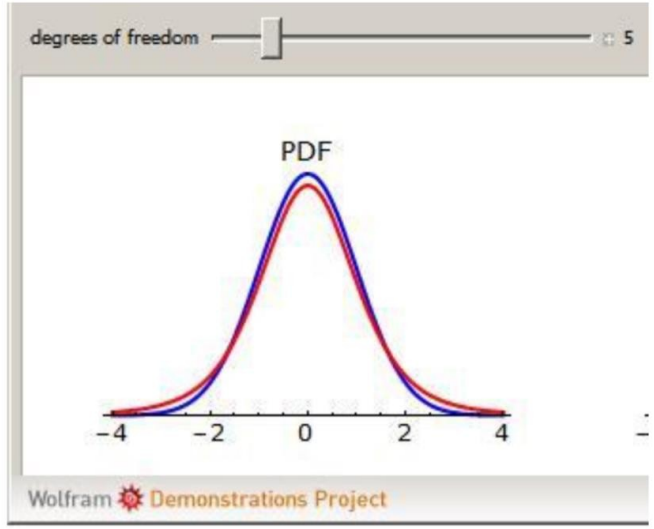
t-Distribution

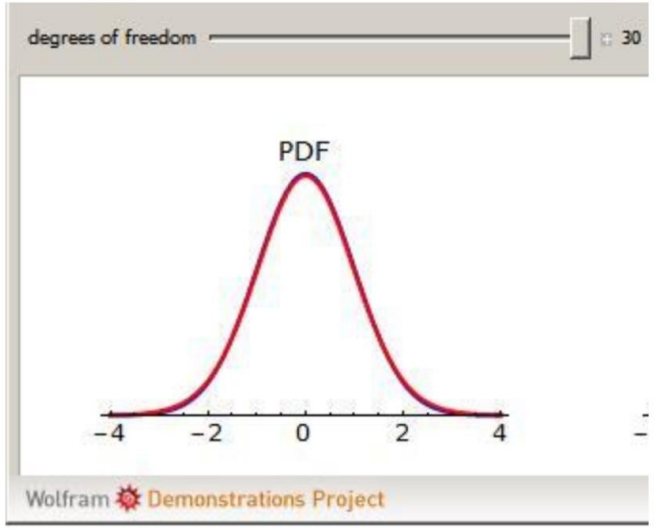


t-Distribution

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 <u>population variance is **not known**</u>, 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*.

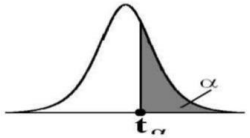
Confidence Interval to Estimate μ

- Population standard deviation UNKNOWN and the population normally distributed.
- $\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
  - 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; $t_{\nu,\alpha}$

$$P(T > t_{\nu,\alpha}) = \alpha$$



	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

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

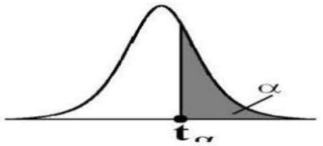
n = 10

$\nu = 10 - 1 = 9$

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, $\nu = 10 - 1 = 9$

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

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

$99.47 \leq \mu \leq 103.01$

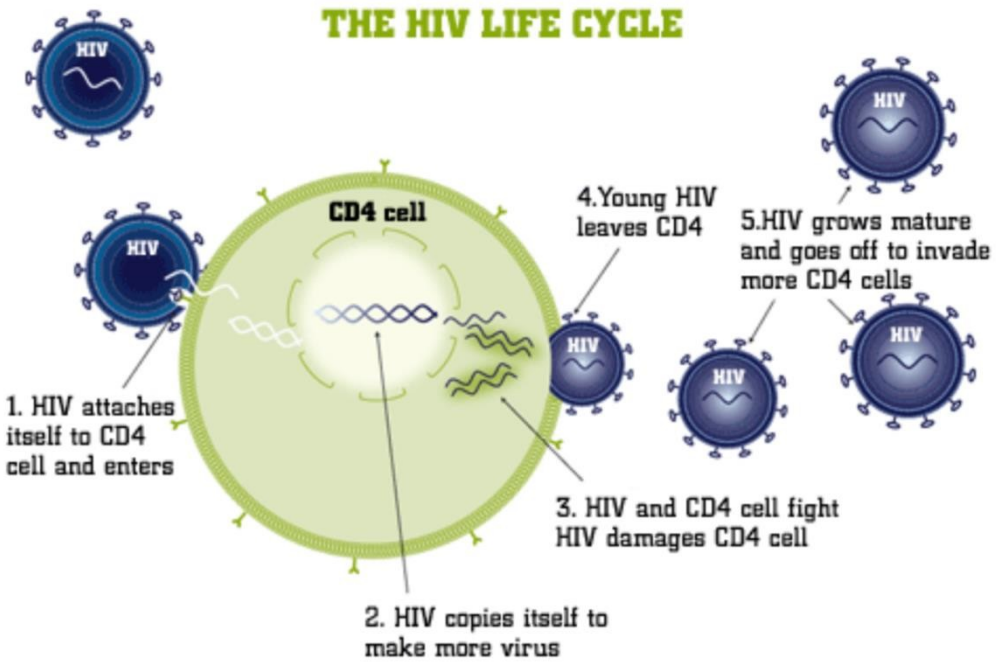
The batch mean is 101.24 mg with an error of +/-1.77 mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.47 mg and 103.01 mg.

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

t-Distribution - Example

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

$$CI(0.05):\ \bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$321.4 - t_{25-1,\frac{0.05}{2}} \frac{73.8}{\sqrt{25}} \leq \mu \leq 321.4 + t_{25-1,\frac{0.05}{2}} \frac{73.8}{\sqrt{25}}$$

$$t_{24,0.025} = 2.064$$

$$\therefore\ CI(0.05): [290.85, 351.95]$$

t-Distribution - Example

What does $CI(0.05)$: $[290.85, 351.95]$ mean in the business context given that US Government classifies AIDS under three official categories of CD4 counts:

- Asymptomatic: $\geq$ 500 cells/μl
- AIDS related complex (ARC): 200-499 cells/μl
- AIDS (Stage 3 infection): < 200 cells/μl

Application of t-test

- **Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.**

- Step 1: Write your null hypothesis statement (How to state a null hypothesis).
- The accepted hypothesis is that there is no difference in sales, so:
  H_0: μ = $100.

Application of t-test

- Step 2: Write your alternate hypothesis. This is the one you're testing. You think that there *is* a difference (that the mean sales increased), so:
  H_1: $\mu > \$100$.

- Step 3: Identify the following pieces of information you'll need to calculate the test statistic. The question should give you these items:

- **Given:**
- **The sample mean**($\bar{x}$). This is given in the question as $130.
- **The population mean**(μ). Given as $100 (from past data).
- **The sample standard deviation**(s) = $15.
- **Number of observations**(n) = 25.

Application of t-test

- Step 4: Insert the items from above into the t score formula.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- t = (130 – 100) / ((15 / √(25))
  t = (30 / 3) = 10
  This is your **calculated t-value**.

Application of t-test

- Step 5: Find the t-table value. You need two values to find this:
- The alpha level: given as 5% in the question.
- The degrees of freedom, which is the number of items in the sample (n) minus 1: 25 – 1 = 24.
- Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1.711.This is your one-tailed critical t-value.
- What this critical value means is that we would expect most values to fall under 1.711. If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.
- Step 6: Compare Step 4 to Step 5. The value from Step 4 **does not** fall into the range calculated in Step 5, so we can reject the null hypothesis. The value of 10 falls into the rejection region.
- In other words, it's highly likely that the mean sale is greater. The sales training was probably a success.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

EXAMPLE FOR T – TEST

- A car company claims that their Super Spiffy Sedan averages 31 mpg. You randomly select 8 Super Spiffies from local car dealerships and test their gas mileage under similar conditions.

- You get the following MPG scores:

- MPG: 30 28 32 26 33 25 28 30
- Does the actual gas mileage for these cars deviate significantly from 31 (alpha = .05)?

Testing Hypotheses about a Variance

Sample estimate of population variance is given by

$$\boxed{s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Multiplying the variance estimate by *n-1* gives the sum of squares. Dividing by population variance gives a random variable distributed as chi-squared with *n-1* degrees of freedom.

$$\therefore \boxed{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}$$

$$\textit{Recall } \chi^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$$

Testing Hypotheses about a Variance

A manufacturing company produces bearings of 2.65 cm in diameter. A major customer requires that the variance in diameter be no more than 0.001 cm^2. The manufacturer tests 20 bearings using a precise instrument and gets the below values. Assuming the diameters are normally distributed, can the population of these bearings be rejected due to high variance at 1% significance level?

Data: 2.69, 2.66, 2.64, 2.59, 2.62, 2.63, 2.69, 2.66, 2.63, 2.65, 2.57, 2.63, 2.70, 2.71, 2.64, 2.65, 2.59, 2.66, 2.62, 2.57

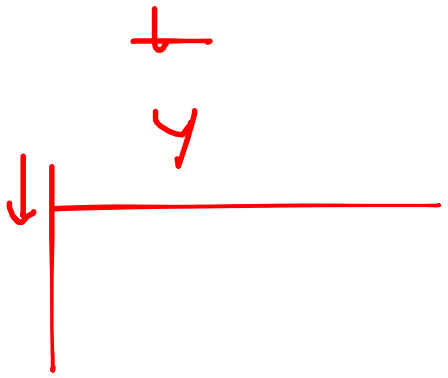
Testing Hypotheses about a Variance

What are null and alternate hypotheses?

H_0: $\sigma^2 \leq 0.001$; H_1: $\sigma^2 > 0.001$

How many degrees of freedom?

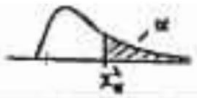
Since n=20, df=19.



Testing Hypotheses about a Variance

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



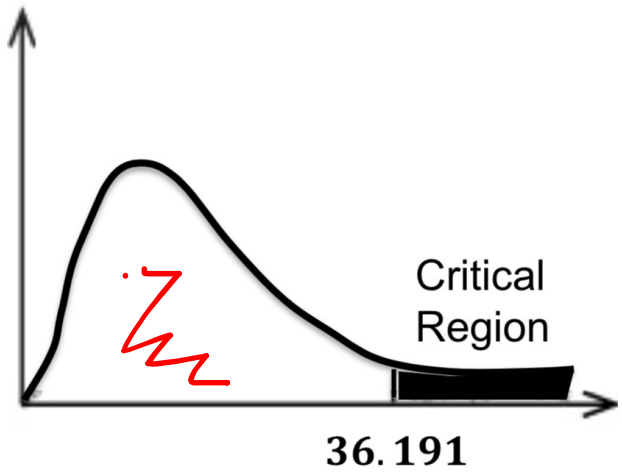
ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315

$$\chi^2_{0.01,19} = 36.191$$

Testing Hypotheses about a Variance

What is the observed χ^2 value?

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19 * 0.001621}{0.001} = \boxed{30.8}$$



Is it in critical region? $\chi^2_{0.01,19} = 36.191$

No.

Will you reject or fail to reject the null hypothesis?

Fail to reject.

What is the business decision?

The population variance is within specification limits required by the customer and hence the bearings can be shipped.

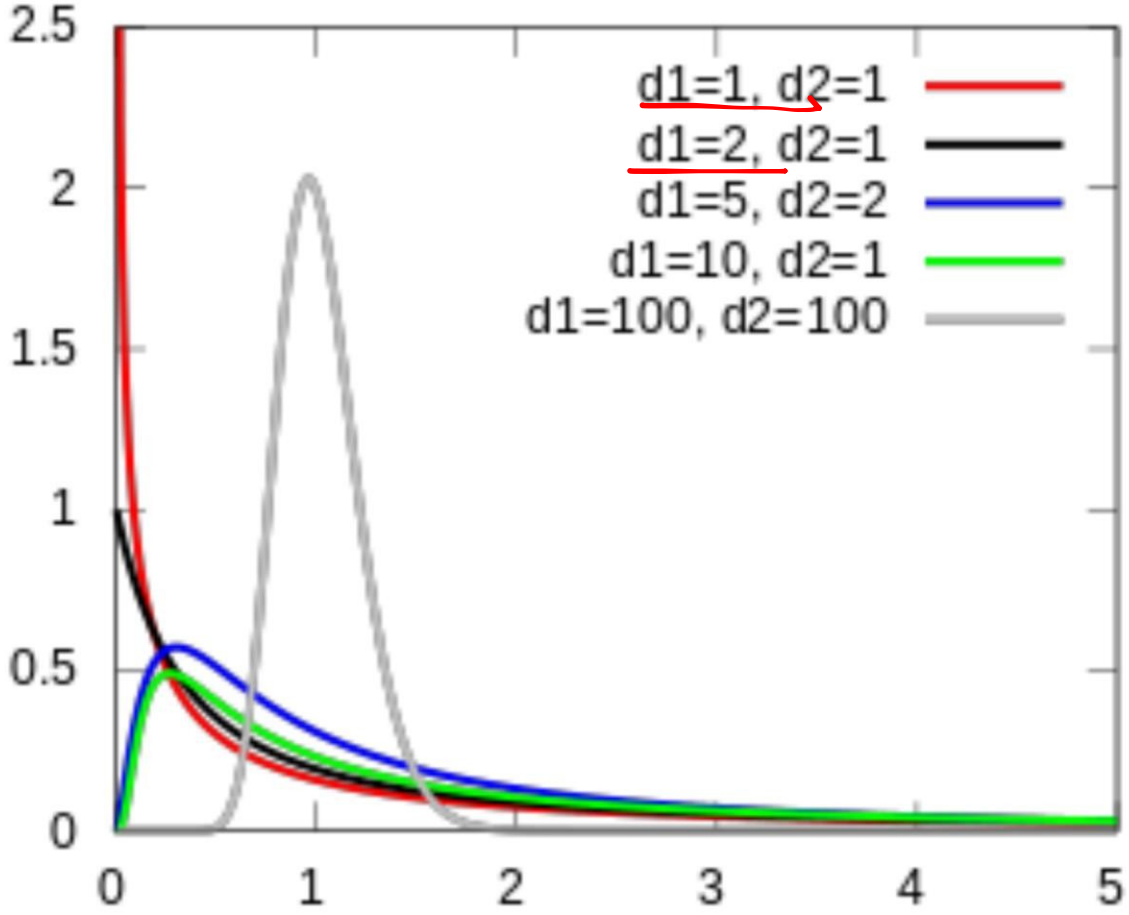
F distribution

- χ^2 was useful in testing hypotheses about a single population variance.
- Sometimes we want to test hypotheses about difference in variances of two populations:
  - Is the variance of 2 stocks the same?
  - Do parts manufactured in 2 shifts or on 2 different machines or in 2 batches have the same variance or not?
  - Is the powder mix for tablet granulations homogeneous?
  - Is there variability in assayed drug blood levels in a bioavailability study?
  - Is there variability in the clinical response to drug therapy of two samples?

F distribution

- Ratio of 2 variance estimates: $\boxed{F = \frac{s_1^2}{s_2^2} = \frac{est.\sigma_1^2}{est.\sigma_2^2}}$
- Ideally, this ratio should be about 1 if 2 samples come from the same population or from 2 populations with same variance, but sampling errors cause variation.
- *Recall* $\chi^2 = \frac{\cancel{(n-1)}s^2}{\cancel{\sigma^2}}$. So, F is also a ratio of 2 chi-squares, each divided by its degrees of freedom, i.e.,

F distribution



Hypothesis test for 2 sample variances

A machine produces metal sheets with 22mm thickness. There is variability in thickness due to machines, operators, manufacturing environment, raw material, etc. The company wants to know the consistency of two machines and randomly samples 10 sheets from machine 1 and 12 sheets from machine 2. Thickness measurements are taken. Assume sheet thickness is normally distributed in the population.

The company wants to know if the variance from each sample comes from the same population variance (population variances are equal) or from different population variances (population variances are unequal).

How do you test this?

Hypothesis test for 2 sample variances

Data

Machine 1		Machine 2	
22.3	21.9	22.0	21.7
21.8	22.4	22.1	21.9
22.3	22.5	21.8	22.0
21.6	22.2	21.9	22.1
21.8	21.6	22.2	21.9
		22.0	22.1
$s_1^2 = 0.11378$	$n = 10$	$s_2^2 = 0.02023$	$n = 12$

Ratio of sample variances, $F = \frac{s_1^2}{s_2^2} = \frac{0.11378}{0.02023} = 5.62$

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$

Is it a one-tailed test or a two-tailed test?

Two-tailed.

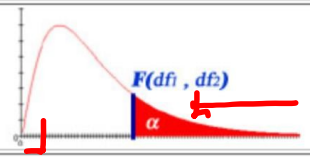
What are numerator and denominator degrees of freedom?

$\nu_1 = 10 - 1 = 9; \nu_2 = 12 - 1 = 11$

Hypothesis test for 2 sample variances

Reading an F-table.

F Table for α = 0.025



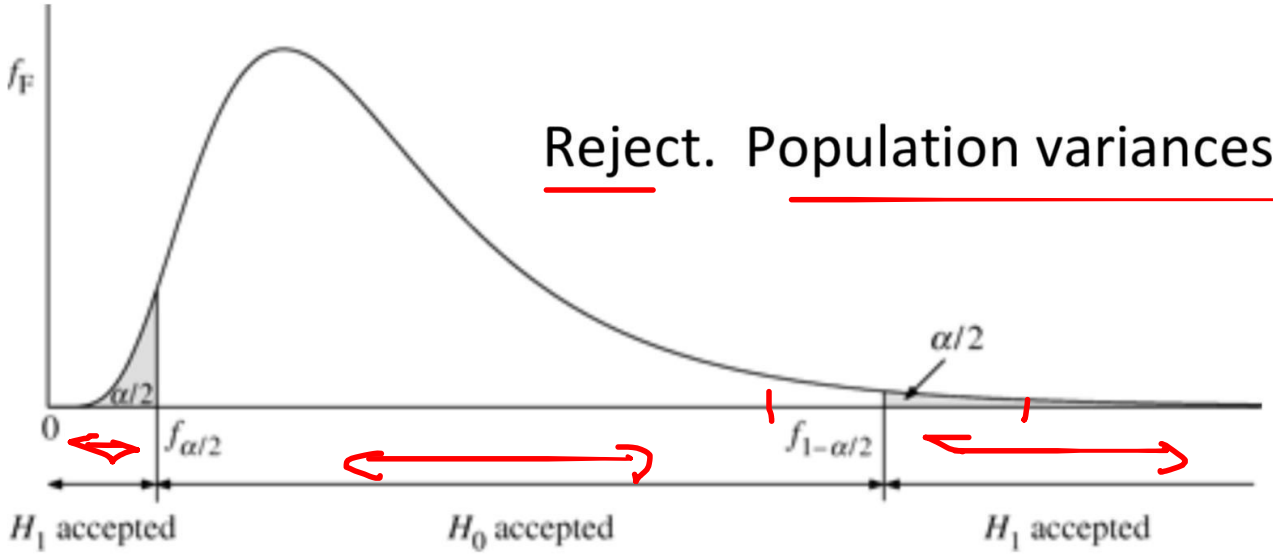
/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
df_2=1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.800	1014.020
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.490
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189	14.3366	14.2527	14.1674	14.1241	14.081	14.037	13.992	13.947
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.7512	8.6565	8.5599	8.5109	8.461	8.411	8.360	8.309
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.5245	6.4277	6.3286	6.2780	6.227	6.175	6.123	6.069
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.3662	5.2687	5.1684	5.1172	5.065	5.012	4.959	4.904
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.4150	4.362	4.309	4.254	4.199
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9472	3.894	3.840	3.784	3.728
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639	3.8682	3.7694	3.6669	3.6142	3.560	3.505	3.449	3.392
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6209	3.5217	3.4185	3.3654	3.311	3.255	3.198	3.140
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879	3.5257	3.4296	3.3299	3.2261	3.1725	3.118	3.061	3.004	2.944
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.2773	3.1772	3.0728	3.0187	2.963	2.906	2.848	2.787

$$F_{0.025,9,11} = 3.5879;\ F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787$$

Hypothesis test for 2 sample variances

$$F_{0.025,9,11} = 3.5879;\ F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787;\ F_{observed} = 5.62$$

Will you reject the null hypothesis or not?



Reject. Population variances are not equal.

Hypothesis test for 2 sample variances

What are the business implications?

Variance in machine 1 is higher than in machine 2. Machine 1 needs to be inspected for any issues.

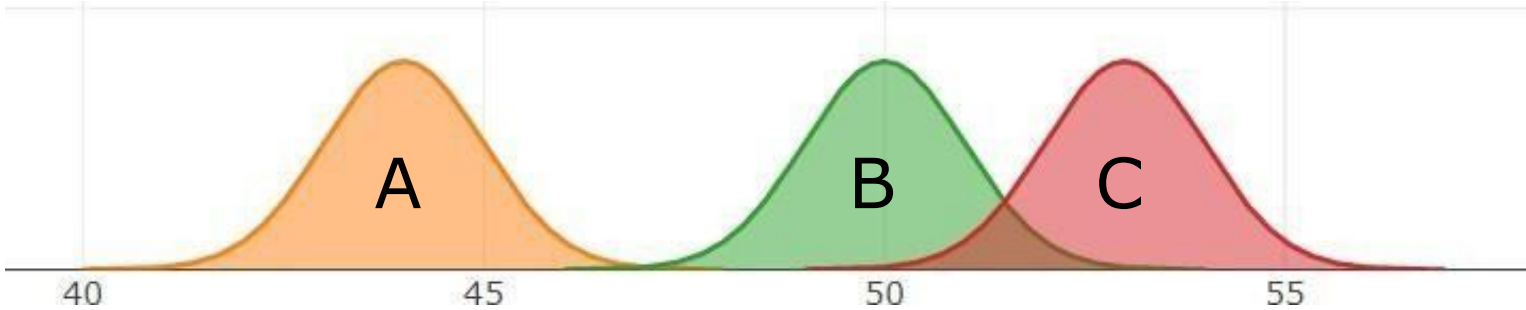
Applications of F Distribution

- Test for equality of variances.
- Test for differences of means in ANOVA.
- Test for regression models (slopes relating one continuous variable to another, e.g., Entrance exam scores and GPA)

ANOVA
ANALYSIS OF VARIANCE

ANOVA

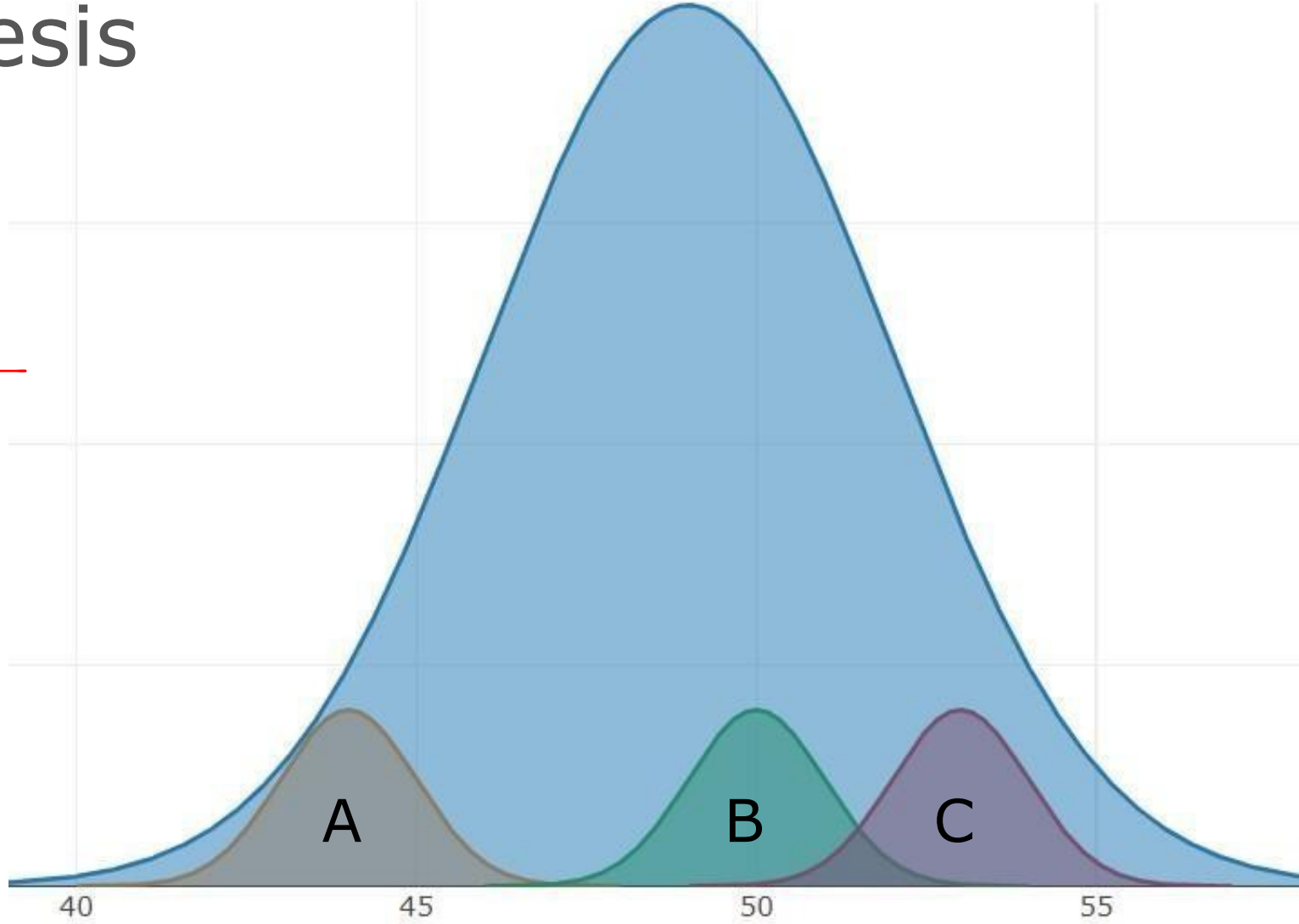
- In the previous section we tested samples to see if they are the representative of the population.
- What if we had three (or more) samples?
- Could we find if they are from same population?



ANOVA

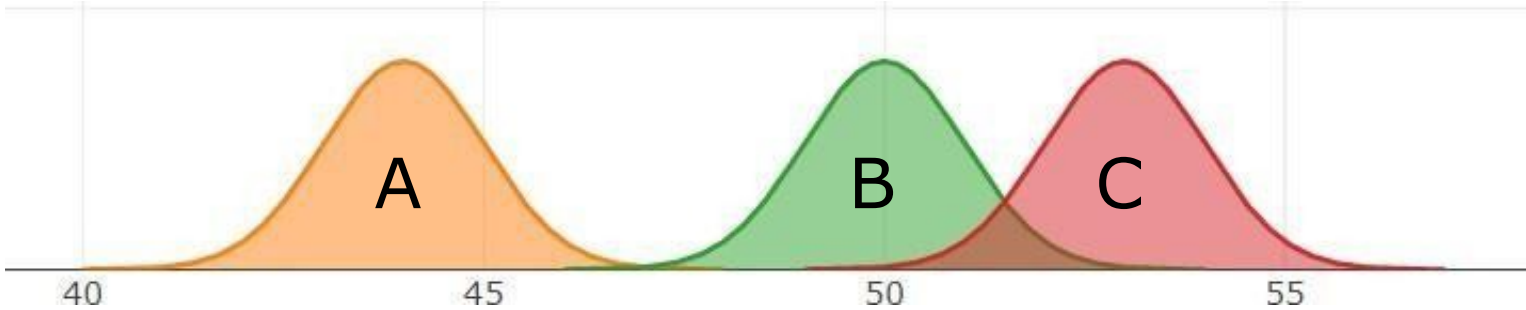
- Our null hypothesis would look like:

$$H_0: \mu_A = \mu_B = \mu_C$$



ANOVA

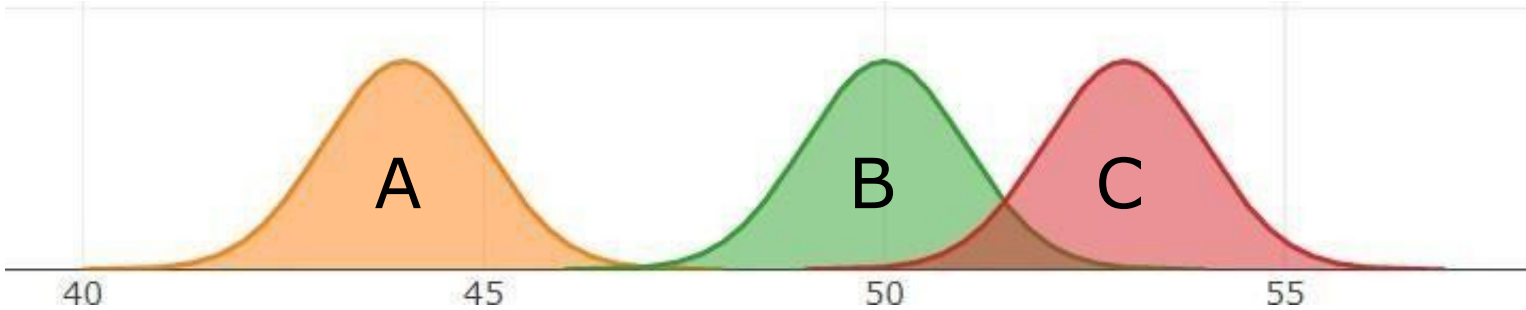
- This is where ANOVA comes in!
- We compute an F value, and compare it to a critical value determined by our degrees of freedom (the number of groups, and the number of items in each group)



ANOVA

Let’s work with somedata:

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43



ANOVA

First calculate the sample mean

Next calculate the overall mean

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

$\frac{44 + 50 + 53}{3}$

ANOVA

$F =$

ANOVA considers two types of variance:

- Between Groups
  - how far group meansstray from the total mean
- Within Groups
  - how far individual values stray from their respective group mean

ANOVA

The F value we're trying to calculate is simply the ratio between these two variances!

$$F = \frac{VarianceBetweenGroups}{VarianceWithinGroups}$$

B

W

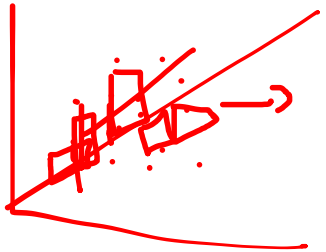
ANOVA

Recall the equation for variance:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{SS}{df}$$

Here $\Sigma(x - \bar{x})^2$ is the "sum of squares" SS and $n - 1$ is the "degrees of freedom" df

ANOVA



$(x-\mu) = \sigma^2$

So the formula for the F value becomes:

$$F = \frac{VarianceBetweenGroups}{VarianceWithinGroups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

SSG=SumofSquaresGroups

SSE=SumofSquaresError

df_{groups}=degreesof reedom(groups)

df_{error}=degreesoffreedom(er or)

ANOVA

$SSG = 420$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Sum of Squares Groups

$(\mu_A - \mu_{TOT})^2 = (44-49)^2 = 25$

$(\mu_B - \mu_{TOT})^2 = (50-49)^2 = 1$

$(\mu_C - \mu_{TOT})^2 = (53-49)^2 = 16$

42

Multiply by the number of items in each group:

$42 \times 10 = 420$

ANOVA

$SSG = 420$

$df_{groups} = 2$

Degrees of Freedom Groups

$$df_{groups} = n_{groups} - 1$$
$$= 3 - 1$$
$$= 2$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

Sum of Squares ~~Error~~ Win

$SSG = 420$

$df_{groups} = 2$

$SSE = 3300$

$(x_A-\mu_A)^2$	$(x_A-\mu_A)^2$	$(x_B-\mu_B)^2$	$(x_B-\mu_B)^2$	$(x_C-\mu_C)^2$	$(x_C-\mu_C)^2$
49	64	144	16	9	1
256	121	529	36	100	0
64	25	361	361	25	100
1	25	196	1	1	144
16	441	49	49	16	100
1062			**1742**		**496**
				TOTAL	**3300**

$(37-44)^2$
$=(-7)^2$
$=49$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43
$\mu_{A,B,C}$ **44**	**50**	**53**
μ_{TOT} **49**		

ANOVA

$SSG = 420$
$df_{groups} = 2$
$SSE = 3300$
$df_{error} = 27$

Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$
$$= (10 - 1) * 3$$
$$= 27$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$SSG = 420$
$df_{groups} = 2$
$SSE = 3300$
$df_{error} = 27$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Plug these into our formula:

$$F = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}} = \frac{\frac{420}{2}}{\frac{3300}{27}} = \frac{210}{122.2} = \mathbf{1.718}$$

ANOVA WITH EXCEL DATA ANALYSIS

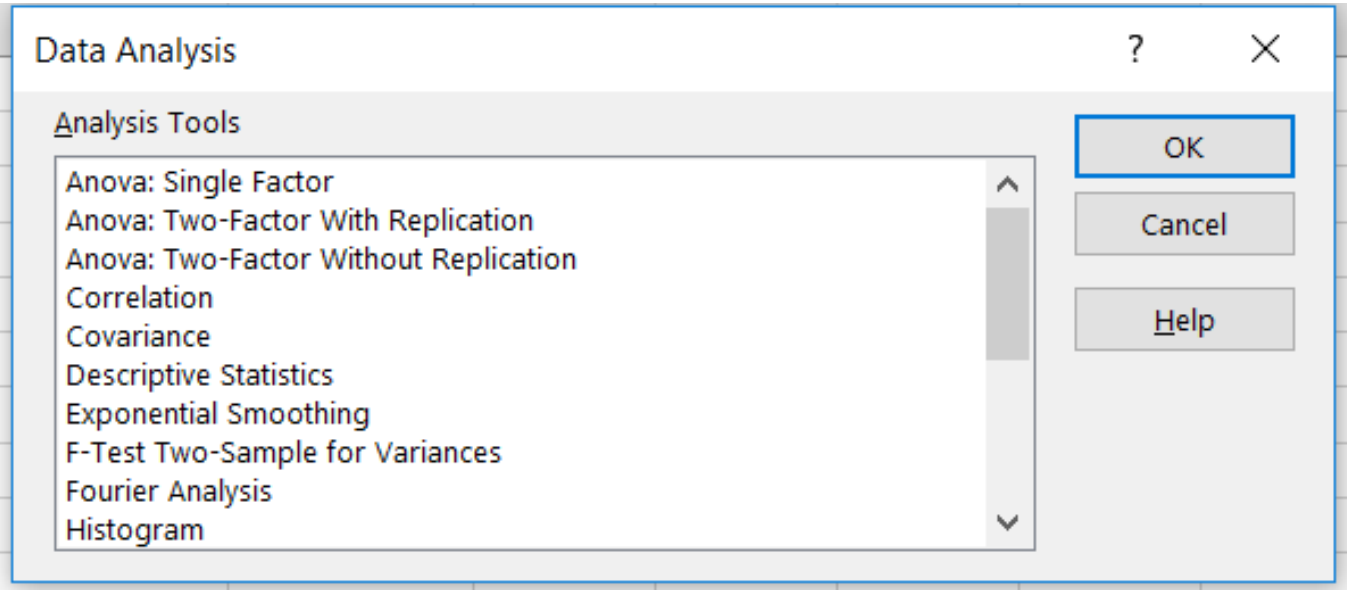
Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
GroupA	10	440	44	118
GroupB	10	500	50	193.5555556
GroupC	10	530	53	55.11111111

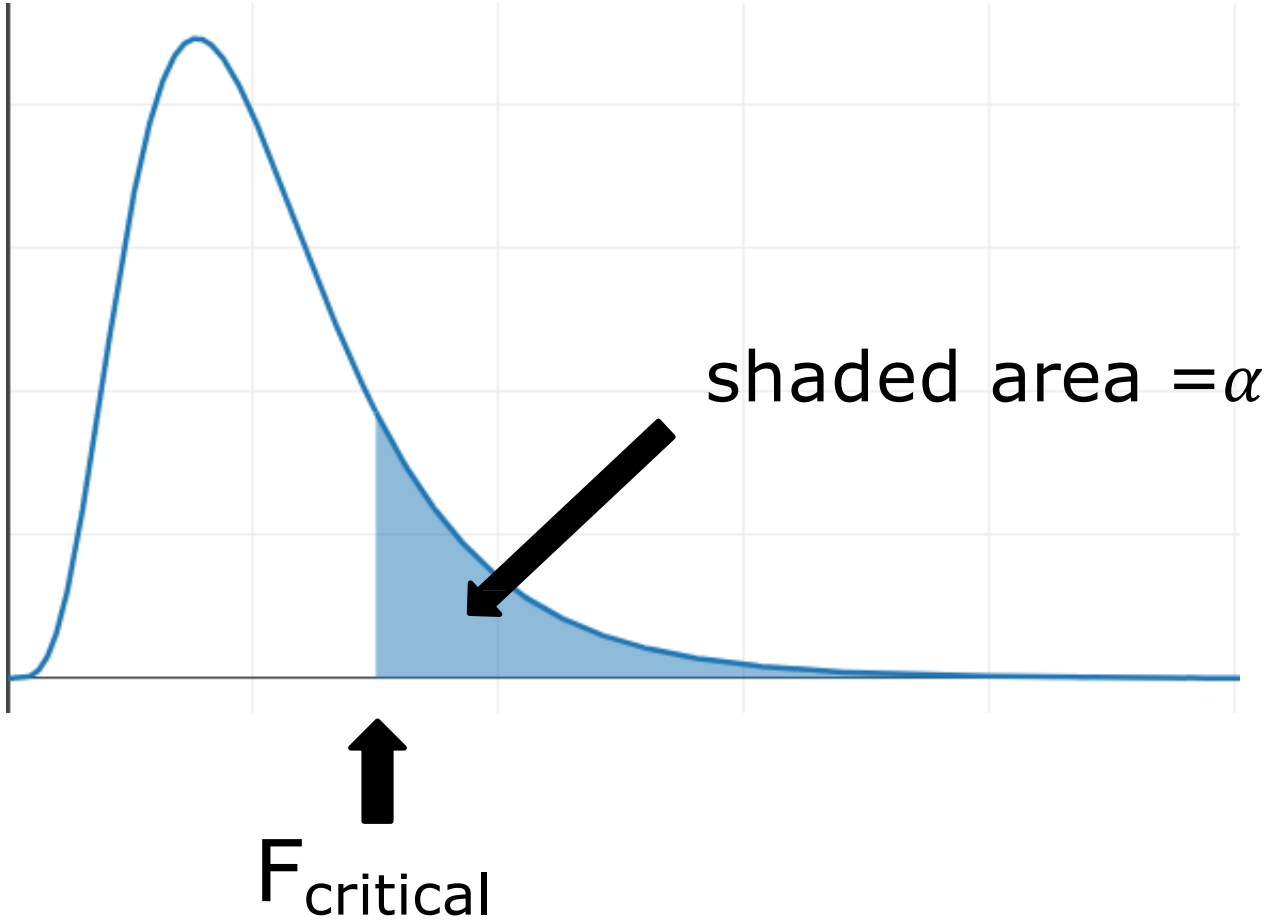
ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	420	2	210	1.71[illegible]1818	0.198430533	3.354130829
Within Groups	3300	27	122.2222			
Total	3720	29				



F DISTRIBUTION

F-DISTRIBUTION



F-DISTRIBUTION

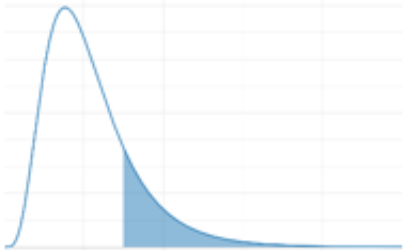
Look up our critical value from an F-table

use a table set for 95% confidence

find numerator df

find denominator df

critical value =3.35



		F-Table Upper Tail Area of 0.05				
		Numerator df				
		1	2	3	4	5
denominator df	25	4.24	3.39	2.99	2.76	2.60
	26	4.23	3.37	2.98	2.74	2.59
	27	4.21	3.35	2.96	2.73	2.57
	28	4.20	3.34	2.95	2.71	2.56
	29	4.18	3.33	2.93	2.70	2.55
	30	4.17	3.32	2.92	2.69	2.53

F-SCORES IN MSEXCEL

- In Microsoft Excel, the following function returns an F-score:

α	df1	df2	Formula	Output Value
0.05	2	27	=FINV(A2,B2,C2)	3.3541308285292

F-SCORES IN PYTHON

```
>>> from scipy import stats
>>> stats.f.ppf(1-.05,dfn=2,dfd=27)
3.3541308285291986
```

ANOVA

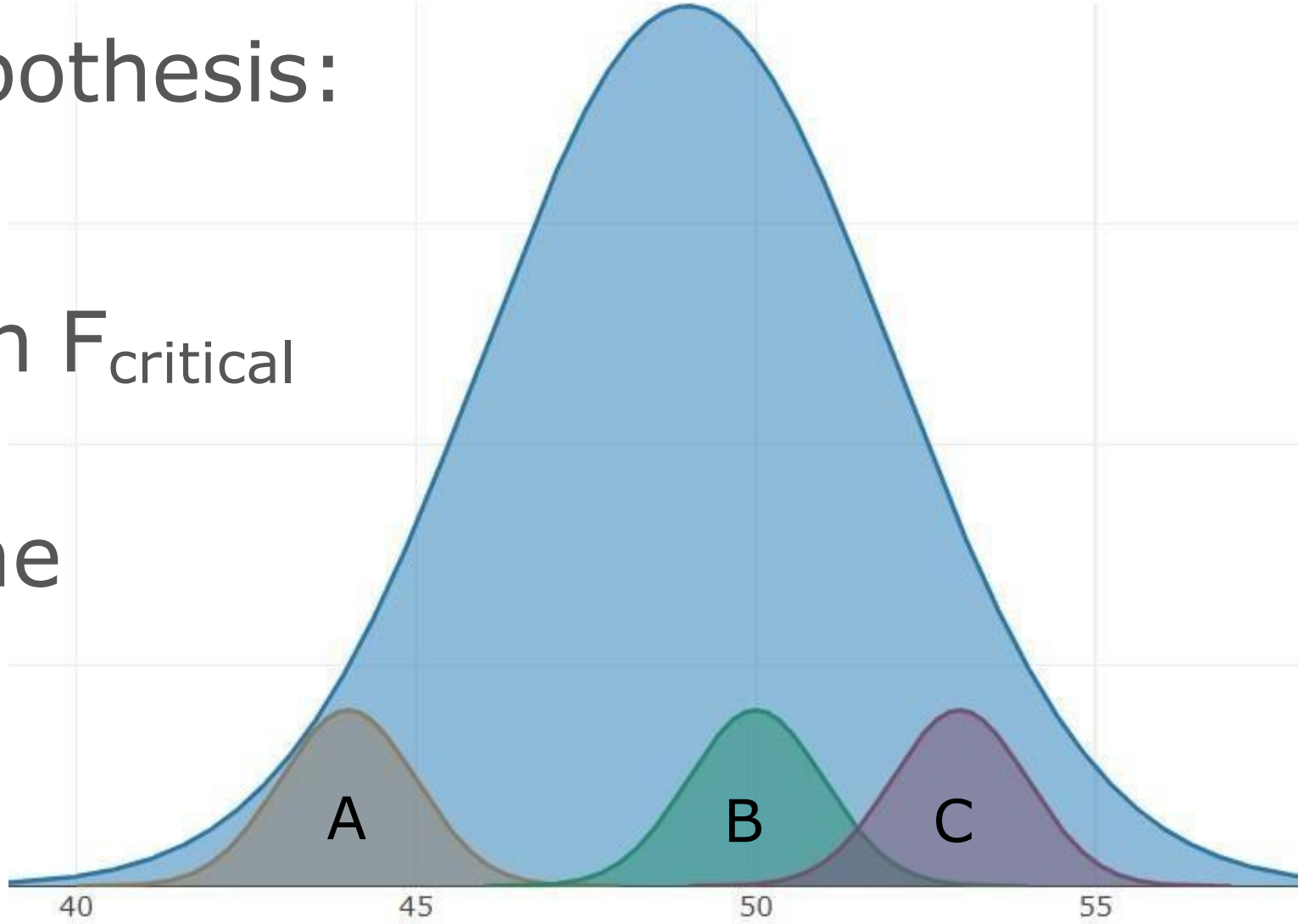
Recall our null hypothesis:

H_0: $\mu_A = \mu_B = \mu_C$

Since F is less than $F_{critical}$

1.718<3.354

wefail to reject the null hypothesis!



STOCK MARKET EXAMPLE

A stock analyst randomly selected 8 stocks from each of 3 industries, viz., Financial, Energy and Utilities. She compiled the 5-year rate of return for each stock.

The analyst wants to know if, at 0.05 Significance Level, there is a difference in the rate of return for any of the industries.

STOCK MARKET EXAMPLE

	5-Year Rates of Return		
	Financial	Energy	Utilities
	10.76	12.72	11.88
	15.05	13.91	5.86
	17.01	6.43	13.46
	5.07	11.19	9.9
	19.5	18.79	3.95
	8.16	20.73	3.44
	10.38	9.6	7.11
	6.75	17.4	15.7
xbar	**11.585**	**13.846**	**8.913**
s	**5.124**	**4.867**	**4.530**

What is the null hypothesis?

$\mu_1 = \mu_2 = \mu_3.\ \alpha = 0.05$
All 3 industries have the same average rate of return.

What is the alternate hypothesis?

At least one of the industries has a different rate of return than the others.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Financial	8	92.68	11.585	26.2528
Energy	8	110.77	13.8463	23.6879
Utilities	8	71.3	8.9125	20.5247

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	97.593	2	48.7965	2.07747	0.1502	3.4668
Within Groups	493.26	21	23.4885			
Total	590.85	23				

Thought Process on When to Use a Particular Test

- What do you want to do?
  - Description – Summary statistics, Various plots, Correlations
  - Prediction – Linear regression, Logistic regression
  - Intervention (differences between groups) – *t*-test, Chi-square, ANOVA
- Is the dependent variable Categorical or Numerical?
  - Nominal – Chi-square, Logistic regression
  - Ordinal – Chi-Square
  - ~~Dichotomous – Logistic regression~~
  - Numerical – *t*-test, ANOVA, Correlation, Multiple regression