Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Confidence Intervals, Hypothesis Testing, t-Distribution

Dr. Sridhar Pappu

Executive VP – Academics, INSOFE

December 23, 2017


FLASHBACK
I REMEMBER IT LIKE IT WAS YESTERDAY...
CMCOTT. 04/13/12 #138

Activity – R

According to the US Bureau of the Census, about 75% of the commuters in the United States drive to work alone. Suppose 150 US commuters are randomly sampled.

- What is the probability that fewer than 105 commuters drive to work alone?
- What is the probability that between 110 and 120 (inclusive) commuters drive to work alone?
- What is the probability that more than 95 commuters drive to work alone?

Answers: 0.0657, 0.6485, 0.9993

Activity – R

According to National Center for Health Statistics of the US, the distribution of serum cholesterol levels for 20-74 year old males has a mean of 211mg/dl with a standard deviation of 46mg/dl.

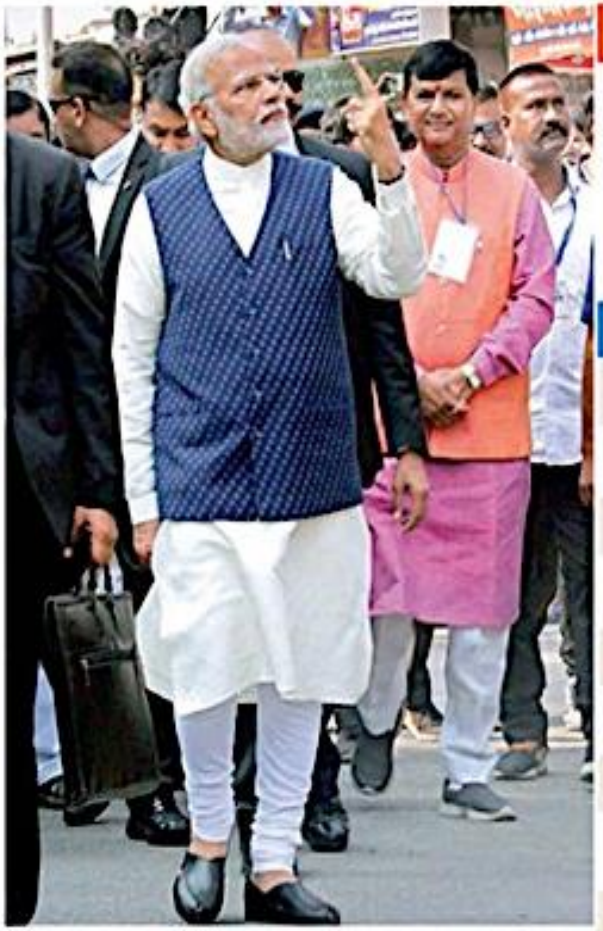
- What is the probability that the serum cholesterol level of a male is >230mg/dl?
- What is the probability that the average serum cholesterol level of a random sample of 25 males will be >230mg/dl?

Answer: 34.0%, 1.9%

Present Day

INFERENTIAL STATISTICS

CONFIDENCE LEVELS AND CONFIDENCE INTERVALS



Prime Minister Narendra Modi shows his finger marked with indelible ink after casting his vote in the second phase of the Assembly elections in Ahmedabad on Thursday. The Congress immediately called it a road show and violation of the EC code. ■ Report on Page 8 — PTI

WHAT'S IN STORE

TOTAL SEATS 182

WINNING SEATS 92

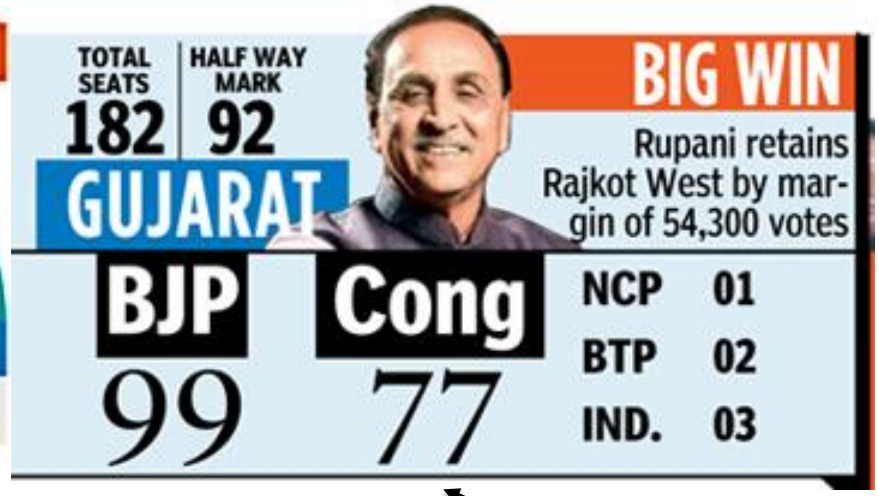
Agency	BJP	Congress
Chanakya	135	47
VDPA	142	37
CSDS-ABP	117	64
Axis	99-113	66
VMR	108-118	61-71
Republic TV	108	74
C-Voter-TV9	108	74
CNX-Newsx	110-120	65-75
NDTV	112	70

In 2012 Gujarat polls, the BJP had won 115 seats, the Congress 61 and others six.

POLLS WERE SEEN AS A LITMUS TEST FOR PM NARENDRA MODI AND RAHUL GANDHI WHO WAS ELEVATED AS PARTY PRESIDENT.

We have set a target of winning 150 seats in Gujarat assembly elections and the party would get that much seats.

— KAILASH VIJAYWARGIYA, BJP general secretary



Exit Polls

Final Result

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.

CSE 7315c

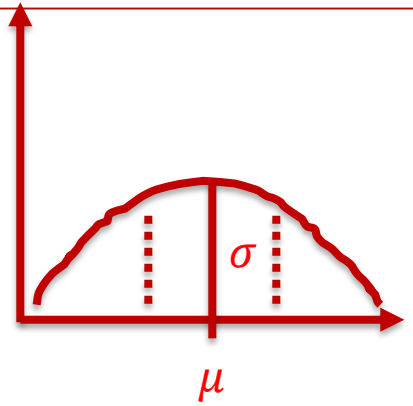
Publish Date	Source	Polling Organisation	NDA	UPA	Other
12 May 2014	[177]	CNN-IBN – CSDS – Lokniti	**276 (±6)**	97 (±5)	148 (±23)
	[177][178]	India Today – Cicero	**272 (±11)**	115 (±5)	156 (±6)
	[177][179]	News 24 – Chanakya	**340 (±14)**	70 (±9)	133 (±11)
	[177]	Times Now – ORG	**249**	148	146
	[177][180]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

Source: http://en.wikipedia.org/wiki/Indian_general_election,_2014
Last accessed: March 27, 2015

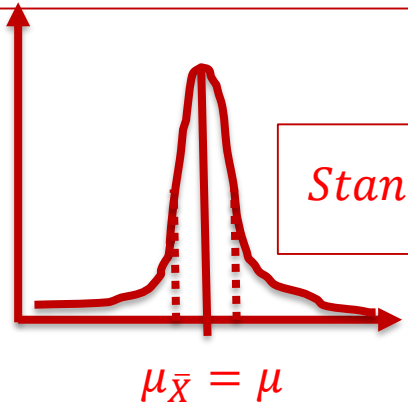
Polling Organisation	NDA	UPA	Other
CNN-IBN – CSDS – Lokniti	276 (±6)	97 (±5)	148 (±23)
India Today – Cicero	272 (±11)	115 (±5)	156 (±6)
News 24 – Chanakya	340 (±14)	70 (±9)	133 (±11)

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

Population distribution

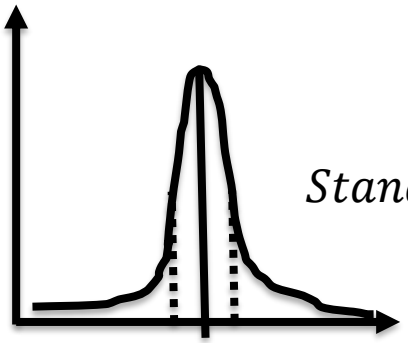


Sampling distribution of sample means



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

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



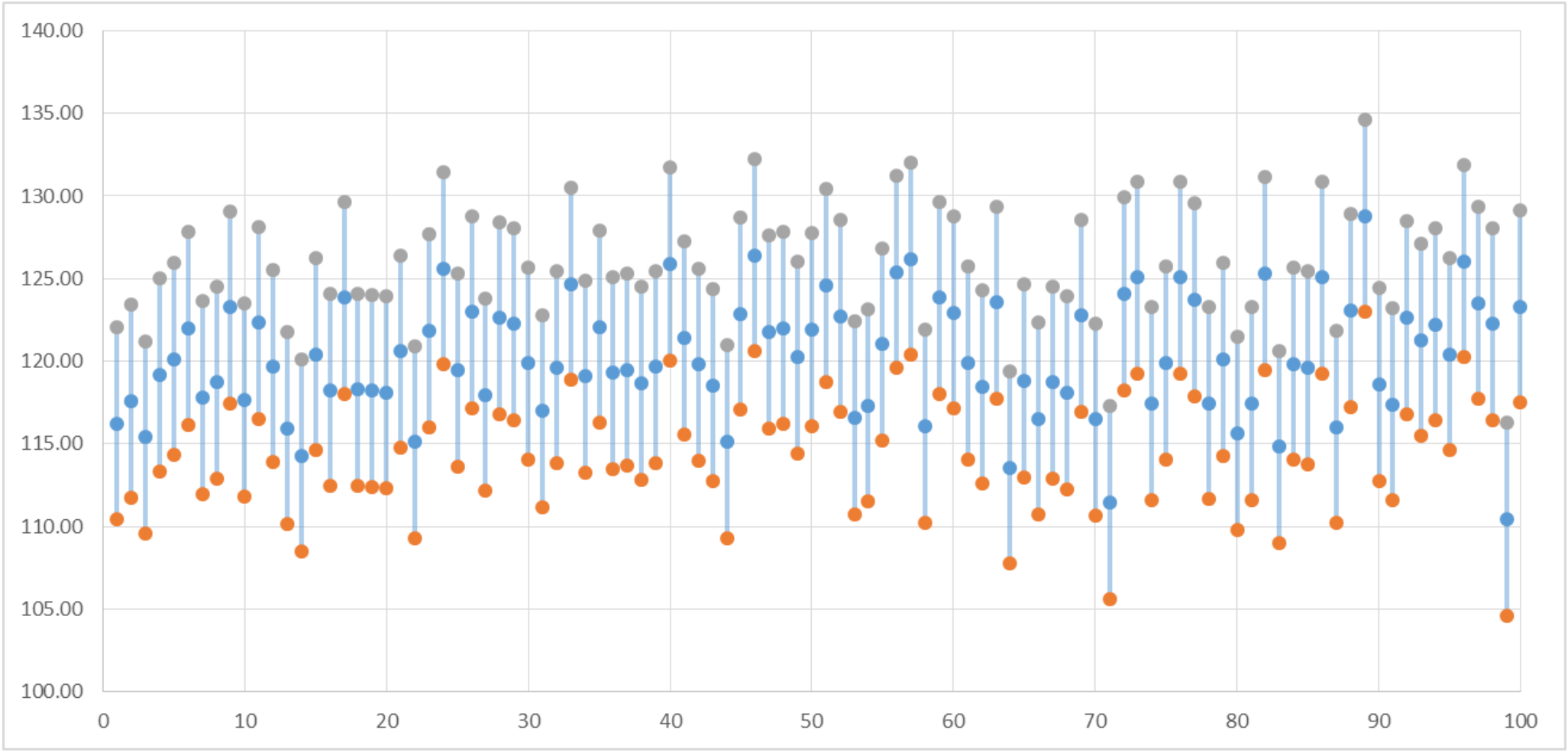
$$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$$

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

Alternatively, 95% of such intervals include the population mean.

Here, 95% is the Confidence Level and the interval is called the Confidence Interval.

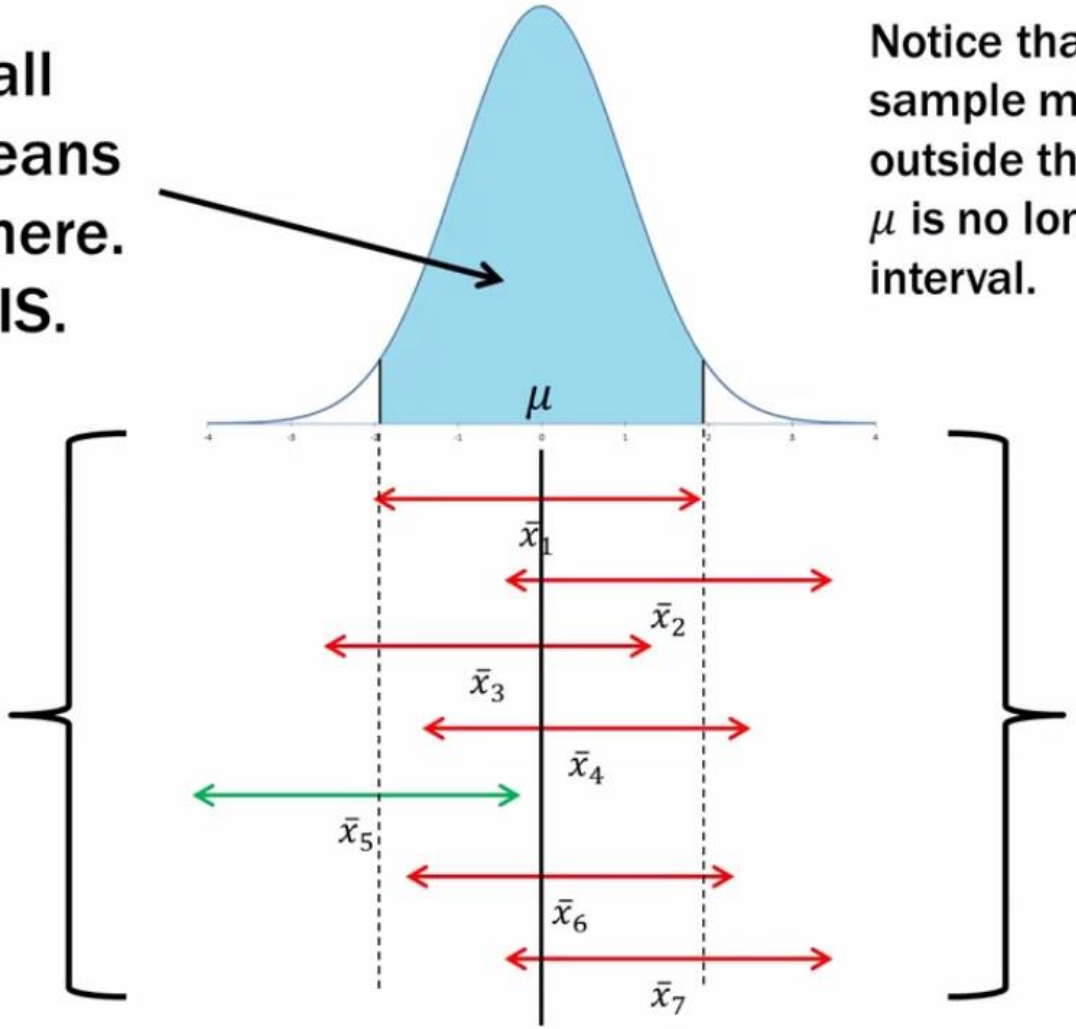
Confidence Level and Interval - Excel



94 of the 100 intervals contain the population mean.

Confidence Level and Interval

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



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

Many samples of the same size. THESE COME FIRST.

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

Polling Organisation	NDA	UPA	Other
CNN-IBN – CSDS – Lokniti	276 (±6)	97 (±5)	148 (±23)
India Today – Cicero	272 (±11)	115 (±5)	156 (±6)
News 24 – Chanakya	340 (±14)	70 (±9)	133 (±11)

SE or Margin of Error?

HUFFPOST POLLSTER

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

AdChoices

← PREVIOUS POLL | NEXT POLL →

POLL UPDATE

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

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

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

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

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%

Poll chart and latest estimates for 2016 National Republican Primary »

le, google.com/+deccanchronicle

IMD does not see any chance of monsoon

6-day delay in mon

DC CORRESPONDENTS
HYDERABAD / NEW DELHI,
MAY 15

Several days after the India Metrological Department raised hopes by predicting an above-average monsoon, it came out with unpleasant news on Sunday of a possible six-day delay in early showers, that were due to hit Kerala on June 1.

"The forecast suggests the monsoon's onset over Kerala this year is likely to be slightly delayed. The southwest monsoon is likely to set over Kerala on June 7, with a model error of plus or minus four days," said the IMD, which has got it right in the past 11 years. June 1 is the official onset date of the monsoon in India.

This will lead to a delay of six days in AP and two days in TS. IMD-Hyderabad assistant meteorologist M. Narsimha Rao said the normal onset date in AP is June 1, the same as of Kerala, and June 5 for Telangana.

He said there could be delays in the monsoon advancing, even after setting in over Kerala due to climatic conditions.

"The monsoon might set in over AP on June 7, the same day as Kerala, or later. The dates can be predicted only after it has set in on Kerala." Last year, the monsoon set in on AP on June 10 and advanced to Telangana on June 13 in Telangana. Rao said, "The dates predicted by IMD have a range of plus or minus 4 days and the prediction is done daily. It might change depending on the change in weather systems."

However, Accuweather website reported no rain warning, only "partly cloudy" sky for June 7, when the monsoon should set in over Nellore.

Page 4: Rain in Nellore likely on June 10

Saturday's thunderstorm brought trees and elect
crashing down in the city. A car lies crushed as a
it on the main road at Padmaraonagar. — S.Su

24-h
hits

DC CORR
HYDERAB

Hundre
kaput
in the
power
than
ATM
bran
five
po

Skymet hints at poor monsoon

J. UMAMAHESHWAR RAO | DC
VISAKHAPATNAM, MARCH 27

THE AGENCY HAS predicted 95 per cent rainfall (with an error margin of +/-5 per cent) of the long period average of 887 mm for the season.

The private weather forecasting agency Skymet has predicted that the southwest monsoon, that lasts from June to September, will be below normal. The first half of the season may see better rainfall than the latter half, Skymet said.

The agency predicted 95 per cent rainfall (with an error margin of +/-5 per cent) of the long period average of 887 mm for the season.

The India Meteorological Department will come out with its monsoon forecast next month. Skymet said pre-monsoon rains would be less during April that would lead to an intense heating of the land mass. Pre-monsoon activities may pick up pace during May, Skymet CEO Jatin Singh said.

Page 4: Skymet: El Nino recurrence likely

| 19 APRIL 2017 | HYDERABAD

Skymet predicts less than normal rainfall

DC CORRESPONDENT with agency inputs
NEW DELHI, APRIL 17

Southwest monsoon is likely to be "near normal" this year with possibility of a "good distribution" of rainfall acr-oss the country, the weatherman said.

A strong El Nino phenomenon causes sea temperatures to rise significantly, and has adverse effects on marine and aquatic life, agriculture and the quality of water supplies.

"The country will receive 96 per cent of Long Period Average with an error model of plus or minus 5 per cent," Ramesh said, while releasing the monsoon forecast.

Anything between 96 and 104 per cent of the LPA is considered "normal", while under 96 per cent rainfall is categorised as "below normal".

Interestingly, Skymet, a private weather forecasting agency, has predicted a "below normal rainfall" this year, with western India likely to experience a shortfall.

The IMD has not issued a region wise forecast yet. Ramesh said it will make a more detailed prediction in its second forecast in June.

However, rainfall may be somewhat deficient in the northeast and parts of south India.

2016 witnessed normal rainfall across the country, barring deficient precipitation in states like Karnataka, Kerala, Andhra Pradesh, Tamil Nadu, and the northeastern states.

Monsoon to be better than expected: IMD

New Delhi, June 6: The monsoon is likely to be better than earlier expected, according to the Met department which revised its initial forecast on Tuesday, upgrading it marginally.

The India Meteorological Department (IMD) director general, K.J. Ramesh, said the revision to 98 per cent precipitation of the Long Period Average (LPA), was done because of reduced chances of occurrence of an El-Nino, a phenomenon associated with the heating of the Pacific waters. In its initial forecast released in April, the IMD said the country could receive rainfall 96 per cent of the LPA.

"We are expecting a good rainfall across the country this year. July is likely to receive 96 per cent of the LPA while August is expected to witness precipitation of 99 per cent of the LPA," Ramesh said.

Swept by a heatwave, the North Indian plains and several parts of Central India are expected to witness a drop in the temperature in the next two days due to a western disturbance, which will bring thundershowers to this belt. Ramesh said rainfall in Central India is likely to be 100 per cent of the Long Period Average (LPA) and 99 per cent in the southern peninsula, where several parts are reeling under drought.

Northeast and Northwest India, a region that has been receiving deficient monsoon for the three consecutive years, is likely to get 96 per cent of the rainfall of the LPA.

According to meteorological parlance, anything between 96-104 is considered as "normal" rainfall and below it is "deficient".

Rainfall in the range of 104 to 100 is "above normal" and anything that surpasses it is considered as "excess". The IMD chief also allayed fears of occurrence of an El-Nino.

"The US' National Oceanic and Atmospheric Administration and the Australia's Bureau of Meteorology have also ruled out chances of an El-Nino," Ramesh said.

El-Nino is phenomena related to heating up of the Pacific waters and is said to have an adverse impact on the monsoon. *-PTI*

RAIN STATS

Quantitatively, monsoon rainfall is likely to be 98% of the Long Period Average (LPA). The LPA of the seasonal rainfall over the country as a whole for the period 1951-2000 is 89 cm

FORECAST PROBABILITY

7% chance of deficient rainfall (below 90%)

28% 28% chance of below normal rainfall (90-96%)

50% chance of Normal rainfall (96-104%)

SKIES | OPEN ■ **Rains in Andaman and Nicobar Islands 3 days ahead of onset date**

Monsoon may hit Kerala on June 1

New Delhi, May 14: The Southwest Monsoon has covered the Nicobar Islands and the entire south Andaman Sea, three days ahead of its normal onset date, the Indian Metereological Department (IMD) said on Sunday.

"In view of the strengthening and deepening of southwesterly winds, persistent cloudiness and rainfall, southwest monsoon has advanced into some parts of southeast Bay of Bengal, Nicobar Islands, entire south Andaman Sea and parts of north Andaman Sea today," the India Meteorological Department said.

IMD Director General K.G. Ramesh, however, said it was too early to forecast whether the monsoon would hit the Kerala coast ahead of schedule. The normal onset date over Kerala, termed as the official arrival of the seasonal rainfall in India, is June 1.

Mr Ramesh said the prevailing conditions do not suggest that monsoon could hit Kerala early simply because it has arrived early in the Andaman and Nicobar Islands.

The normal onset date for Southwest Monsoon over the Andaman and Nicobar Islands is May 17.

Mahesh Palawat, Chief Meteorologist with the Skymet, a private weather forecasting agency, said monsoon is likely to hit Kerala on June 1 with an error of one or two days.

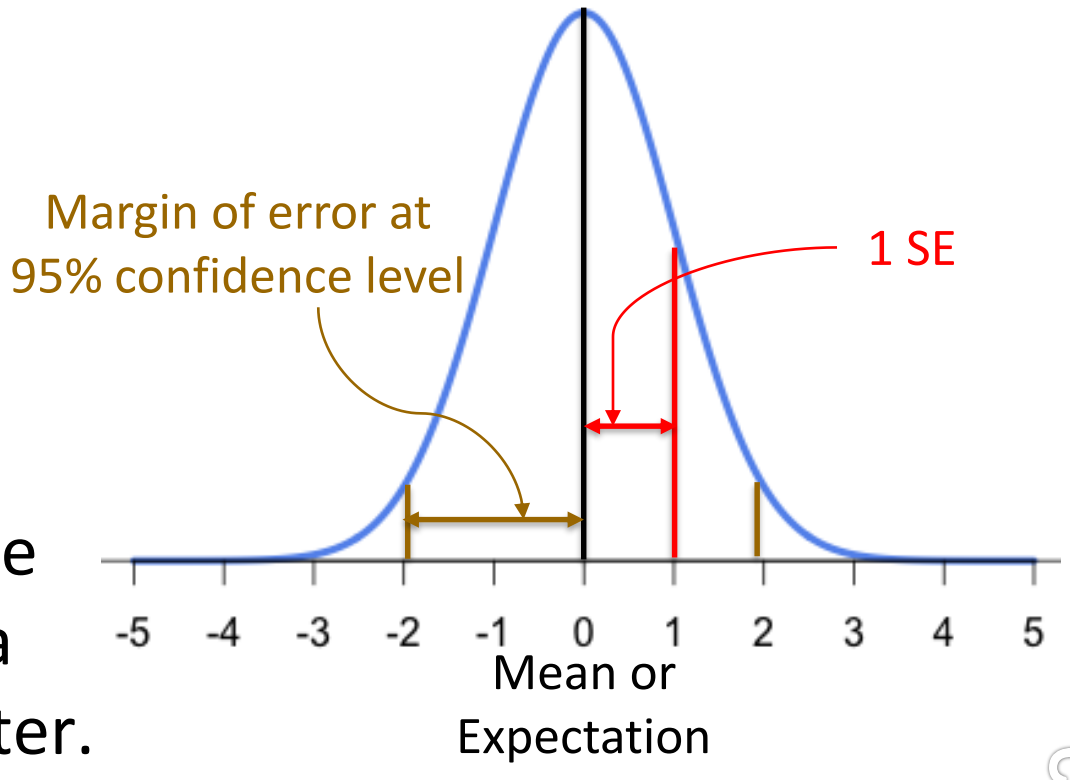
"Conditions are favourable for further advance of southwest monsoon into some parts of southwest Bay of Bengal, some more parts of southeast Bay of Bengal, remaining parts of Andaman Sea, Andaman and Nicobar Islands and some parts of east-central Bay of Bengal during the next 72 hours," the IMD said. — *PTI*

■ Skymet, a private weather forecasting agency, said Monsoon is likely to hit Kerala on June 1 with an error of one or two days, while IMD says it's early to predict.

SE, Margin of Error, Confidence Interval and Sample Size

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

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

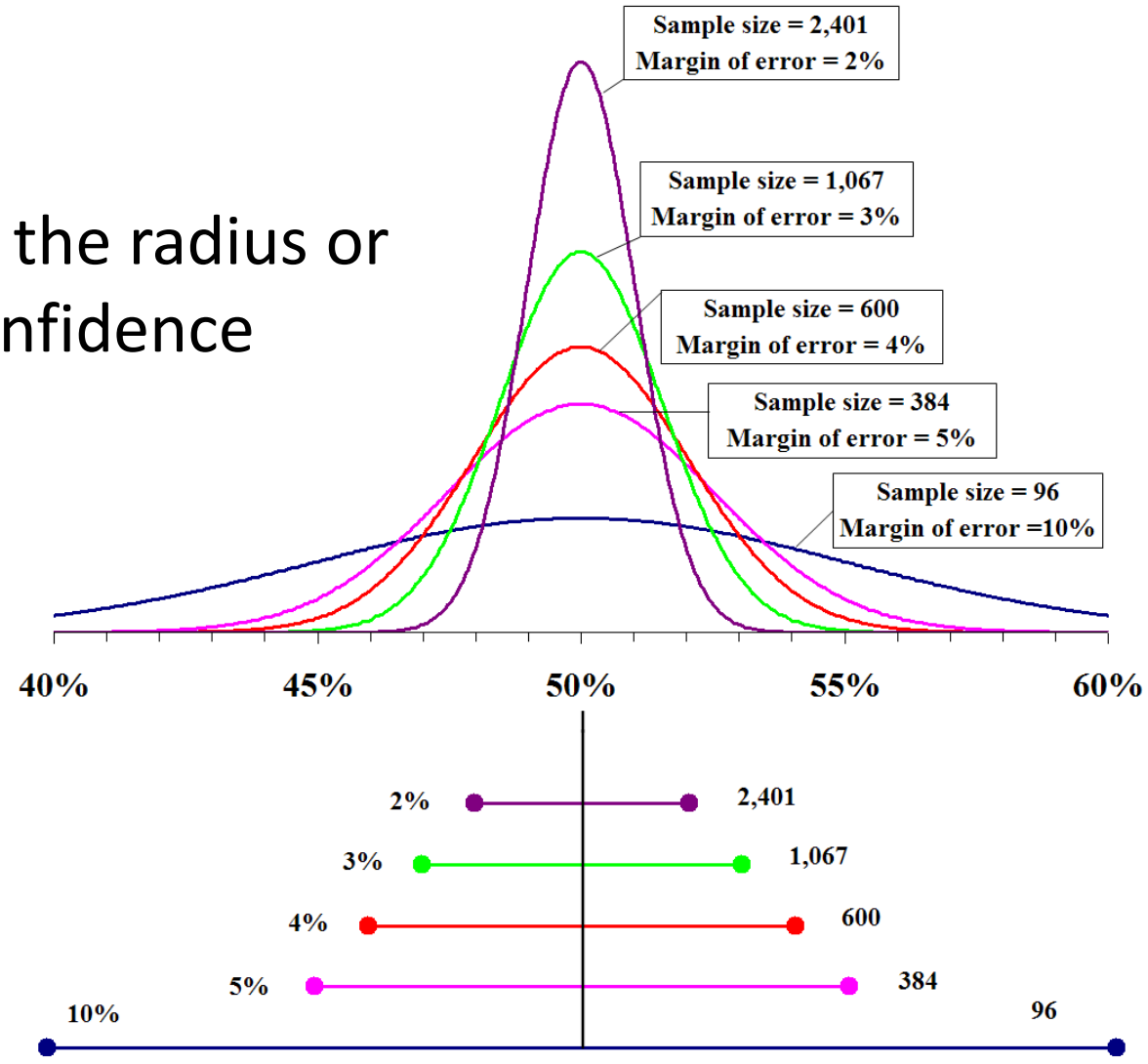


Margin of error is the **maximum expected difference** between the true population parameter and a sample estimate of that parameter.

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

SE, Margin of Error, Confidence Interval and Sample Size

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



Source: https://en.wikipedia.org/wiki/Margin_of_error
Last accessed: June 18, 2015

to have 3 zones

UK opinion polls give knife edge lead to Bremain at 51%

London, June 23: Millions of Britons today braved rains to cast their vote in the knife-edge referendum to decide whether the country will stay in or leave the 28-nation bloc, even as the latest survey gave the Remain camp its first solid lead in the bitterly-contested poll.

Brisk voting was reported and members of the public posted photographs of busy polling stations across the country.

Both sides of the campaign have appealed to a record number of registered voters — more than 46 million, including 1.2

British PM David Cameron and his wife Samantha after casting their votes.

million British Ind for a big turnout as P Minister David Cam made his final appe "get out there and

Remain" and reject the "untruths" of the camp in favour of 'Brexit' or Britain's exit from the European Union (EU).

Based on average of last six polls, the UK daily *Telegraph* said the remain is leading with 51 per cent vote. The exit is trailing with 49 per cent vote.

A late boost for the Remain camp came from an Ipsos Mori poll, giving it a 52 per cent against 48 per cent for Brexit. The

UK waits with ba

Brexit result remained too close to call with opinion polls showi

London, June 23: Millions of Britons began voting on Thursday in a bitterly-fought, knife-edge referendum that could tear up the island nation's EU membership and spark the greatest emergency of the bloc's 60-year history.

A record 46.5 mill voters have registere decide Britain's future in the 28-nation Europ Union, which was b out of a determination to unite in lasting peace after the carnage of two

ber of the European Union or leave the European Union?"

In one of the last opinion polls before the vote, "Remain" took a lead with 48 per cent against "Leave" on 42 per cent. Two other polls published on the eve of the referen- dum had put the "Leave" camp ahead by one or two points, well within the margin of error.

Using the hashtag #ivot- ed, some people posted mobile phone images of their completed ballot

FATE OF A NATIO

The referendum ballot paper asks a 'Yes' or 'No' question: "Should the United Kingdom remain a member of the European Union or leave the European Union?"

After the refer-

/deccannews, twitter.com/deccanchronicle, google.com/+deccanchronicle

EUROPAIN

Cameron exits over Brexit, says UK needs new leadership

London, June 24: Britain has voted to leave the European Union, forcing the resignation of Prime Minister David Cameron and dealing the biggest blow since World War II to the European project of forging greater unity.

Global financial markets plunged on Friday as results from a referendum defied bookmakers' odds to show a 52-48 per cent victory for the campaign to leave a bloc Britain joined more than 43 years ago.

The pound fell as much as 10 per cent against the dollar to touch levels last seen in 1985, on fears the decision could hit invest-

Nigel Farage, the leader of the UK Independence Party, reacts in celebration at a "Leave.EU" party. — PTI

Frenzied markets in a freefall, currency pounded

WHAT'S IN STORE NOW

With Brexit done, what's in store for U and the European Union now? The Bre summit will be held June 28 and 29.

REPORT ON PAGE

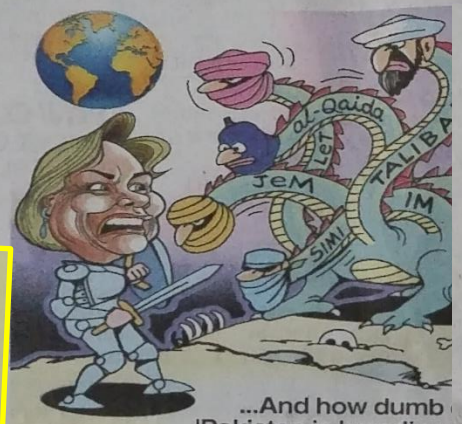
LONDONE SIGN FOR

Tens of thousa

CSE 7315c

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

Chidanand.Rajghatta
@timesgroup.com



...And how dumb
'Pakistan is breeding

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

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

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

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

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

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

"More to the point, Trump is a bigoted petulant eg ac with the temperam hyena. She's exper competent, and int What's going on?" Rei dered. What's going o polls never tell the fu much less the end resu servatives claim there reservoir of Trump ters who do not 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 polls to new algorithms tha to account past resul fic issues, demograp Some even project possible result in percentages to get around the certitude. Nate Silver, a baseball statistician who moved to politics after accurately calling the 2008 and 2012 elections, puts Hillary Clinton's in-chief, imitating something that had worked for Margaret Thatcher. The last pitch misfired badly. Polls did not account for it. It's the kind of pitfall both the camps keep watching for as much as the polls.

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

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

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

CSE 7315c

nie rally for Clinton; Don brings in Melania

MONKEY KING SAYS TRUMP WILL WIN

A Chinese monkey described as the "king of prophets" has picked Donald Trump for the US presidency. Geda had successfully predicted the winner of football's European Championship final earlier this year. — AFP


OW MUCH OF THE EXIT
POLLS DO YOU BELIEVE?
CNN NEW
new
U.S. GOOFED
00:02:12
TRUMP EXIT
POLL GAFFE
HIDDEN TRUM
VOTERS - MAN
DIDN'T ADMIT
THEY VOTED
TRUMP
CNN-NEWS18 STUDIO
RAJYA SABHA MP
LIVE
RESULT MATRIX

say what he doesn't mean.

If Modi wanted to avenge the Uri

other than the people of Jammu & Kashmir who repulsed them. And in 1999,

artistes and musicians to exit India.

The writer is a political commentator

'8 in 10 Indians have a favourable view of Modi ... 50% were critical of his Pakistan policy before Kashmir unrest'

*The Washington-based Pew Research Center recently released a new India survey that shows that over two years into his tenure as prime minister, most Indians remain upbeat about Narendra Modi even as they see India playing a larger role in the world. **Bruce Stokes**, director of global economic attitudes at Pew Research Center, spoke to **Nalin Mehta** about the survey's findings on Indian political attitudes, why Congress and Sonia Gandhi's favourability ratings have also increased in the past year and what India thinks about Modi's handling of Pakistan:*

■ Your survey found that a strong majority of 81% of Indians continue to have a favourable view of Modi. How does that compare to leaders in similar countries?

We found about 8 in 10 Indians have a favourable view of Modi. That was down 6 percentage points from last year but when you are talking of a stratosphere of around 80% it doesn't really matter that much. In comparative terms, the popularity of the US president in our most recent poll was 52%. Donald Trump says 82% of Russians like Putin. I am not sure where he got that data but in functioning democracies this approval rating of the Indian prime minister is really good and consistent.

■ But isn't there a partisan gap when people assess Modi's performance?

A majority of people who favour the Congress party, also have a favourable view of Modi. However, while most people have a very favourable view of the current PM, Congress people have only a somewhat favourable view. Fifty per cent or so of Congress-leaning people we interviewed had a very favourable or somewhat favourable view of Modi but only half of that 50% had a very favourable view, whereas BJP people overwhelmingly said very favourable.

When you ask specific questions about his conduct in office, about whether he gets things done, whether he understands people's issues, you see a much more partisan gap. BJP people are more likely to say yes than Congress people. This gap between BJP and Congress supporters has become bigger. Congress people are more critical of Modi this year than they were last year and interestingly BJP people are happier with him this year than they were last year.

■ What about the PM's handling of Pakistan?

We asked a number of questions about how he was handling relations with Pakistan, China and the US. The greatest criticism was in his handling of Pakistan. Fifty per cent of the population was critical. This survey was done before the Kashmir unrest this year so we don't know what people will say today. The public in India has an overwhelmingly negative view of Pakistan. They criticised Modi's handling of it.

■ Sonia Gandhi's approval rating has also gone up from 58% last year to 65% in 2016, and Rahul Gandhi's from 62% to 63% in your surveys. Have you been surprised?

This is a party that ran the country for most of its existence. My intuition would tell me that there is a core bedrock of people who are Congress people, whose parents were Congress people. A lot of them in the last election were clearly frustrated. My guess is their frustration with Congress has waned a bit now.

■ Arvind Kejriwal's approval rating dropped 10%. Why, do you think?

AAP had seemed to emerge as a third alternative but other than Delhi they have not materialised as a nationwide movement. This was a nationwide survey and there was quite a high level of 'I don't know' answers about AAP. For example, when we asked about Kejriwal, 23% said 'we don't know' compared to 3% about Modi. A lot of people either didn't know who we were talking about or hadn't thought about it. In a lot of places, especially rural and southern India, AAP is not a factor.

Q&A

■ But your polling sample is only 2,464 people nationwide. How can you be so confident with this sample size?

In our experience the design of a survey matters, not the number of people interviewed. The Economist calls us the gold standard of public opinion surveys because of our methodology. In the US we have for the last 20 years been the most accurate predictor of elections. In India, we predicted a landslide for Modi. We were right and every other pollster was wrong. I would love to interview 25,000 people but it's a question of resources. India is an emerging power and the world needs to know what Indians think as it affects what the Indian government does.

dilbert

■ But your polling sample is only 2,464 people nationwide. How can you be so confident with this sample size?

In our experience the design of a survey matters, not the number of people interviewed. The Economist calls us the gold standard of public opinion surveys because of our methodology. In the US we have for the last 20 years been the most accurate predictor of elections. In India, we predicted a landslide for Modi. We were right and every other pollster was wrong. I would love to interview 25,000 people but it's a question of resources. India is an emerging power and the world needs to know what Indians think as it affects what the Indian government does.

SE, Margin of Error, Confidence Interval and Sample Size

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

Expectation of sample proportions = p

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

SE, Margin of Error, Confidence Interval and Sample Size

In a poll by CNN/ORC conducted between November 27 – December 1, 2015, a survey of 930 randomly sampled registered voters predicted that 49% would vote for Hillary Clinton.

What is the margin of error at 95% confidence level (z = 1.96)?

Check qnorm(0.975, 0, 1). Why 0.975?

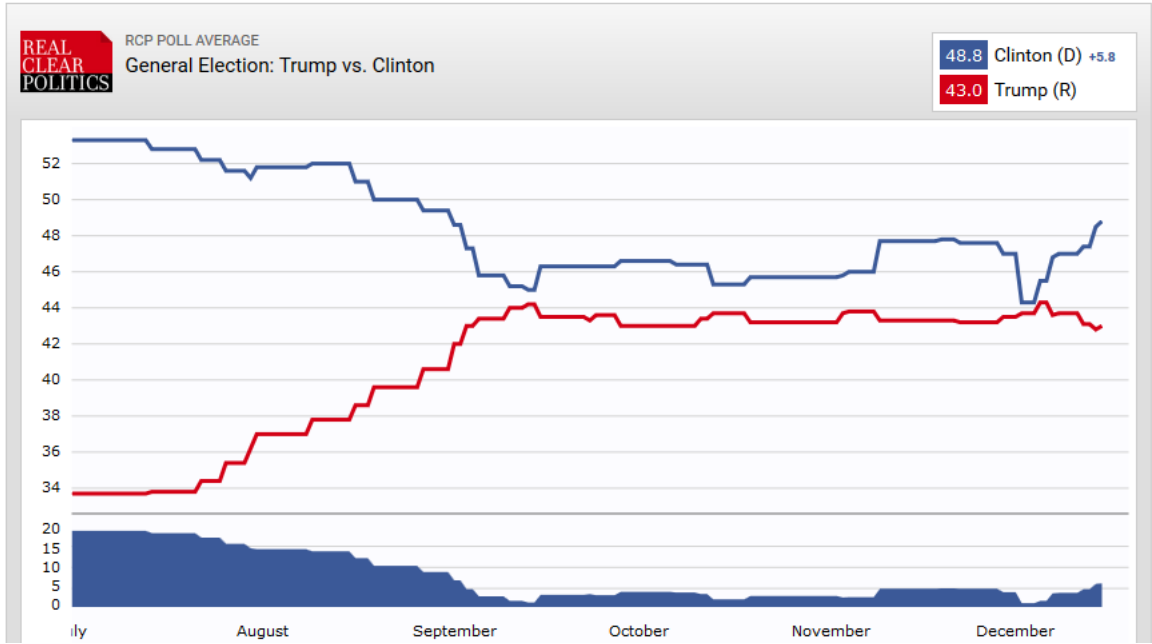
RealClear Politics

Polls | Election 2016 | Video | Wri

Polling Data

Poll	Date	Sample	MoE	Clinton (D)	Trump (R)	Spread
RCP Average	11/23 - 12/13	--	--	48.8	43.0	Clinton +5.8
ABC News/Wash Post	12/10 - 12/13	851 RV	4.0	50	44	Clinton +6
NBC/WSJ	12/6 - 12/9	849 RV	3.4	50	40	Clinton +10
USA Today/Suffolk	12/2 - 12/6	1000 LV	3.0	48	44	Clinton +4
CNN/ORC	11/27 - 12/1	930 RV	3.0	49	46	Clinton +3
Quinnipiac	11/23 - 11/30	1473 RV	2.6	47	41	Clinton +6

All General Election: Trump vs. Clinton Polling Data



SE, Margin of Error, Confidence Interval and Sample Size

$$\text{Margin of error} = 1.96 * \sqrt{\frac{0.49 * 0.51}{930}} \approx 3.2\%$$

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

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

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

Other ways of estimating the data size

- The rule of thumb
  - Count the total number of levels (assume 10 levels for numeric)
  - Multiply with number classes
  - Multiply with 75-150

Example

- Will a patient adhere or not?
  - Age (young, middle, old), income (low, medium, high), gender (male, female), education (high school; college; university)

Data need

Attributes	Levels	Classes	Rule of Thumb (x75)
2	2	2	600
5	5	3	5625
10	5	4	15000
20	5	2	15000
50	5	3	56250
100	5	4	150000

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

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

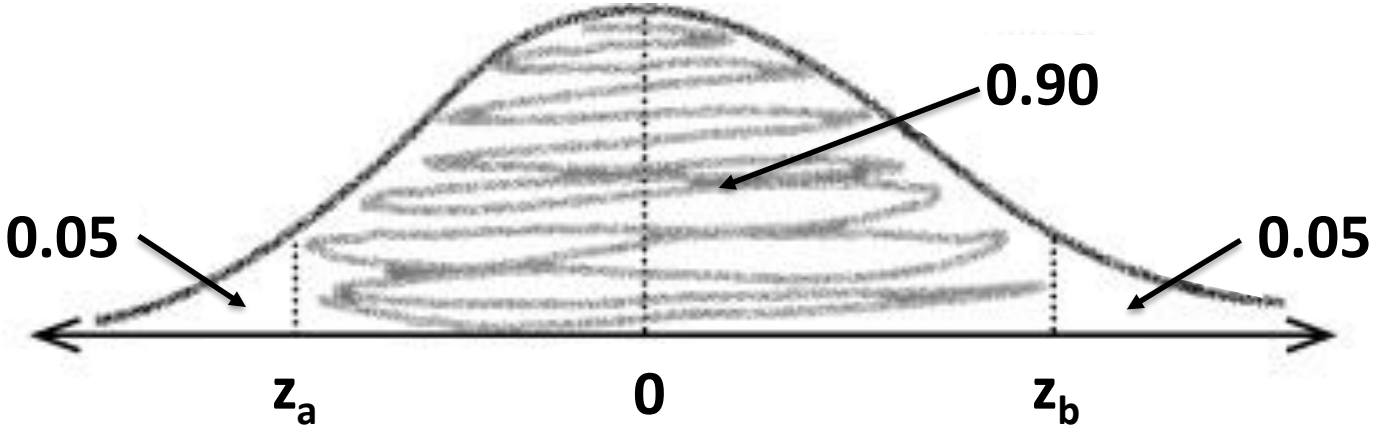
$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

$\therefore$ 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
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

From probability tables using interpolation, we get $z_a = -1.645$ and $z_b = 1.645$.

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

Confidence Intervals

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

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

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

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

The analyst is 90% confident that if a census of all US companies trading with firms in India were taken at the time of the survey, the actual population mean number of trading years of such firms would be between 8.545 and 12.365 years.

Shortcuts for Calculating Confidence Intervals

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

You took a sample of 50 Gems and found that in the sample, the proportion of red Gems is 0.25. Construct a 99% confidence interval for the proportion of red Gems in the population.

$$0.25 - 2.58 * \sqrt{\frac{0.25 * 0.75}{50}} < p < 0.25 + 2.58 * \sqrt{\frac{0.25 * 0.75}{50}}$$

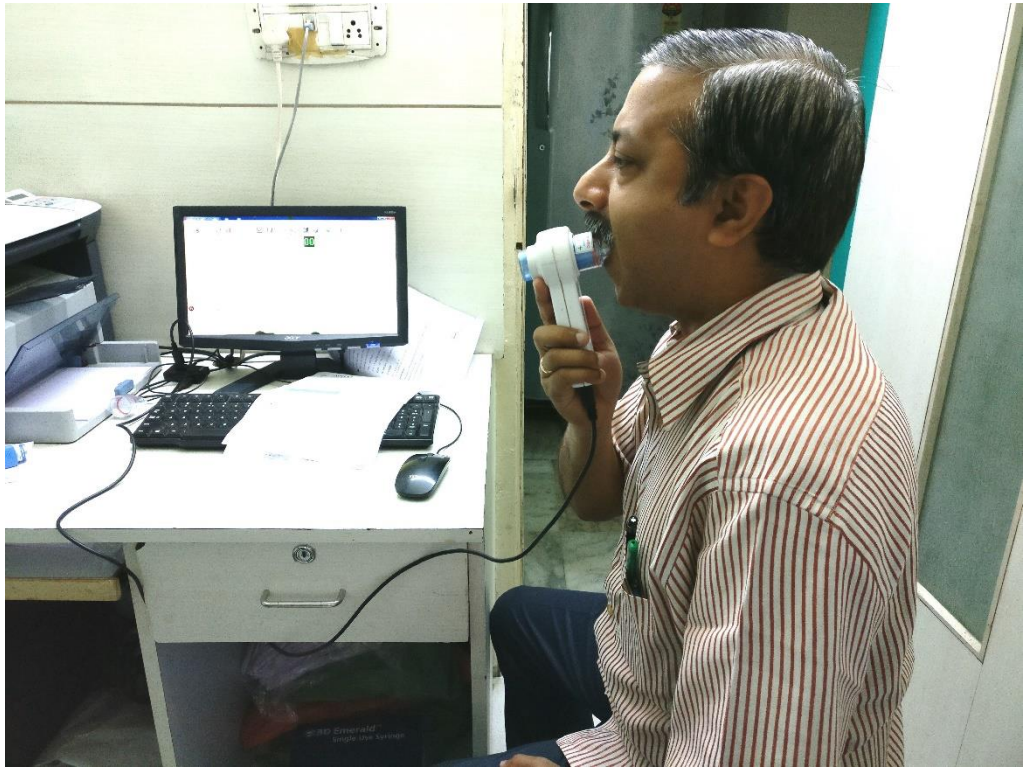
$$0.09 < p < 0.41$$

Shortcuts for Calculating Confidence Intervals

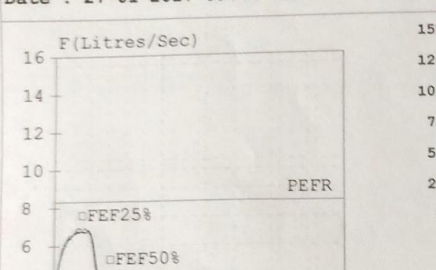
The lung function in 57 people is tested using FEV1 (Forced Expiratory Volume in 1 Second) measurements. The mean FEV1 value for this sample is 4.062 litres and standard deviation, *s* is 0.67 litres. Construct the 95% Confidence Interval.

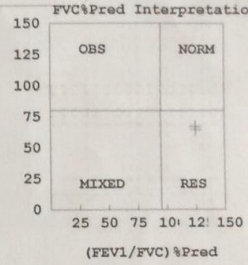
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

Shortcuts for Calculating Confidence Intervals



Pred.Eqns: RECORDERS
Date : 27-01-2017 09:03 AM
ID: 03132 Temp : degrees





FVC Results

Parameter		Pred	M.Pre	%Pred	M.Post	%Pred	%Imp
FVC	(L)	03.15	02.06	065	02.11	067	+02
FEV1	(L)	02.52	02.06	082	02.11	084	+02
FEV1/FVC	(%)	80.00	100.00	125	100.00	125	---
FEF25-75	(L/s)	03.55	04.75	134	04.84	136	+02
PEFR	(L/s)	08.32	06.88	083	06.76	081	-02
FIVC	(L)	-----	02.28	---	02.31	---	+01
FEV.5	(L)	-----	01.92	---	01.93	---	+01
FEV3	(L)	03.06	02.06	067	02.11	069	+02
PIFR	(L/s)	-----	05.57	---	05.32	---	-04
FEF75-85	(L/s)	-----	02.14	---	02.00	---	-07
FEF.2-1.2	(L/s)	06.21	06.19	100	06.19	100	---
FEF 25%	(L/s)	07.60	06.49	085	06.52	086	---
FEF 50%	(L/s)	05.31	06.44	121	06.32	119	-02
FEF 75%	(L/s)	02.38	02.57	108	02.63	111	+02
FEV.5/FVC	(%)	-----	93.20	---	91.47	---	-02
FEV3/FVC	(%)	97.14	100.00	103	100.00	103	---
FET	(Sec)	-----	00.74	---	00.78	---	---
ExplTime	(Sec)	-----	00.10	---	00.11	---	---
Lung Age	(Yrs)	047	055	117	055	117	---
FEV6	(L)	03.15	-----	---	-----	---	---
FIF25%	(L/s)	-----	05.36	---	05.28	---	--01
FIF50%	(L/s)	-----	05.49	---	05.07	---	--08
FIF75%	(L/s)	-----	04.40	---	03.94	---	--10

Pre Test COPD Severity
Test within normal limits
Post Test COPD Severity
Test within normal limits

Pre Medication Report Indicates
Mild Restriction as (FEV1/FVC)%Pred >95 and FVC%Pred <80
Post Medication Report Indicates
Mild Restriction as (FEV1/FVC)%Pred >95 and FVC%Pred <80

Dr S. Kishore Kumar R.
MBBS, DTCD
Regd. No: 39866

FEV1 values of 57 male medical students

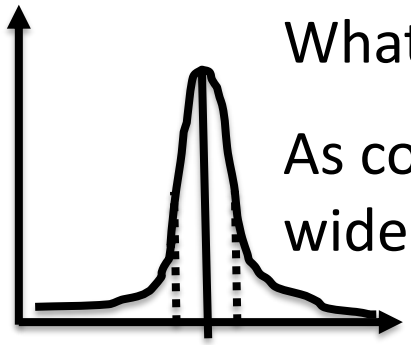
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

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?

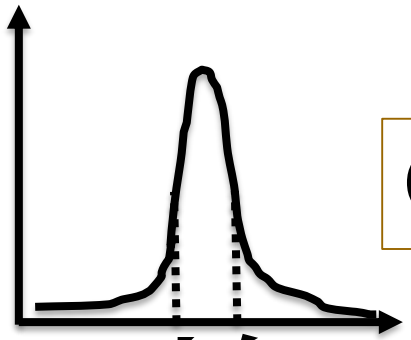
As confidence level increases, the confidence interval becomes wider and *vice-versa*.

What happens to the confidence interval as sample size changes?

As sample size increases, the confidence interval becomes narrower.

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

The Connection



Confidence Interval = Sample Statistic ± Margin of Error

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

$$Max(\bar{x} - \mu)$$

Margin of error = z * Standard Error

(Recall the standardization formula and CLT)

Depends on the confidence level

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

Probability density.

Area under the curve between the limits.

Probability that a certain % of samples will contain the population mean within this interval.

Standard deviation of the population: A sort of "average" deviation from the mean.

Interview Question

If you toss a coin 20 times and get 15 heads, would you say the coin is biased?

Let us apply our learning thus far…

Q. What distribution is it?

A. Binomial; X~B(20,0.5) assuming the coin is fair.

Q. What is the expectation?

A. $np = 10$

Q. What is the standard deviation?

A. $\sqrt{npq} = \sqrt{5} = 2.236$

Q. How many standard deviations away from the mean is 15?

A. $\frac{15-10}{2.236} = 2.236$

Q. What is the probability of getting 15 or more heads?

A. $P(X \geq 15) = P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20) = 0.021$ *pbinom(14, 20, 0.5, lower.tail = FALSE, log.p = FALSE)*

Q. Can it be approximated to a normal distribution?

A. $np = 10\ and\ nq = 10$. Since both are greater than 5, it can be approximated to X~N(10,5)

Q. What is the probability of getting 15 or more heads?

A. $\mathrm{P}(\mathrm{X} > 14.5) = 1 - P(X < 14.5)$

Q. What is the z-score?

A. $\frac{14.5-10}{\sqrt{5}} = 2.01. \therefore P = 1 - 0.9778 = 0.022$

pnorm(14.5,10,sqrt(5), lower.tail = FALSE)

pnorm((14.5-10)/sqrt(5),0,1, lower.tail = FALSE)

INFERENTIAL STATISTICS

HYPOTHESIS TESTS

Hypothesis tests give a way of using samples to test whether or not statistical claims are likely to be true or not.

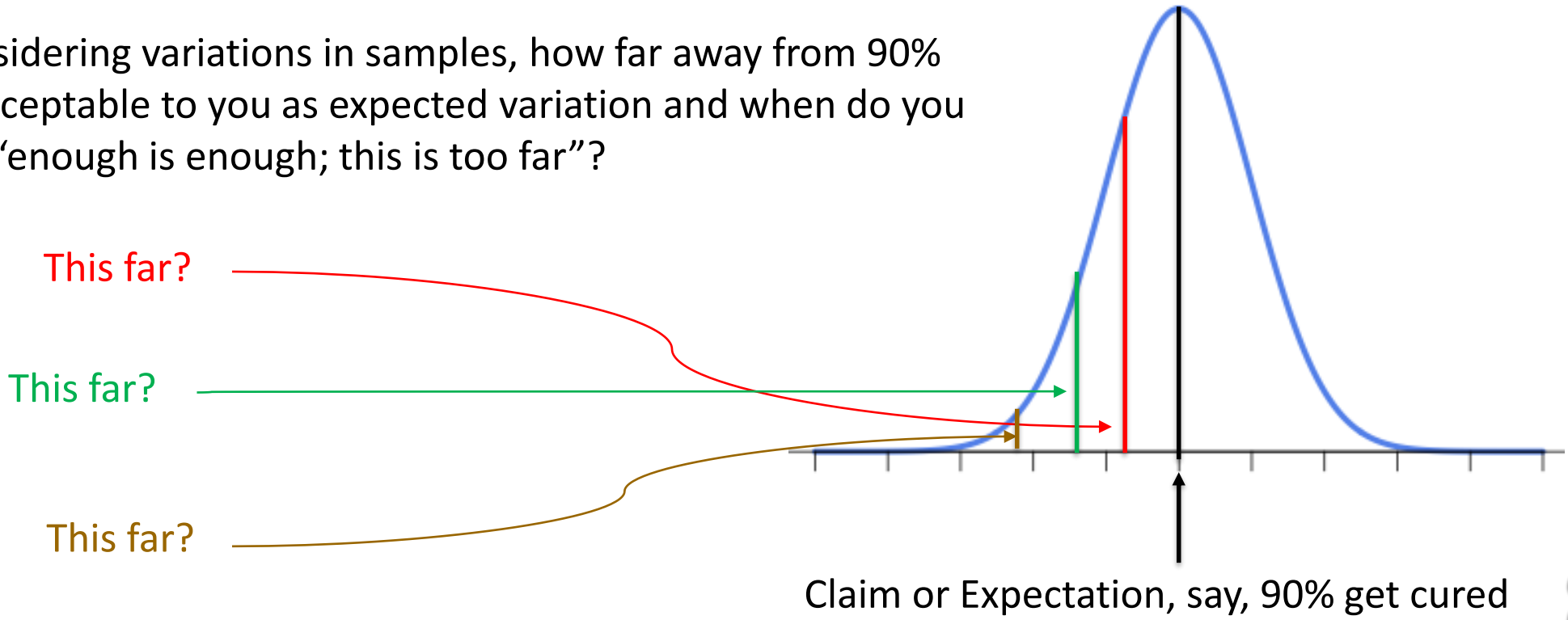
Hypothesis tests give a way of using samples to test whether or not statistical claims are likely to be true or not.

Dr. Unsnora prescribes SnoreCull to 15 of her patients and records whether it cured them or not after 2 weeks. She found that 11 were cured and 4 were not.

If the drug maker claimed that 90% get cured, 13.5 or 14 patients should have been cured. Is the company making false claims or is the doctor's sampling biased?

Hypothesis Testing Process

Considering variations in samples, how far away from 90% is acceptable to you as expected variation and when do you say “enough is enough; this is too far”?



Step 1: Decide on the hypothesis

SnoreCull cures 90% of the patients within 2 weeks.

This is called Null Hypothesis and is represented by H_0.

In this case, H_0: $p = 0.9$

If Null Hypothesis is rejected based on evidence, an Alternate Hypothesis, H_1, needs to be accepted. **We always start with the assumption that Null Hypothesis is true.**

In this case, H_1: $p < 0.9$

Examples of Hypotheses

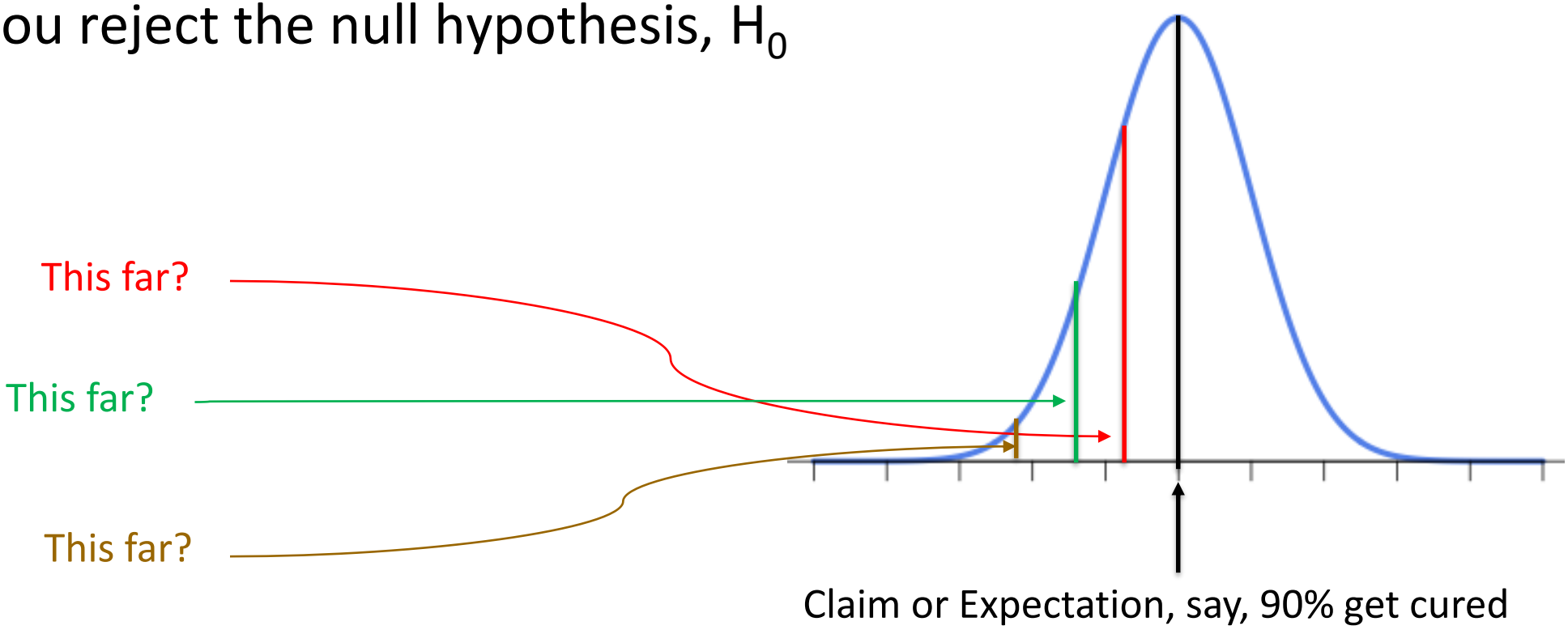
- Two hypotheses in competition:
  - H_0: The NULL hypothesis, usually the most conservative.
  - H_1 or H_A: The ALTERNATIVE hypothesis, the one we are actually interested in.
- Examples of NULL Hypothesis:
  - The coin is fair
  - The new drug is no better (or worse) than the placebo
- Examples of 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

Step 2: Choose your statistic

$X \sim B(15, 0.9)$

Step 3: Specify the Significance Level

First, we must decide on the Significance Level, α. It is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H_0



Step 4: Determine the critical region

If X represents the number of snorers cured, the critical region is defined as $P(X < c) < \alpha$ where $\alpha = 5\%$.

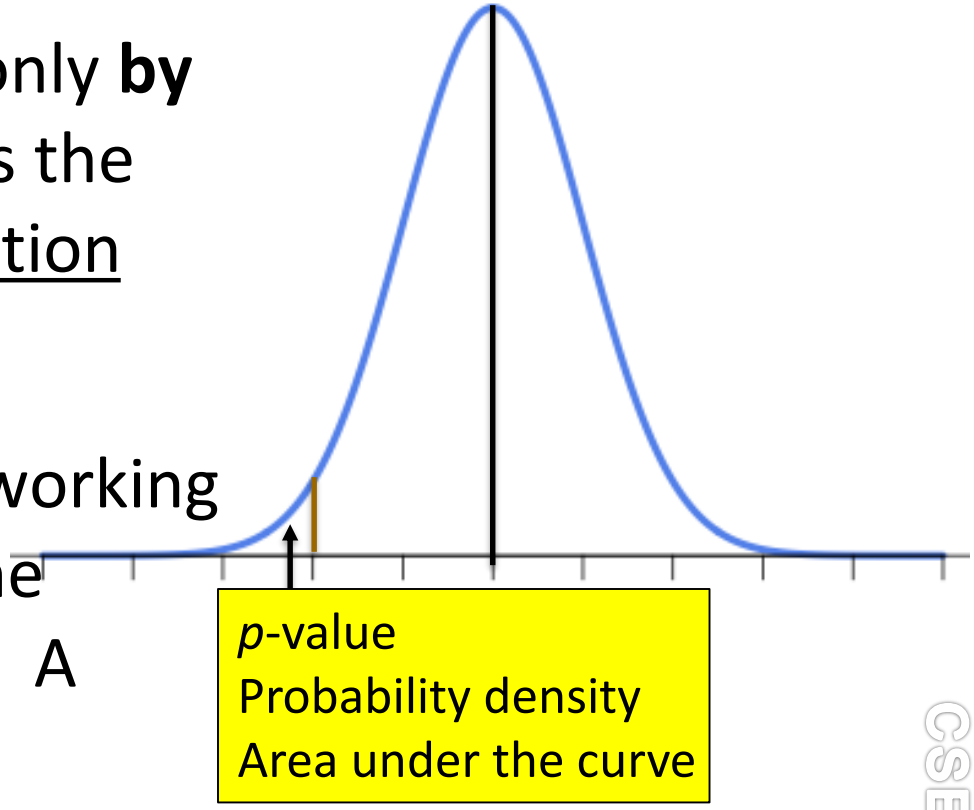
Recall that in a 95% CI, there is a 5% chance that the sample will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 90% snorers are cured, is rejected.

That is the reason 5% or 0.05 is called the Significance Level. In a 99% CI, 0.01 is the Significance Level.

Step 5: Find the *p*-value

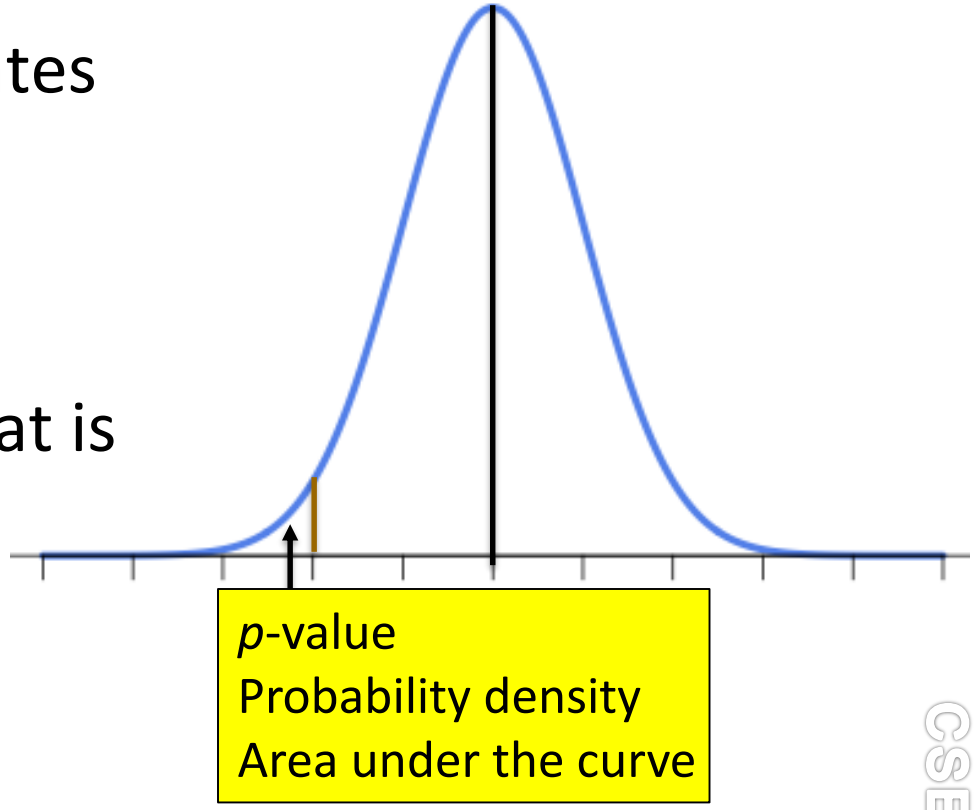
p-value is the probability of getting only **by chance** a value at least as extreme as the one in the sample under the assumption that the null hypothesis is true.

It is a way of taking the sample and working out whether the result falls within the critical region of the hypothesis test. A value in the critical region indicates presence of a real effect when the null hypothesis represents presence of no effect.



Step 5: Find the *p*-value

For example, a *p*-value of 0.01 indicates that, under the assumption that null hypothesis is true, there is only a 1% probability that the observed result occurred **by chance**. This means what is observed is a **real effect** (when null hypothesis represents no effect).



Essentially, this is the value used to determine whether or not to reject the null hypothesis.

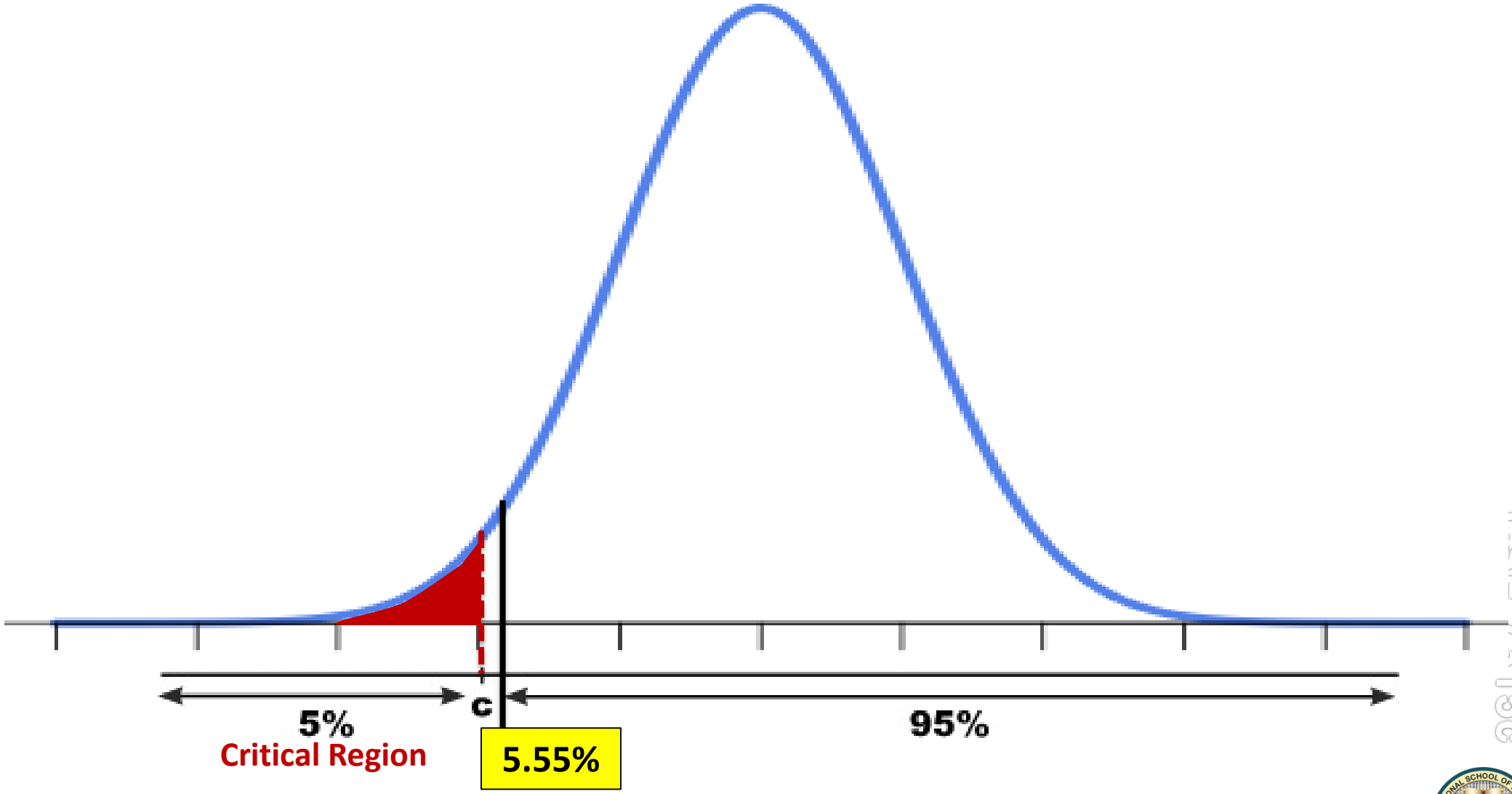
Step 5: Find the *p*-value

In the SnoreCull test done by Dr. Unsnora, 11 people were cured. This means our *p*-value is $P(X \leq 11)$, where X is the distribution of the number of people cured in the sample.

If $P(X \leq 11) < 0.05$ (Significance Level), it indicates that 11 is inside the critical region, and hence H_0 can be rejected.

Given that $X \sim B(15, 0.9)$, $P(X \leq 11) = 1 - P(X \geq 12) = 0.0555$

Step 6: Is the sample result in the critical region?



Step 7: Make your decision

There isn't sufficient evidence to reject the null hypothesis and so, the claims of the company are accepted.

Dr. Unsnora is not convinced and did another test with 100 people where 80 got cured and 20 didn't. What is your decision going to be now?

What are the null and alternate hypotheses?

H_0: p = 0.9

H_1: p < 0.9

What is the test statistic?

X ~ B(100, 0.9) Oh! Dear

What probability distribution can be used to approximate the Binomial distribution?

Since np>5 and nq>5, binomial distribution can be approximated to normal distribution.

What is the probability of 80 or fewer getting cured?

$$z = \frac{80.5 - np_0}{\sqrt{np_0(1-p_0)}} = \frac{80.5 - 90}{\sqrt{100 * 0.9 * 0.1}} = \frac{80.5 - 90}{\sqrt{9}} = -3.17$$

p-value = P(Z < -3.17) = 0.0008

CONTINUITY CORRECTION FACTOR

What is your decision?

Since the *p*-value (0.0008) is less than the Significance Level of 0.05, the null hypothesis can be rejected.

Attention Check

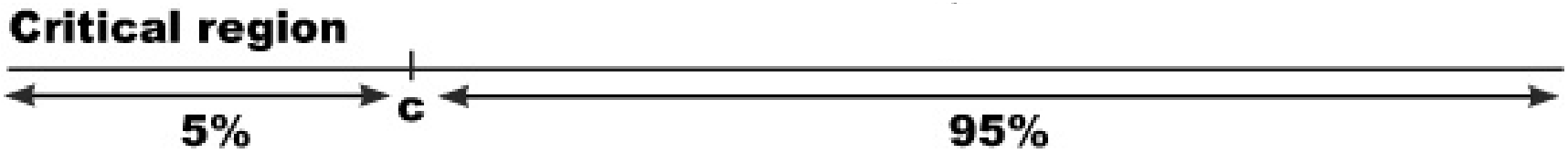
In hypothesis testing, do you assume the null hypothesis to be true or false?

True.

If there is sufficient evidence against the null hypothesis, do you accept it or reject it?

Reject it.

Attention Check



If the *p*-value is less than 0.05 for the above significance level, will you accept or reject the null hypothesis?

Reject it.

Do you need weaker evidence or stronger to reject the null hypothesis if you were testing at the 1% significance level instead of the 5% significance level?

Stronger.

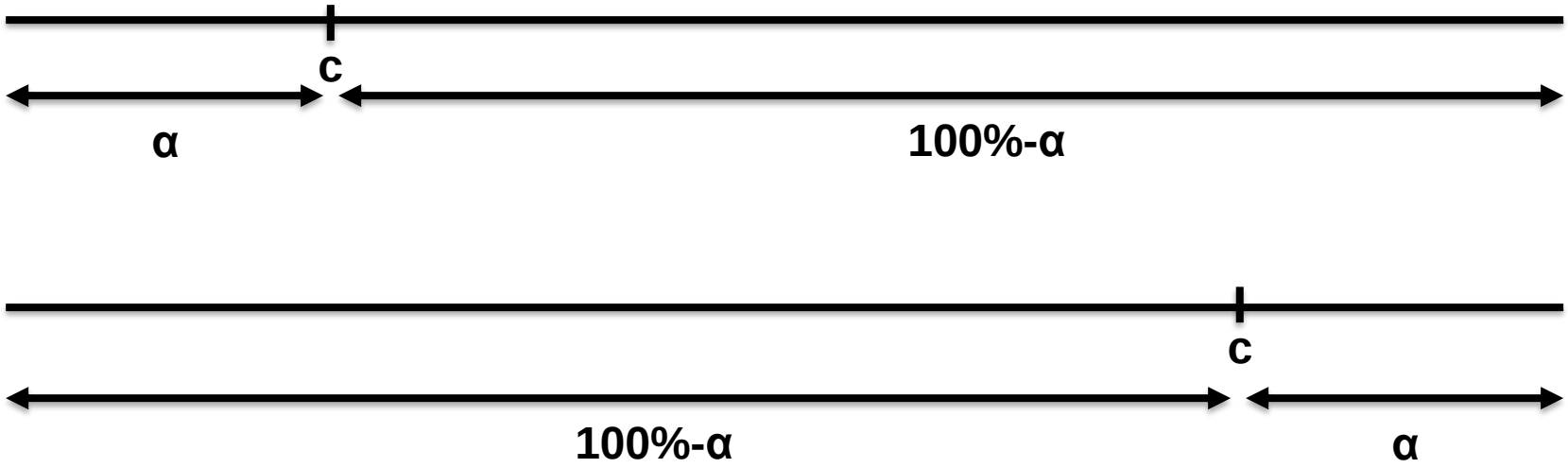
Critical Region Up Close

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.

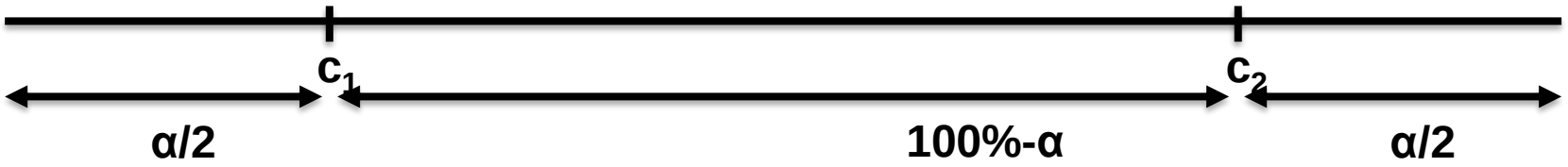


Critical Region Up Close

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.



Critical Region Up Close

For each of the scenarios below, identify what type of test you would require.

- SnoreCull hypothesis test as discussed till now.
  One-tailed/Lower-tailed

- If we were checking whether significantly more or significantly fewer than 90% patients had been cured, i.e., H_1: $p \neq 0.9$.
  Two-tailed test

- The coin is biased.
  Two-tailed test

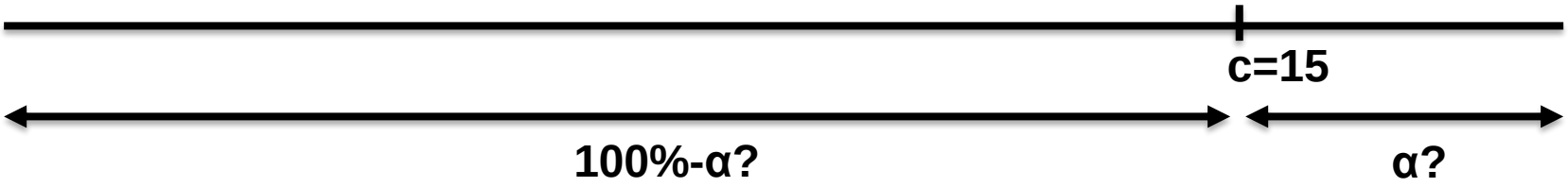
- The coin is biased towards heads with probability 0.8.
  One-tailed/Upper-tailed

The Missing Link in the Interview

Q. What is the probability of getting 15 or more heads?

A. $P(X \geq 15) = P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20) = 0.021$

What can you now say about the coin being biased or not?



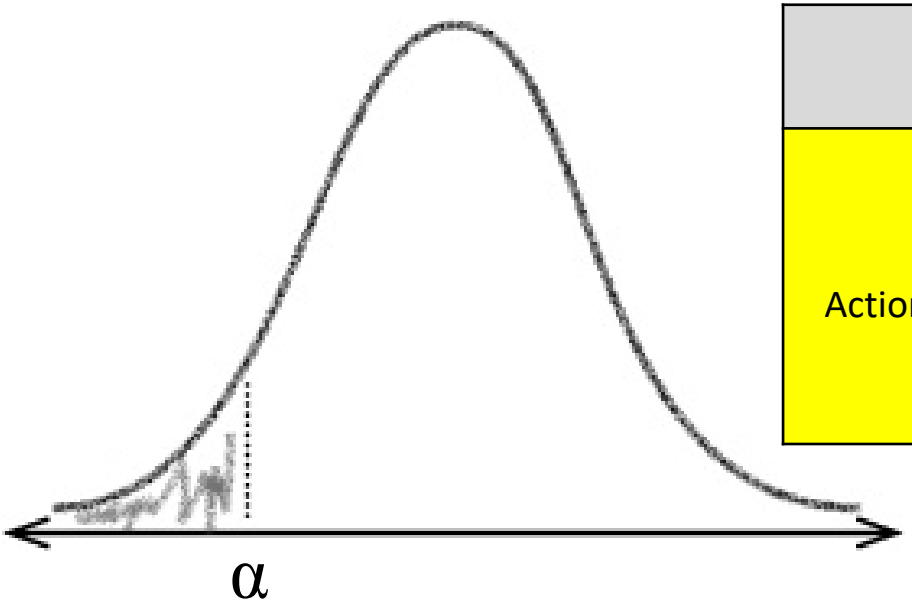
The hypothesis test doesn't answer the question whether the coin is biased or not; it only states whether the evidence is enough to reject the null hypothesis or not ***at the chosen significance level***.

Errors

- Type I: We reject the NULL hypothesis incorrectly
- Type II: We "accept" it incorrectly

		State of Nature	
		Null true	Null false
Action	Fail to reject null (negative)	Correct decision True Negative Specificity P(*Accept* H_0 | H_0 True)	Type II error (β) False Negative P(*Accept* H_0 | H_0 False)
	Reject null (positive)	Type I error (α) False Positive P(Reject H_0 | H_0 True)	Correct decision (Power) True Positive Sensitivity/Recall P(Reject H_0 | H_0 False)

Probability of Getting Type I Error



		State of Nature	
		Null true	Null false
Action	Fail to reject null (negative)	Correct decision True Negative Specificity	Type II error (β) False Negative
	Reject null (positive)	Type I error (α) False Positive	Correct decision (Power) True Positive Sensitivity/Recall

P(Type I error) = α

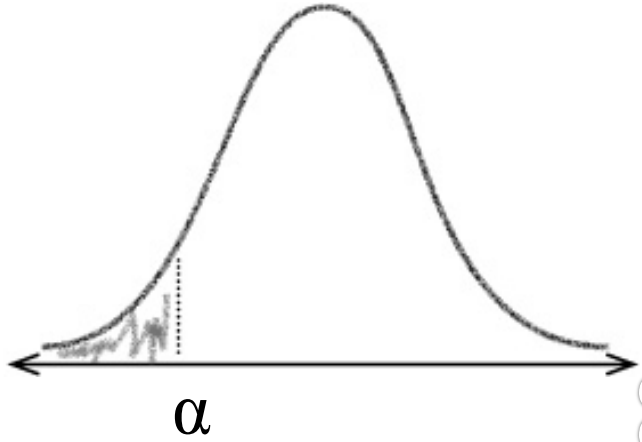
Probability of Getting Type II Error

P(Type II error) = β

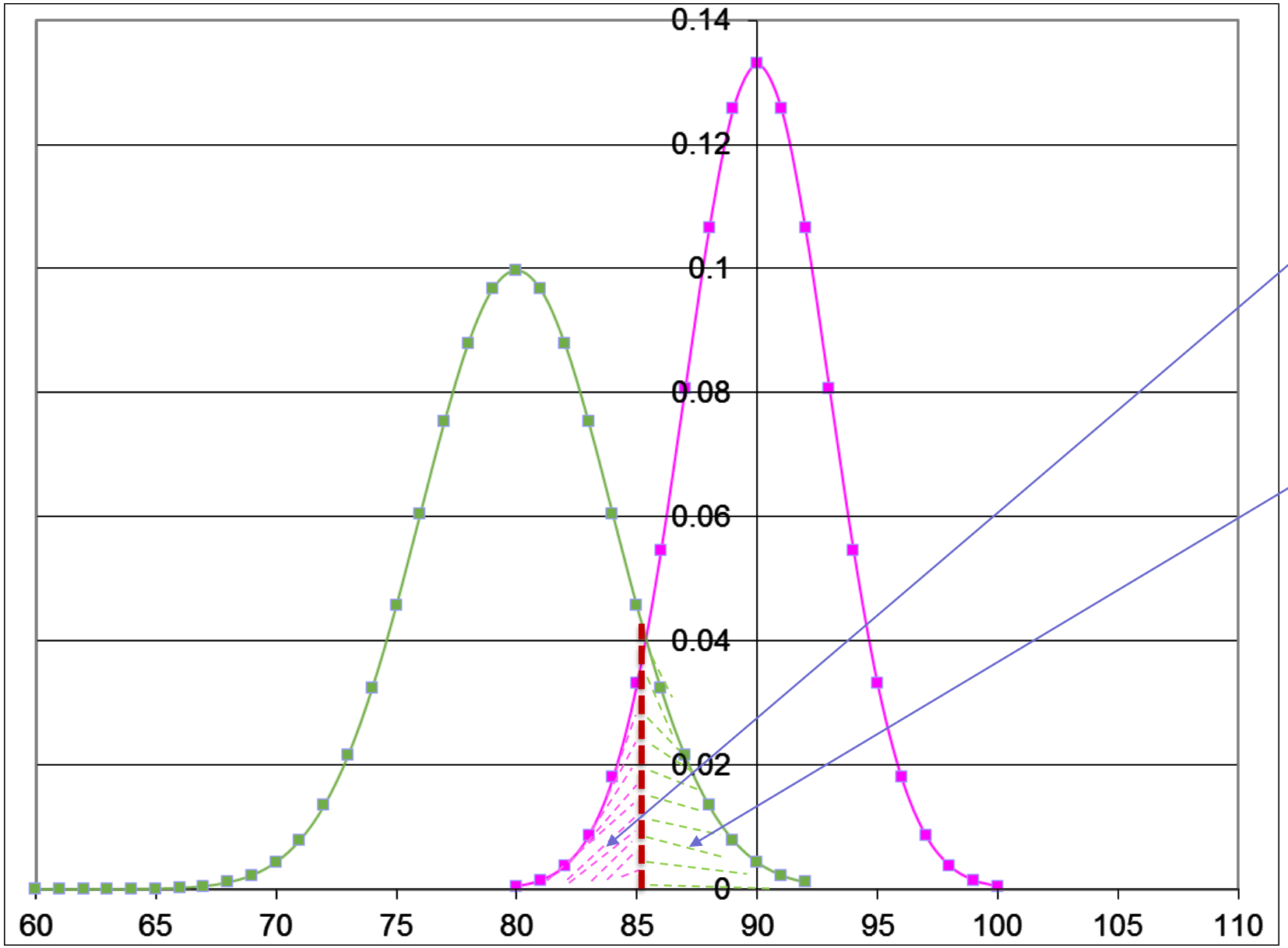
To find β (the difficult way to remember),

1. Check that you have a specific value for H_1.
2. Find the range of values outside the critical region of the test. If the test statistic has been standardized, it needs to be de-standardized for the purpose.
3. Find the probability of getting this range of values, assuming H_1 is true. In other words, find the probability of getting the range of values outside the critical region, but this time using the test statistic described by H_1 and not H_0.

		State of Nature	
		Null true	Null false
Action	Fail to reject null (negative)	Correct decision True Negative Specificity	Type II error (β) False Negative
	Reject null (positive)	Type I error (α) False Positive	Correct decision (Power) True Positive Sensitivity/Recall



Probabilities of Type I and Type II Errors



Probabilities of Errors in Our Example

P(Type I error) = 0.05

To calculate P(Type II error)

H_0: p = 0.9

H_1: p = 0.8

P(Z < z) = 0.05 for 5% Significance Value. From probability tables, z = -1.64.

To de-standardize and find values outside the critical region,

$\frac{X-90}{\sqrt{9}} \geq -1.64; X \geq 85.08$, i.e., we would accept null hypothesis if 85.08 or more people out of 100 had been cured.

Probabilities of Errors in Our Example

Finally, we need to calculate P($X \geq 85.08$), assuming H_1 is true.

$X \sim N(np_0, np_0(1-p_0))$ where n=100 and p_0=0.8.
This gives $X \sim N(80,16)$.

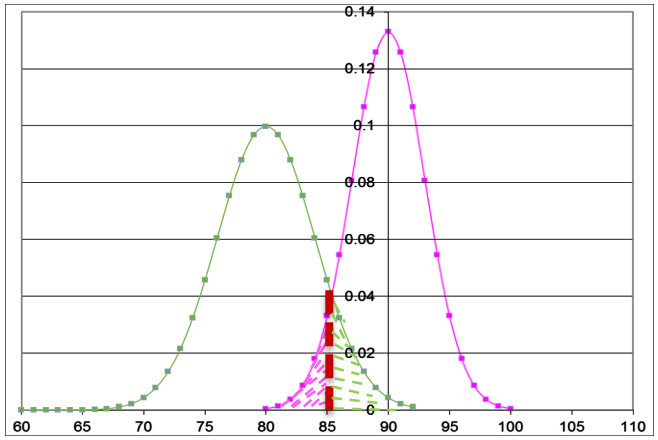
To calculate P($X \geq 85.08$) where $X \sim N(80,16)$, we find

$$z = \frac{85.08 - 80}{\sqrt{16}} = 1.27$$

P(Z $\geq$ 1.27) = 1 – P(Z < 1.27) = 1-0.8980 = 0.102

P(Type II error) = 0.102

The probability of accepting the null hypothesis that 90% are cured when actually 80% are is 10.2%.



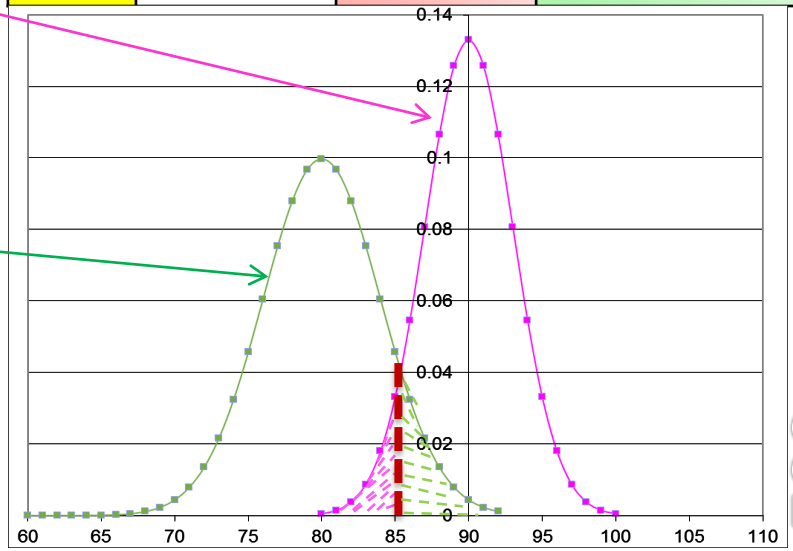
Probability of Getting Type II Error

P(Type II error) = β

To find β (the easy way to remember),

1. On the **NULL** hypothesis distribution, do a *qnorm* to get the critical value corresponding to the selected significance level *(the x-axis value)*.
2. On the **ALTERNATE** hypothesis distribution, do a *pnorm* corresponding to the critical value obtained above to get β *(the area under the curve)*.

		State of Nature	
		Null true	Null false
Action	Fail to reject null (negative)	Correct decision True Negative Specificity	Type II error (β) False Negative
	Reject null (positive)	Type I error (α) False Positive	Correct decision (Power) True Positive Sensitivity/Recall

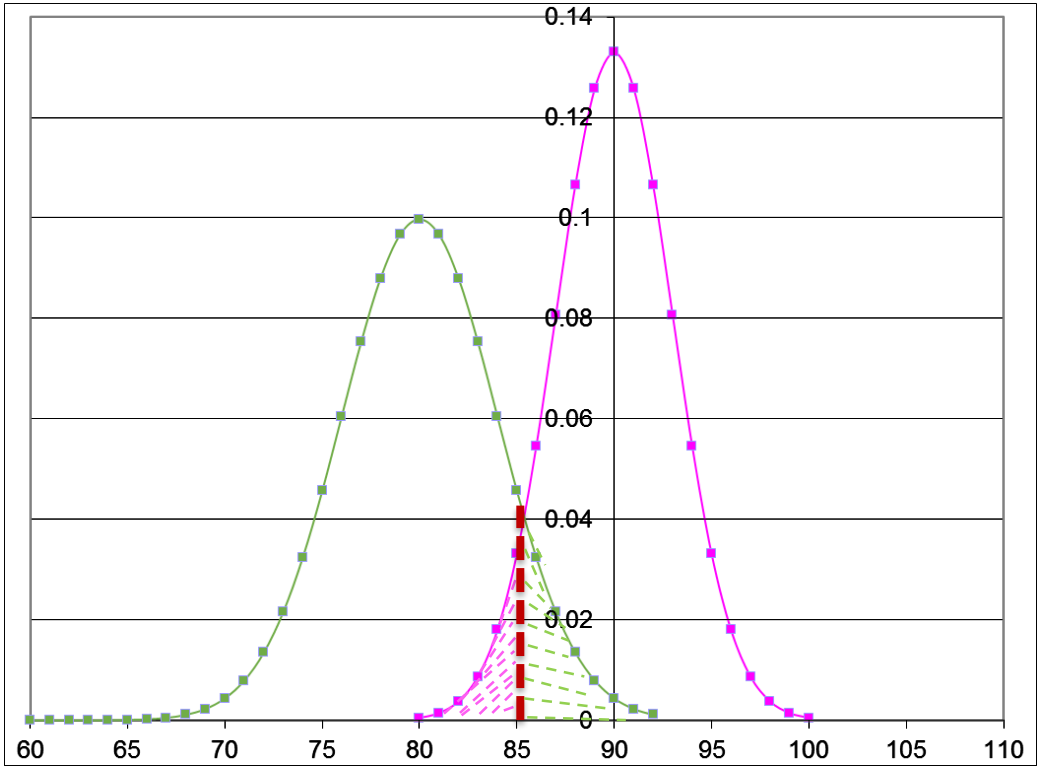


Power of a Hypothesis Test

We **reject null hypothesis correctly** when it is false.

It is actually the opposite of **Type II** error (**we *accept* the null hypothesis incorrectly**), and therefore,

Power = 1 – β = 1-0.102 = 0.898, i.e., the probability that we will make the correct decision in rejecting the null hypothesis is 89.8%.



Hypothesis Testing

A prisoner is on trial and you are on the jury. The jury's task is to assume that the accused is innocent, but if there is enough evidence, the jury needs to convict him.

In the trial, what is the null hypothesis?

The prisoner is innocent (or not guilty).

What is the alternate hypothesis?

The prisoner is guilty.

Hypothesis Testing

What are the possible ways of the jury coming to an incorrect verdict?

If the prisoner is innocent, and the jury gives a ‘guilty’ verdict.
If the prisoner is guilty, and the jury gives an ‘innocent’ verdict.

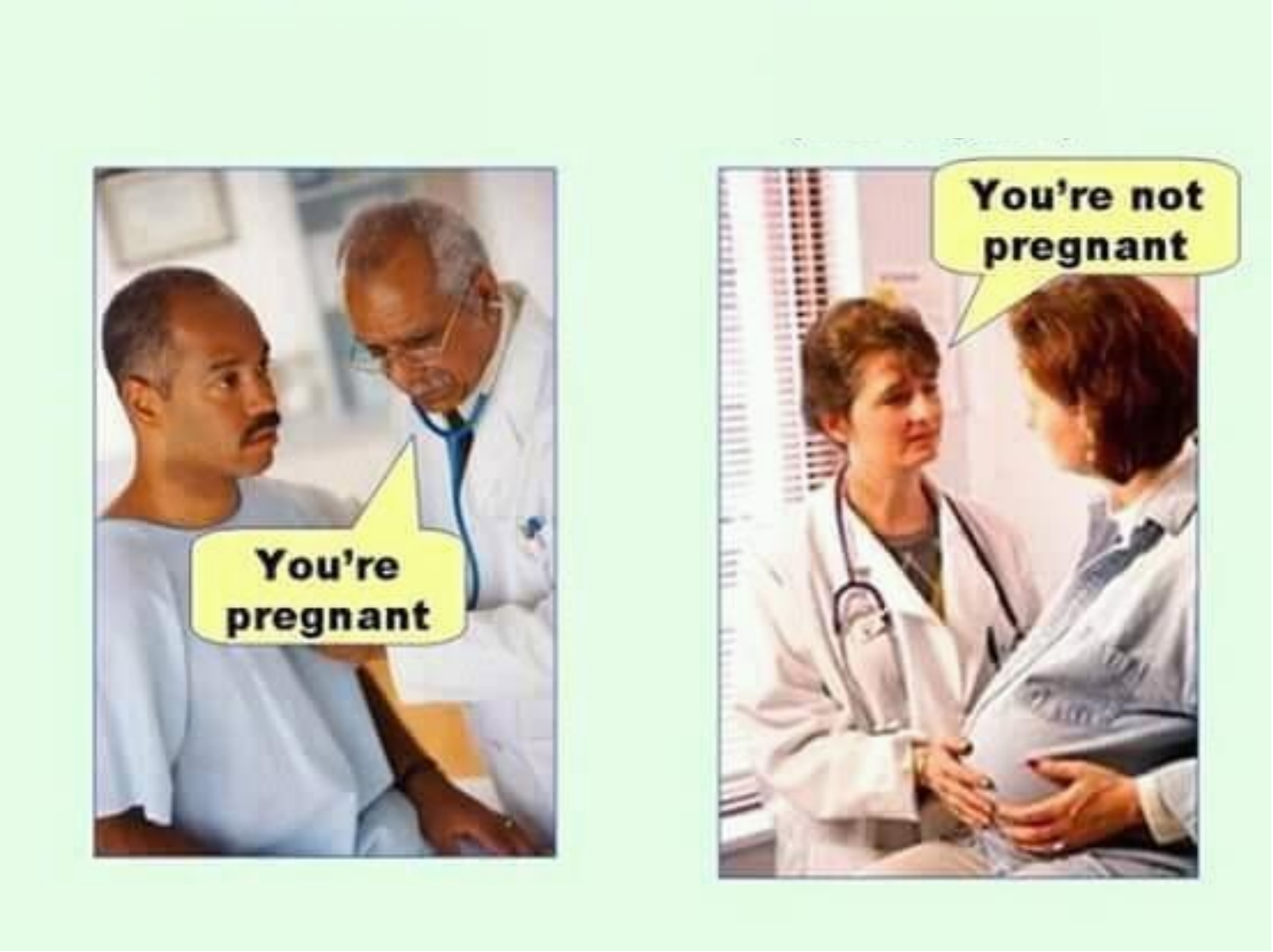
Which one is Type I and which one Type II?

First one is Type I because null hypothesis actually was correct but rejected incorrectly.
Second one is Type II because null hypothesis was false but was accepted incorrectly.

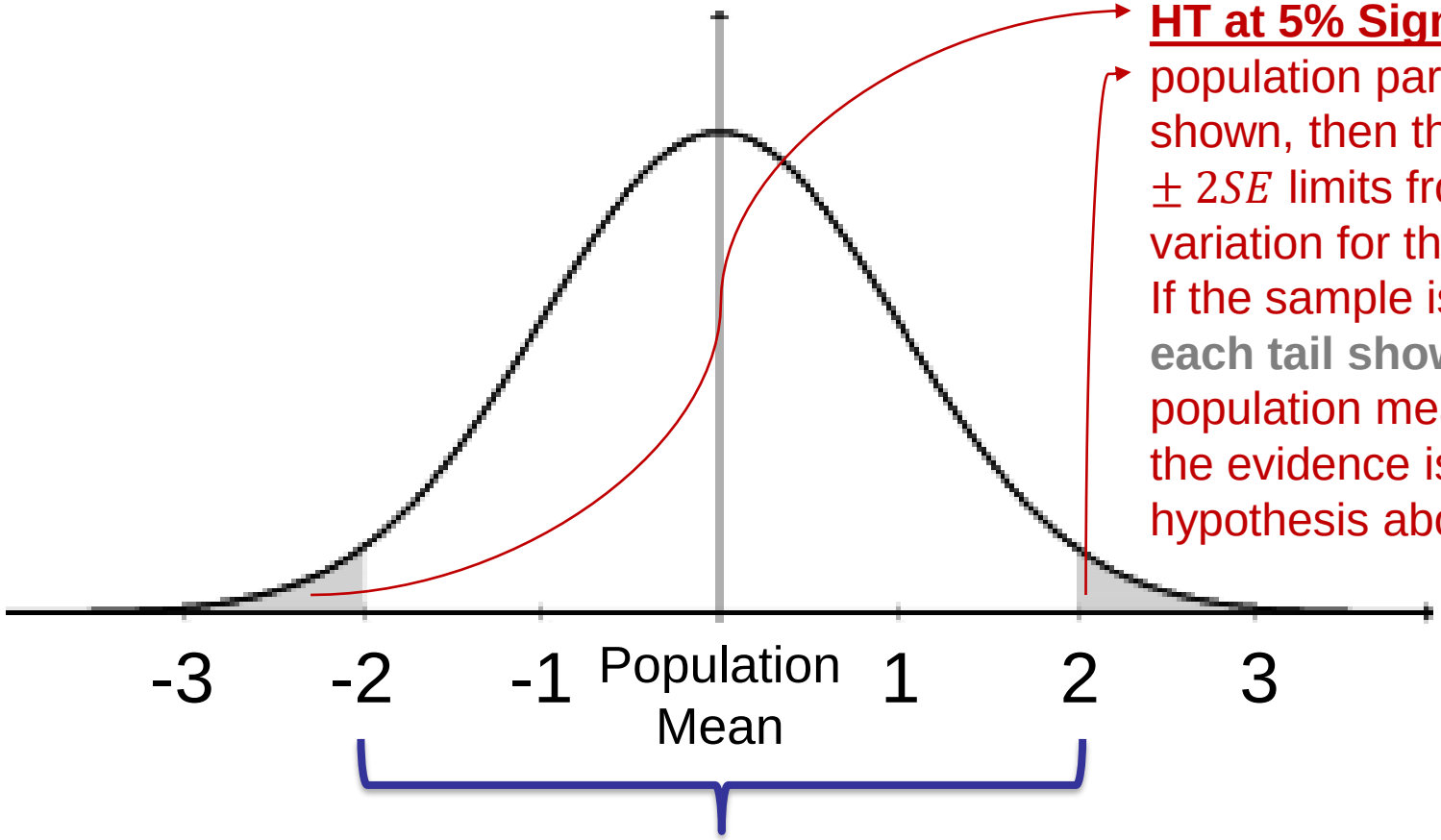
What is the Power of the test?

Since it is opposite of Type II, it will be finding the prisoner guilty when the prisoner is actually guilty, i.e., rejecting the null hypothesis correctly.

Hypothesis Testing



Confidence Intervals and Hypothesis Testing – Two Ways of Inferring the Same



HT at 5% Significance Level: If the true population parameter (e.g., mean) is as shown, then the sample must be within $\pm 2SE$ limits from it (the acceptable normal variation for the null hypothesis to be true). If the sample is in the **5% zone (2.5% in each tail shown in gray)**, then the population mean cannot be as shown (i.e., the evidence is strong to reject the null hypothesis about the population mean).

95% CI: If the true population parameter (e.g., mean) is as shown, then 95% of the samples will contain it within the range $\bar{x} \pm 2SE$. If the sample is in the **5% zone (2.5% in each tail shown in gray)**, then the true population parameter cannot be as shown (i.e., it will not lie in the range $\bar{x} \pm 2SE$.)

Common Test Statistics for Inferential Techniques

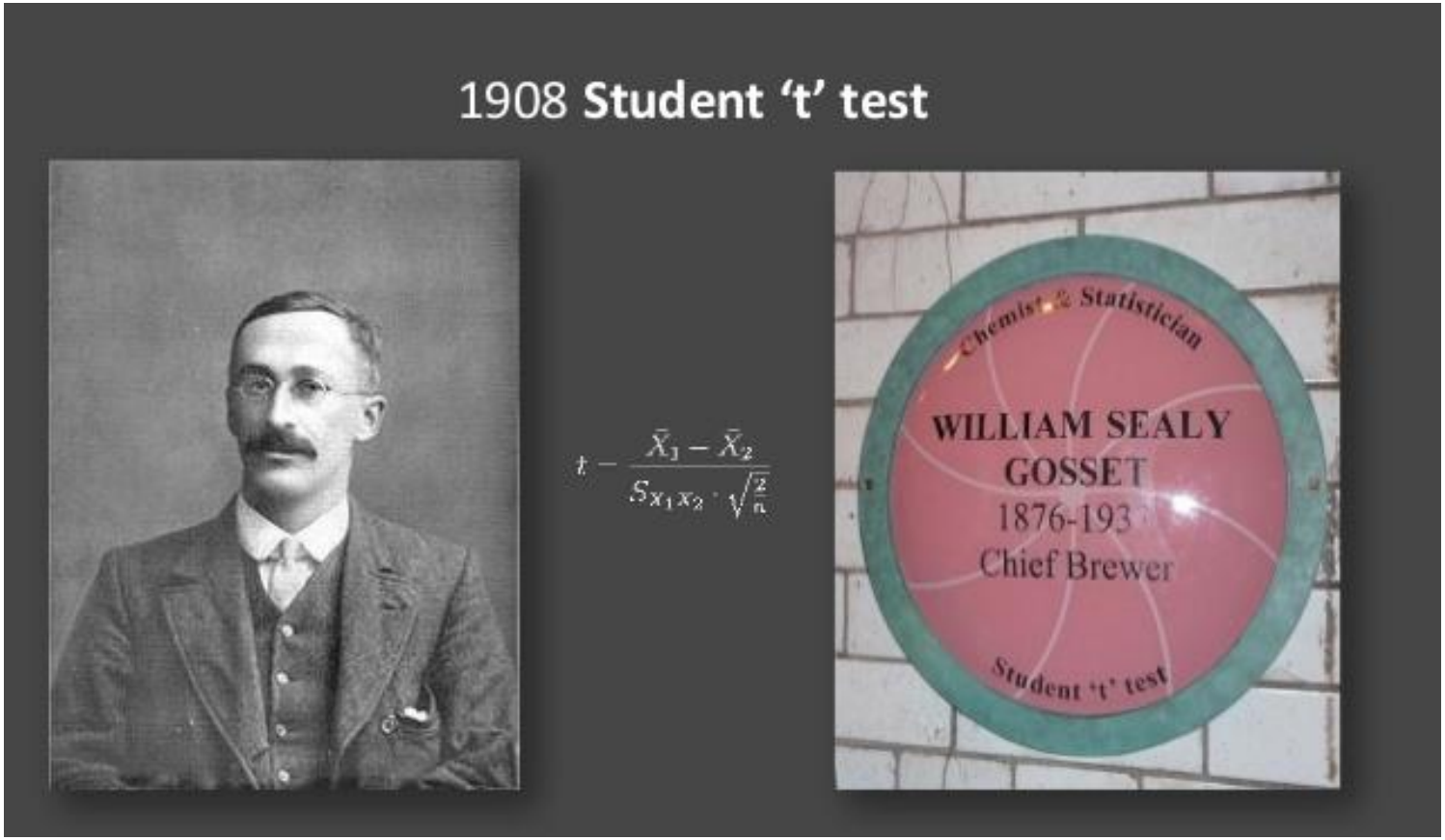
Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistics:

- z
- t
- χ^2 (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

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

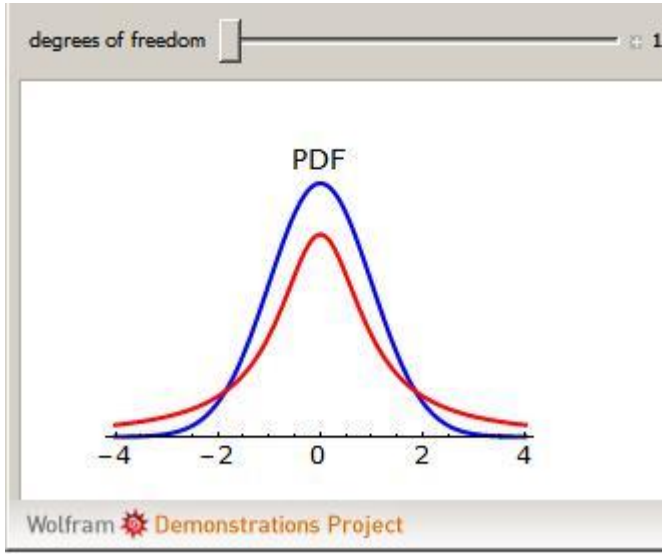
t-Distribution

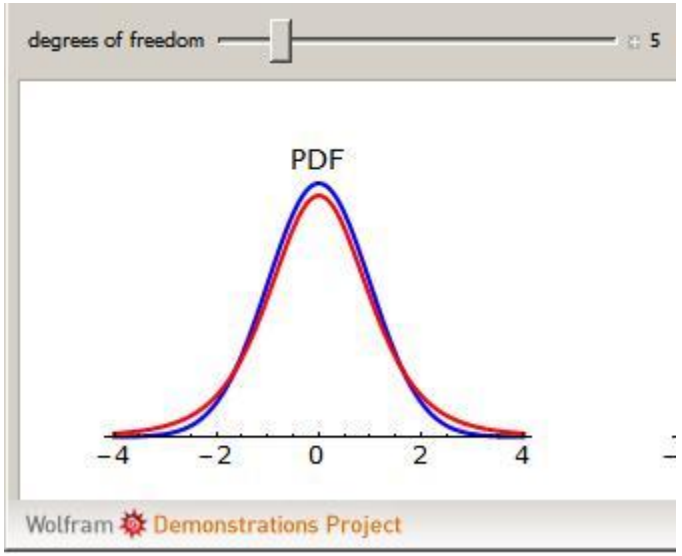


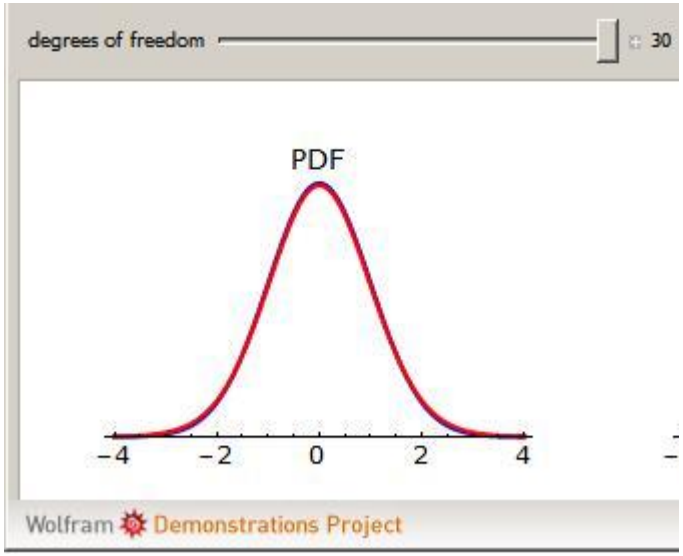
Ref: http://image.slidesharecdn.com/2013-ingenious-ireland-theingeniousirishiet-slideshow-130524065705-phpapp01/95/2013-ingeniousirelandthe-ingenious-irishietslideshow-43-638.jpg?cb=1369825611
Last accessed: October 31, 2015

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.







Ref: "Comparing Normal and Student's t-Distributions" from the Wolfram Demonstrations Project
http://demonstrations.wolfram.com/ComparingNormalAndStudentsTDistributions/
Contributed by: Gary McClelland; Last accessed: August 11, 2017

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*.

* Roger E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, California: Brooks/Cole, 1968.

Properties of *t*-Distribution

- Mean of the distribution = 0
- Variance = $\frac{\nu}{\nu-2}$, where $\nu > 2$
- Variance is always greater than 1, although it is close to 1 when there are many degrees of freedom (sample size is large)
- With infinite degrees of freedom, *t* distribution is the same as the standard normal distribution

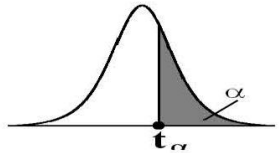
Confidence Interval to Estimate μ

- Population standard deviation UNKNOWN and the population normally distributed.
- $\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ Recall: $t\ statistic\ (or\ t\ score), t = \frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}}$
  - 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 (α/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

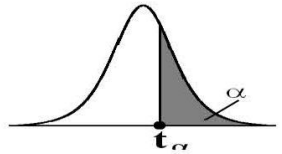
n = 10

$\nu = 10 - 1 = 9$

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

t-table

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



	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

$t_{9,0.025} = 2.262$

t-Distribution - Example

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

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

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

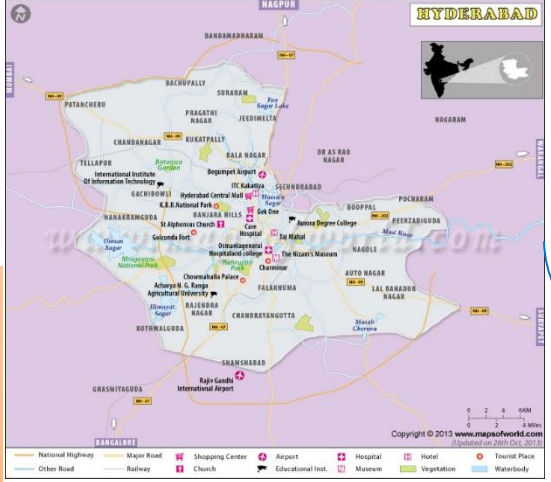
$$101.24 - 2.262 * \frac{2.48}{\sqrt{10}} \leq \mu \leq 101.24 + 2.262 * \frac{2.48}{\sqrt{10}}$$

$$99.47 \leq \mu \leq 103.01$$

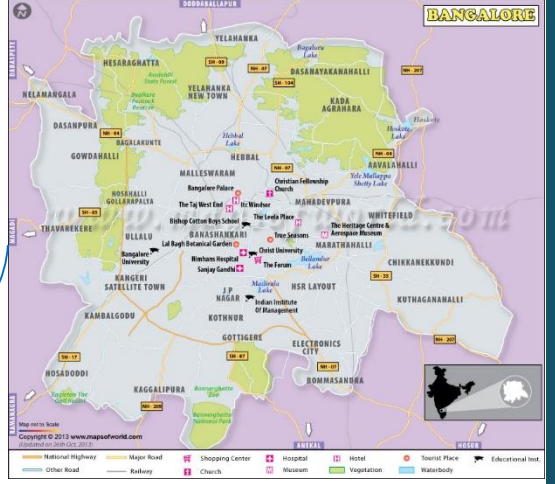
The batch mean is 101.24 mg with an error of +/-1.77 mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.47 mg and 103.01 mg.

t-Distribution – Example – R

```
        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
```



Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering