Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Sampling Distribution of Means, CLT, Confidence Intervals, Hypothesis Testing, t-Distribution

Dr. Sridhar Pappu

Executive VP – Academics, INSOFE

June 17, 2018

Central Limit Theorem (CLT)

SAMPLING DISTRIBUTION OF MEANS

Sampling Distribution of the Means

- The sampling distribution of means is what you get if you consider all possible samples of size *n* taken from the same population and form a distribution of their means.

- Each randomly selected sample is an independent observation.

Central Limit Theorem

- http://onlinestatbook.com/2/sampling_distributions/clt_demo.html
- As sample size goes large and number of buckets are high, the means will follow a normal distribution with same mean (μ) and $\frac{1}{n}$ of variance (σ^2).

Expectation and Variance for $\bar{X}$

$E(\bar{X}) = \mu$

Mean of all sample means of size *n* is the mean of the population.

$Var(\bar{X}) = \frac{\sigma^2}{n}$

Standard deviation of $\bar{X}$ tells how far away from the population mean the sample mean is likely to be. It is called the **Standard Error of the Mean** and is given by

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

$\bar{X}$ ~ N(μ, σ^2/n)

When an Attribute is Not Normal

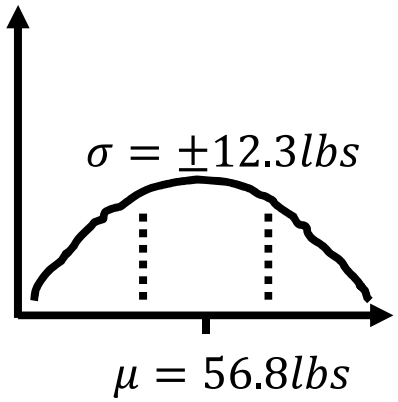
- Let us assume it is a sample from infinite data
- So, if we take many such samples of large sample size (>30 as a thumb rule), the mean values, $\bar{x}$, will be hovering close to the population mean, μ, with a standard deviation, $s = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and *n* is the sample size.

Using the Central Limit Theorem

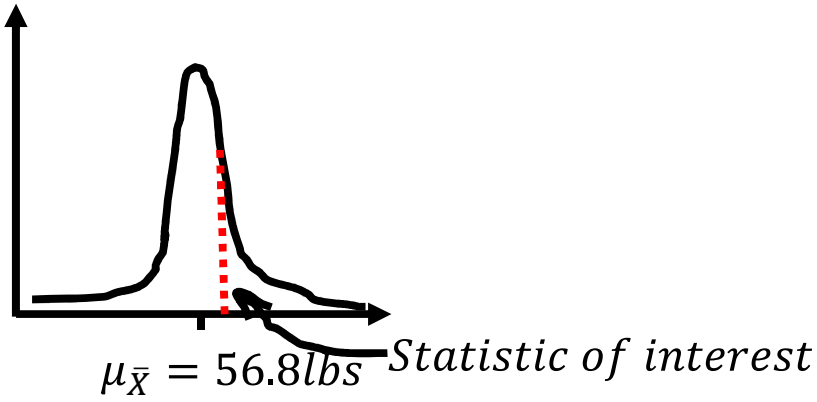
The Aluminum Association of America reports that the average American household uses 56.8 lbs of aluminium in a year. A random sample of 51 households is monitored for one year to determine aluminium usage. If the population standard deviation of annual usage is 12.3 lbs, what is the probability that the sample mean will be $> 60\ lbs$?

Sampling Distribution

$Population\ distribution$

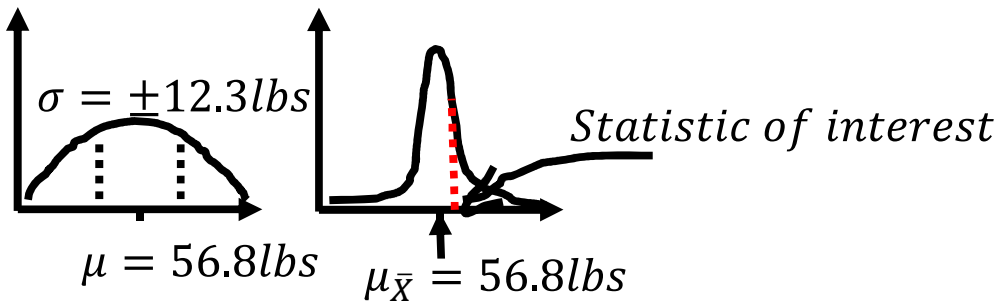


$Sampling\ distribution\ of\ sample\ mean\ when\ n = 51$



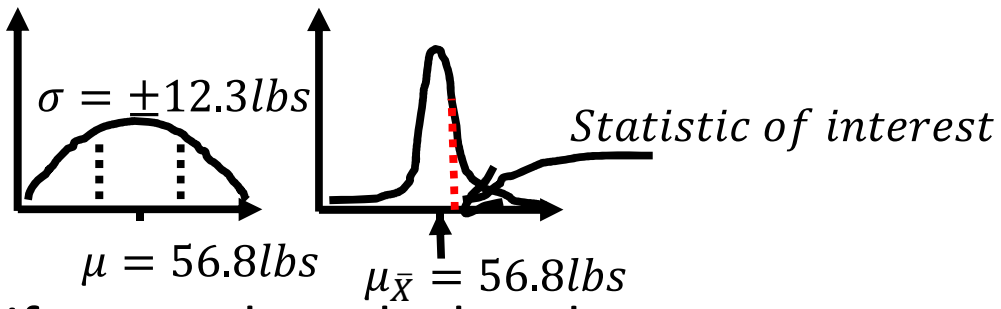
- Step 1: List all known parameters and values
- Step 2: Calculate others, or estimate if cannot be calculated
- Step 3: Find probabilities using tables, Excel or R

Sampling Distribution



- Step 1: List all known parameters and values
  - Population mean, $\mu = 56.8\ lbs$
  - Population standard deviation, $\sigma = 12.3\ lbs$
  - Sample size, $n = 51$
  - Sample mean, $\bar{x} > 60\ lbs$
  - Mean of sample means, $\mu_{\bar{X}} = \mu = 56.8\ lbs$

Sampling Distribution



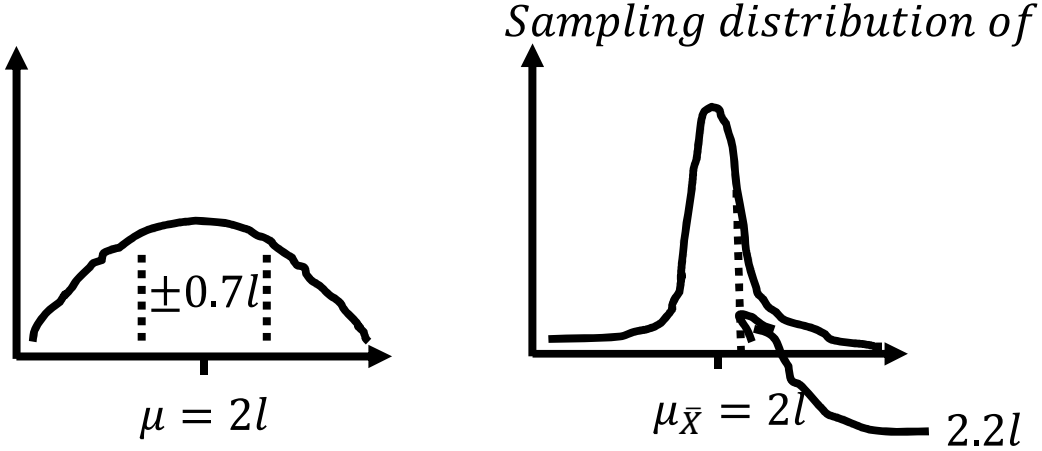
- Step 2: Calculate others, or estimate if cannot be calculated
  - Standard deviation of sample means, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12.3}{\sqrt{51}} = 1.72$
  - $\therefore z = \frac{60 - 5\ .8}{1.72} = 1.86$
- Step 3: Find probabilities using tables, Excel or R
  - Excel: 1-NORM.S.DIST(z,TRUE) = 0.0316
  - R: 1-pnorm(60, 56.8, 12.3/sqrt(51)) = 0.0316 or 1-pnorm(1.86, 0, 1) = 0.0316
  - Please calculate these for:
    - $> 58\ lbs$
    - $> 56\ lbs < 57\ lbs$
    - $< 50\ lbs$

Sampling Distribution

The average male drinks $2l$ of water when active outdoors with a standard deviation of $0.7l$. You are planning a trip for 50 men and bring $110l$ of water. What is the probability that you will run out of water?

$\mu = 2, \sigma = 0.7$

$P(run\ out) \Rightarrow P(use > 110l) \Rightarrow P(average\ water\ use\ per\ male > 2.2l)$



$$1 - pnorm\left(2.2, 2, \frac{0.7}{\sqrt{50}}\right)$$

$$= 0.0217, i.e., 2.17\%$$

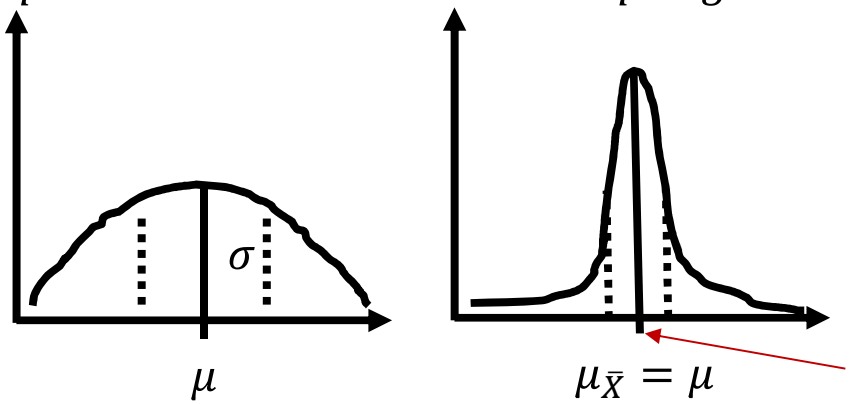
Source: https://www.khanacademy.org/math/probability/statistics-inferential/sampling_distribution/v/sampling-distribution-example-problem
Last accessed: May 9, 2014

Sampling Distribution

You sample 36 apples from your farm's harvest of 200,000 apples. The mean weight of the sample is 112g with a 40g sample standard deviation. What is the probability that the mean weight of all 200,000 apples is between 100 and 124g?

Sampling Distribution

$Population\ distribution$ $Sampling\ distribution\ of\ sample\ mean\ when\ n = 36$



$$\sigma_{\bar{X}}{}^2 = \frac{\sigma^2}{n} = \frac{\sigma^2}{36} \Rightarrow \sigma_{\bar{X}} = \frac{\sigma}{6}$$

What are we trying to find out?

Is this between 100g and 124g?

We need to know if population mean, μ, is within $\pm$12g of the sample mean, $\bar{x}$ *(112g)*, i.e., $\mu = \bar{x} \pm 12$. As $\bar{x}$ can be anywhere in the distribution, we don't have a fixed reference.

But, this is the same as saying that we need to know if sample mean, $\bar{x}$, is within $\pm$12g of the population mean, μ, i.e., $\bar{x} = \mu \mp 12$

Since $\mu = \mu_{\bar{X}}$, we can now use the sampling distribution of the means.

Source: https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1
Last accessed: May 9, 2014

Sampling Distribution

$Population\ distribution$ $Sampling\ distribution\ of\ sample\ mean\ when\ n = 36$

We need to find out how many standard deviations away from $\mu_{\bar{X}}$ is 12g. But, we don't know $\sigma_{\bar{X}}$ because we don't know σ. We use the sample standard deviation, *s* (40g), as the best estimate of population standard deviation. $\sigma \approx s = \pm 40g.\ \therefore \sigma_{\bar{X}} = \frac{\sigma}{6} = \frac{40}{6} = 6.67.$ So 12g is 12/6.67 = 1.8 standard deviations.

The z-table gives the probability as 0.9641 but that is the entire region below +1.8 z.

Find the region between -1.8 and +1.8 z.

0.9282. *Rcode: pnorm(1.8,0,1)-pnorm(-1.8,0,1)*

Source: https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-1
Last accessed: May 9, 2014

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%

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

TOTAL SEATS 182 | HALF WAY MARK 92

GUJARAT

BIG WIN

Rupani retains Rajkot West by margin of 54,300 votes

BJP	Cong	NCP	01
99	77	BTP	02
		IND.	03

Exit Polls

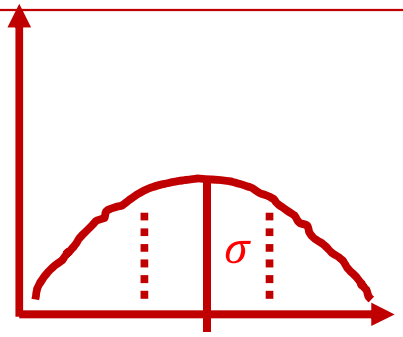
Final Result

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.

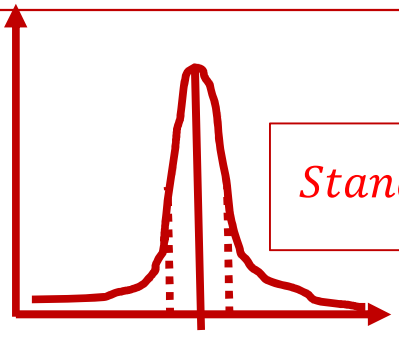
Polling Organisation	NDA	UPA	Other
CNN-IBN – CSDS – Lokniti	276 (±6)	97 (±5)	148 (±23)
India Today – Cicero	272 (±11)	115 (±5)	156 (±6)
News 24 – Chanakya	340 (±14)	70 (±9)	133 (±11)

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



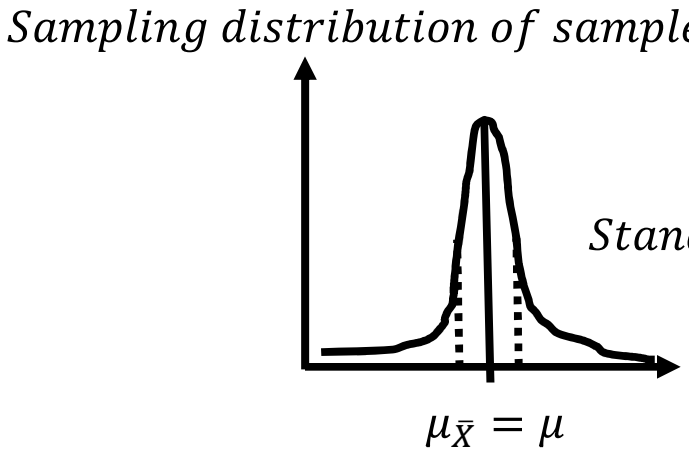
$$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$$

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.

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 and the apples example of Central Limit Theorem)*.

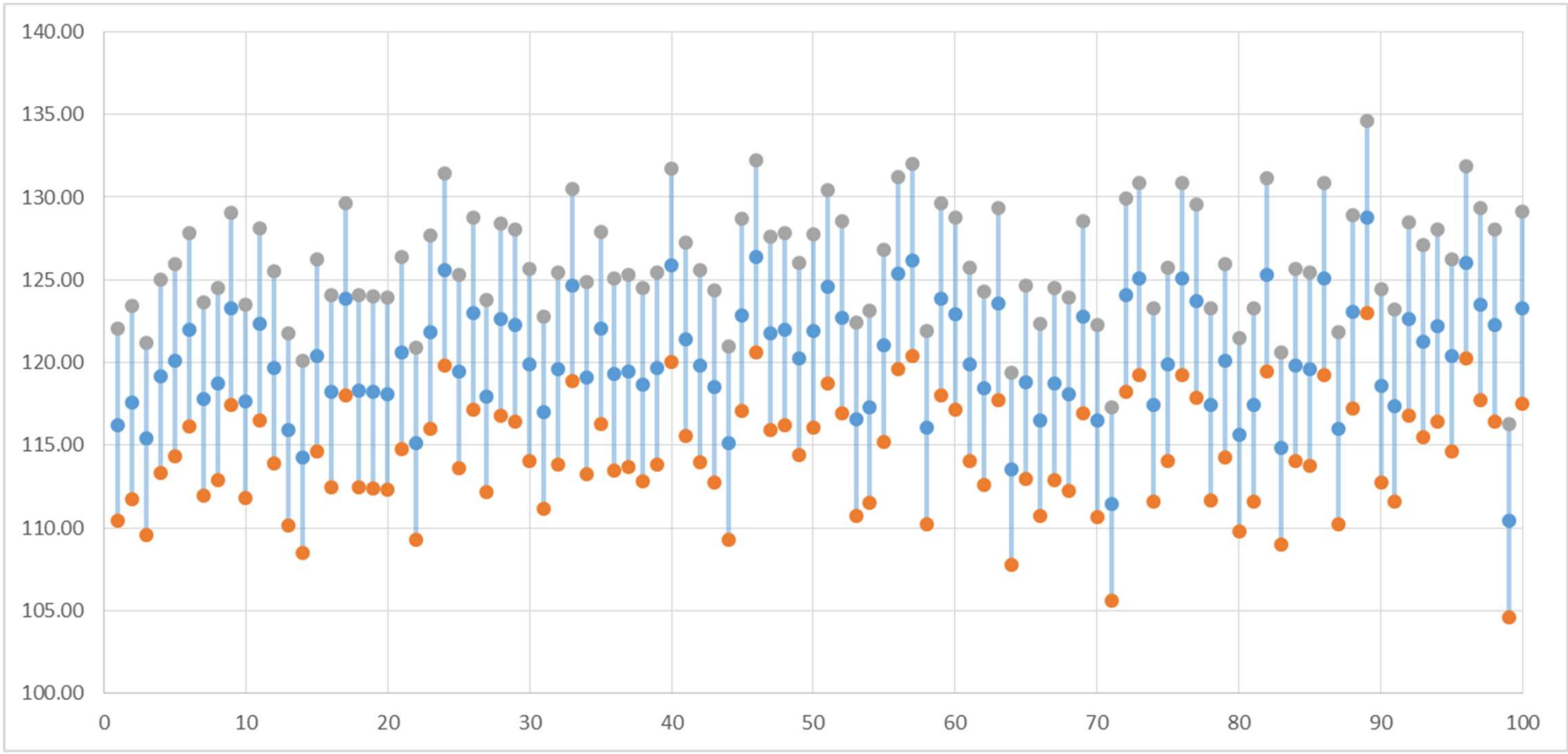
Alternatively, 95% of such intervals include the population mean.

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



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

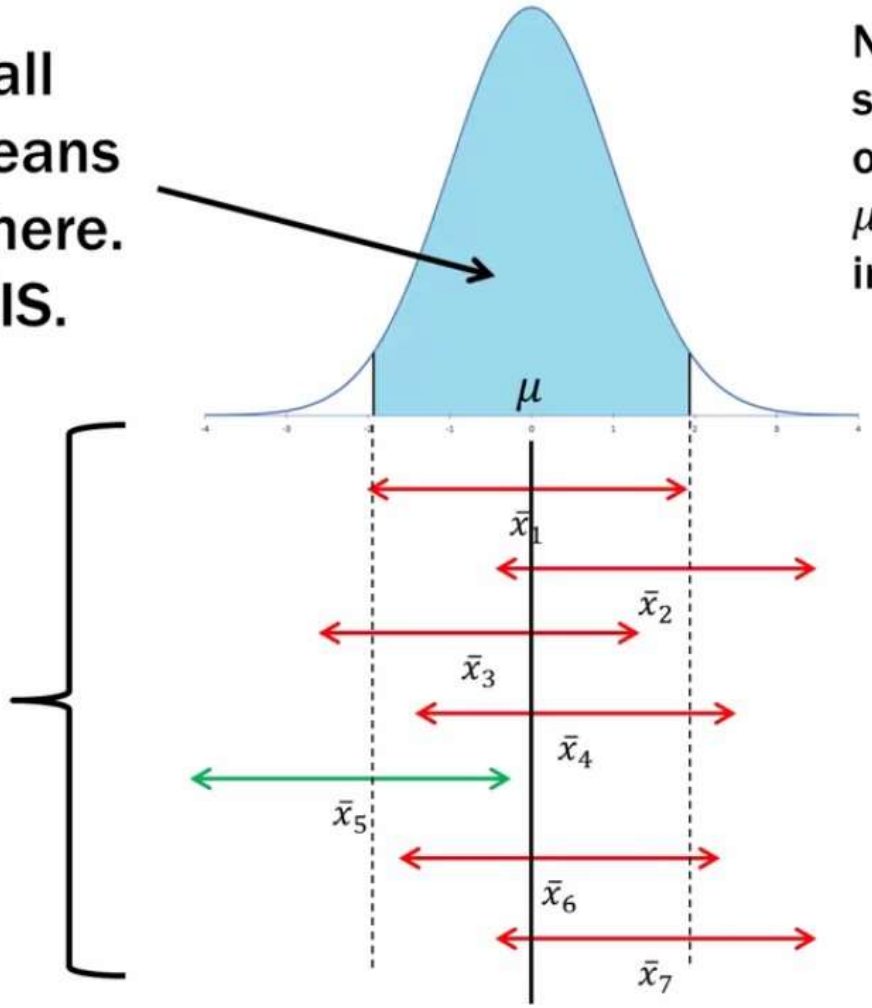
Confidence Level and Interval - Excel



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?

Sampling, Sample Size, MoE, Confidence Intervals

Times Now survey gives Congress 91, BJP 89 seats

TIMES NEWS NETWORK

New Delhi: Karnataka is heading for an absolute photo-finish, with the Congress and the BJP locked in a neck-and-neck race, a Times Now-VMR survey in the poll-bound state has revealed.

▶ Siddaramaiah preferred CM, P2

As things look now, HD Kumaraswamy, JD(S) state president, could well play the role of the kingmaker. The ruling Congress is likely to win 91 of the 224 assembly segments, with the BJP notching up 89 seats. The JD(S) may end up with a tally of 40.

If polls were held now, neither CM Siddaramaiah nor BS Yeddyurappa of the BJP is in a position to form a government as their parties are well short of the 113 seat mark.

The survey, conducted by VMR for Times Now, was done between April 4 and 16. Over, 4,000 respondents were surveyed through stratified random sampling spread across all the six regions of the state. The constituencies were carefully chosen to be representative of the behaviour of the each region and, thus, the entire state. The survey has a margin of error of 3% at the estimation of votes with a 95% confidence level.

WITH 40 SEATS, JD(S) TO PLAY KINGMAKER

PARTY	2018		2013		SWING	
	Seat share	% of vote share	Seat share	% of vote share	No. of seats	% of votes
CONGRESS	91	38.6	122	36.59	-31	+2.01
BJP	89	35.03	40	19.89	+49	+15.14
JD(S)+ BSP	40	21.33	40	20.19	0	+1.14
OTHERS	4	5.04	22	23.33	-18	-18.29
TOTAL	224		224			

Source: Times Now-VMR Survey

Source: The Times of India, Bengaluru Edition, April 24, 2018
Last accessed: April 28, 2018

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%

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 electr
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
brar
five
po

Skymet hints at poor monsoon

J. UMAMAHESHWAR RAO | DC
VISAKHAPATNAM, MARCH 27

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.

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.

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.

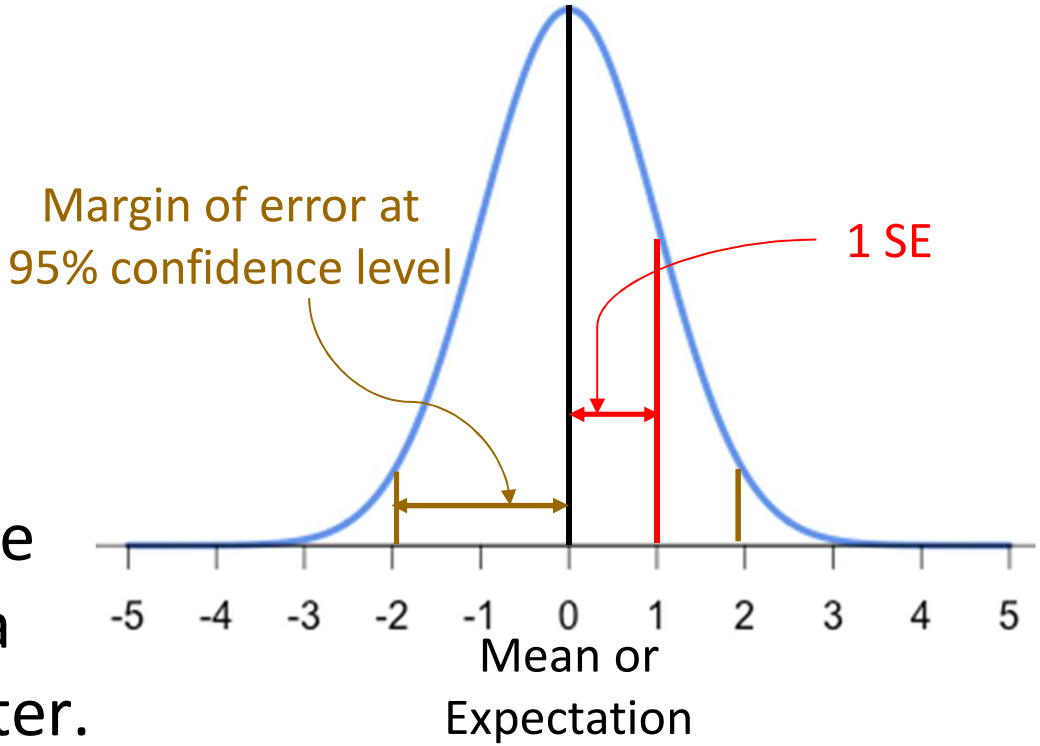
2016 witnessed normal rainfall across the country, barring deficient precipitation in states like Karnataka, Kerala, Andhra Pradesh, Tamil Nadu, and the northeastern states.

SE, Margin of Error, Confidence Interval and Sample Size

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

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

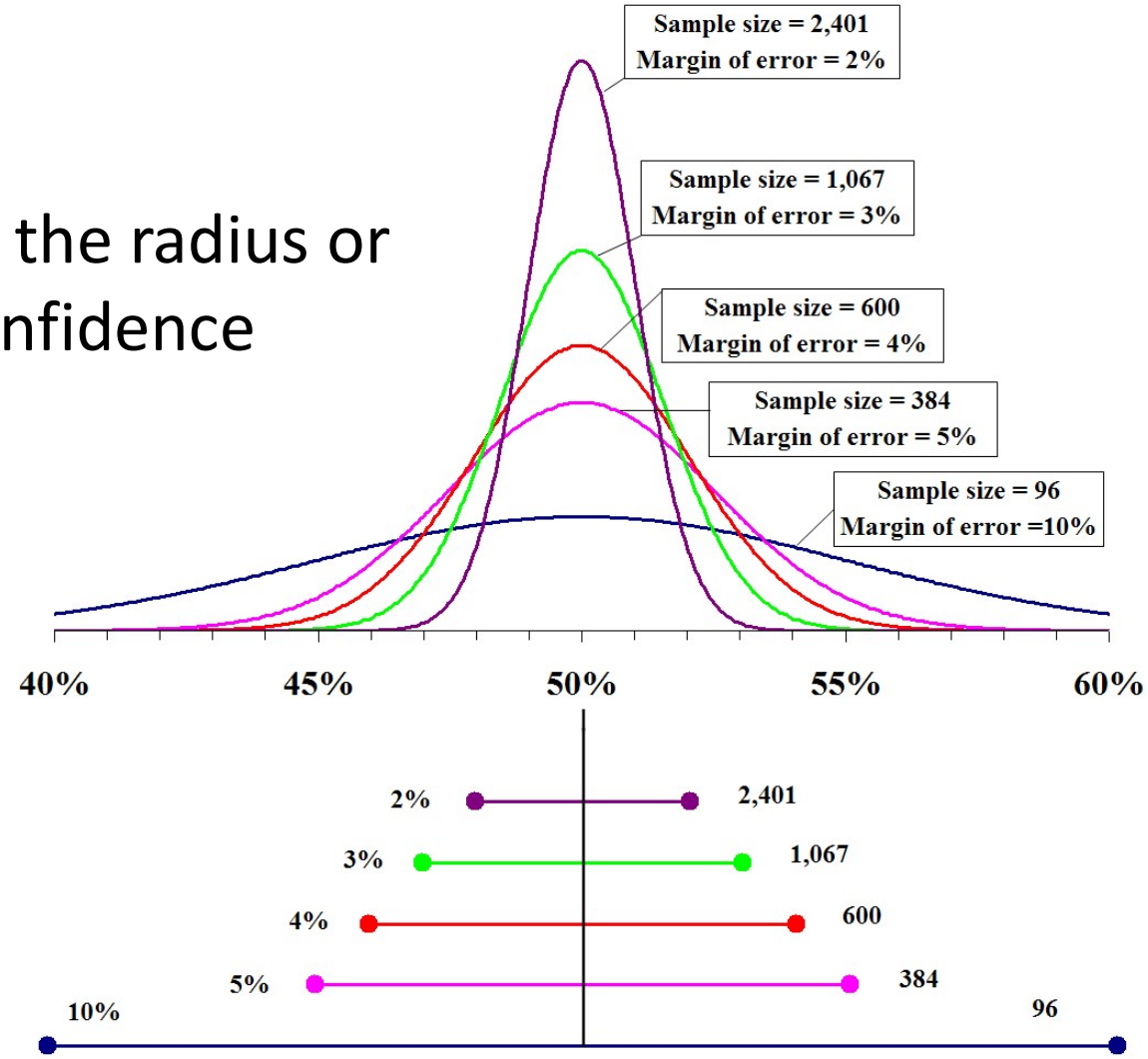
Margin of error is the **maximum expected difference** between the true population parameter and a sample estimate of that parameter.



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


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

he says – or more importantly, doesn't 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,

status for Pakistan, to

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 = $\sqrt{\frac{pq}{n}}$

If you are interested in the derivation (optional):

Recall Binomial distribution

E(X) = np

Var(X) = npq

Proportion, p = X/n

*So, E(p) = E(X/n) = 1/n*E(X) = I/n*np = p*

*Var(p) = Var(X/n) = $1/n^2$*Var(X) = $1/n^2$*npq = 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?

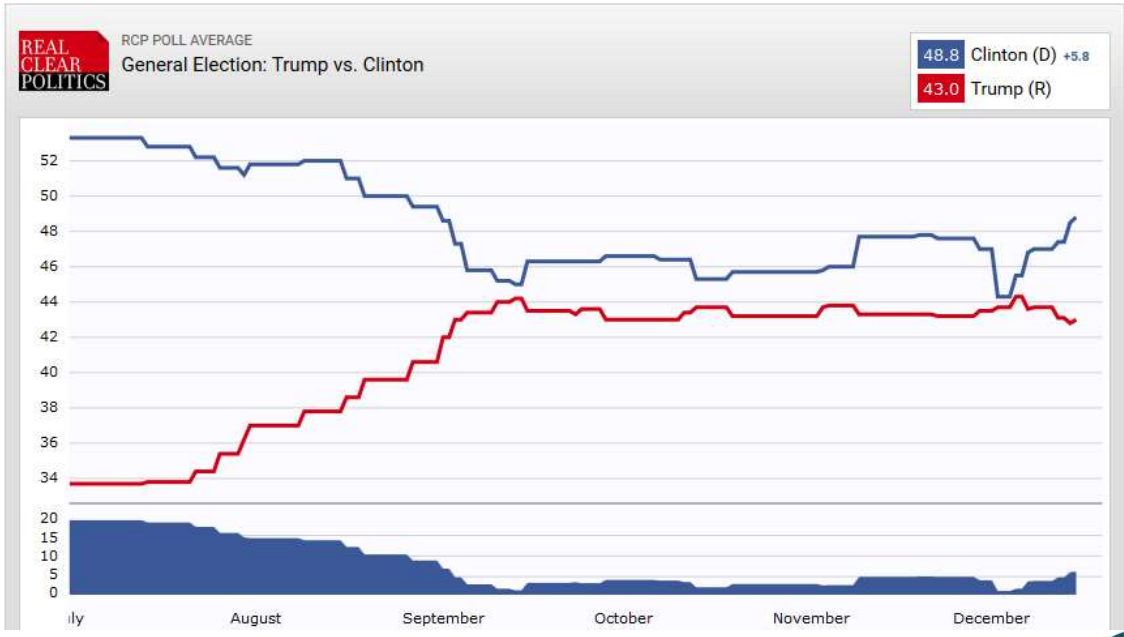
RealClear Politics

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) in independent variables
  - Multiply with number classes (assume 10 levels for numeric) in the dependent variable
  - 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

Levels of the Independent Variables	Classes (Levels of the Dependent Variable)	Rule of Thumb (x75)
4	2	600
25	3	5625
50	4	15000
100	2	15000
250	3	56250
500	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

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

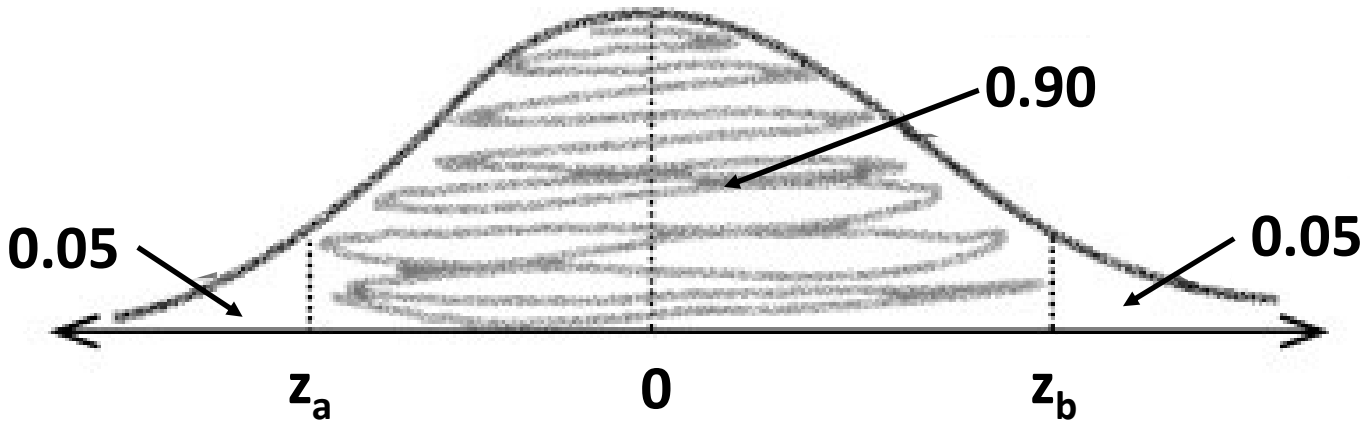
$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

$\therefore$ 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 $P(Z > z_b) = 0.05$

We get $z_a = -1.645$ and $z_b = 1.645$. How?

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.

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

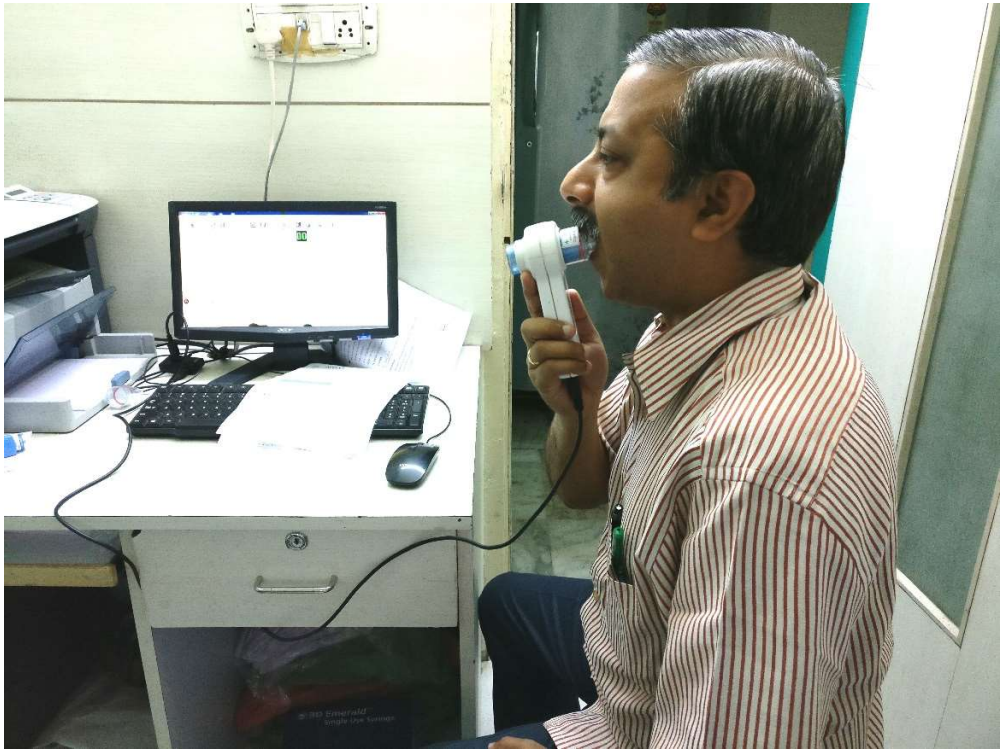
$$0.09 < p < 0.41$$

Shortcuts for Calculating Confidence Intervals

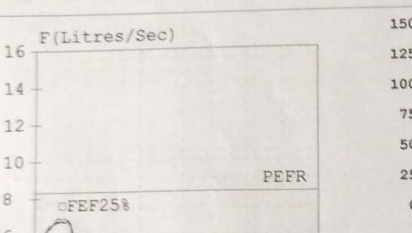
The lung function in 57 people is tested using FEV1 (Forced Expiratory Volume in 1 Second) measurements. The mean FEV1 value for this sample is 4.062 litres and standard deviation, *s* is 0.67 litres. Construct the 95% Confidence Interval.

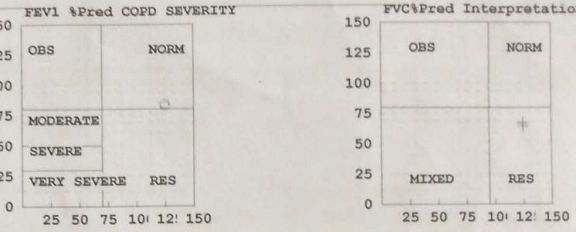
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

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

Normal Spirometry

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

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

FEV1 values of 57 male medical students

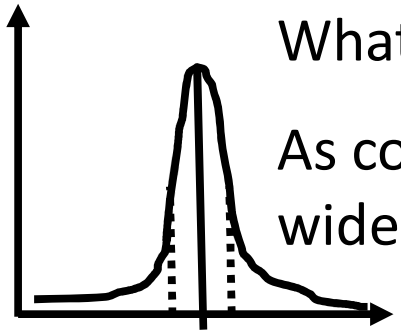
Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

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?

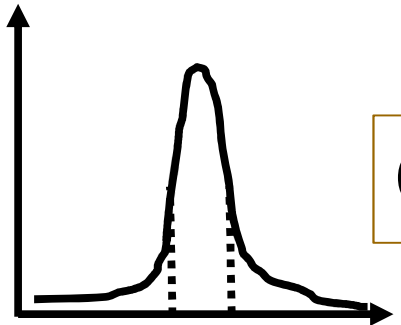
As confidence level increases, the confidence interval becomes wider and *vice-versa*.

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

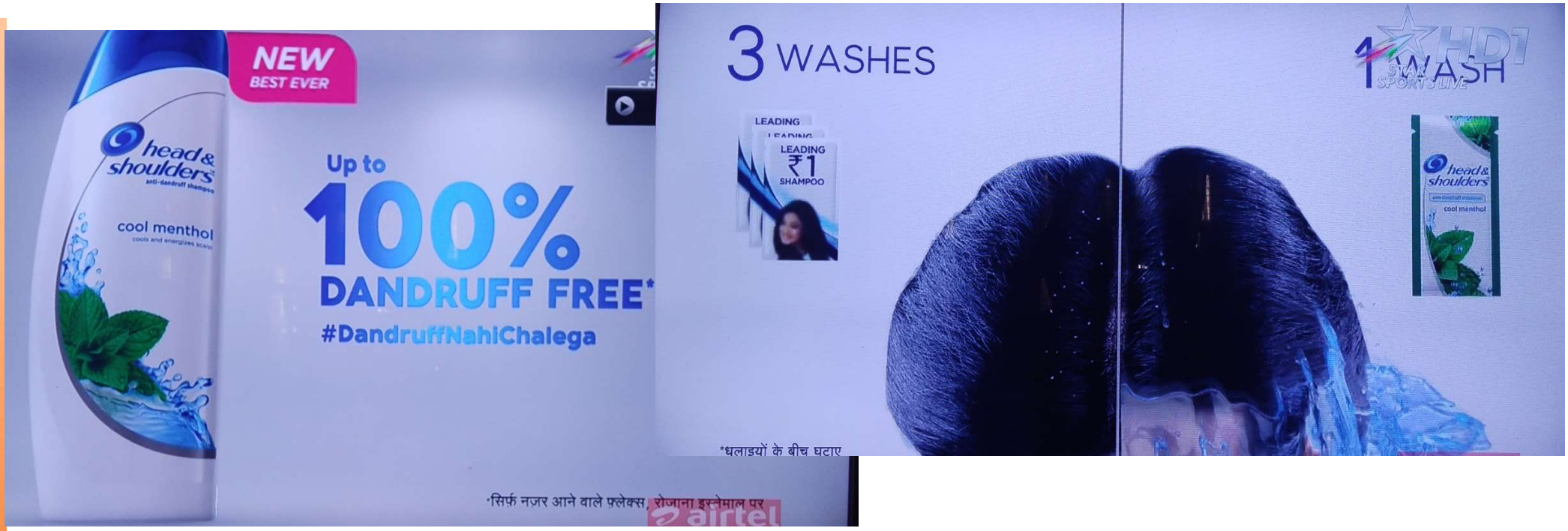
Q. What is the probability of getting 15 or more heads?

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)*

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.



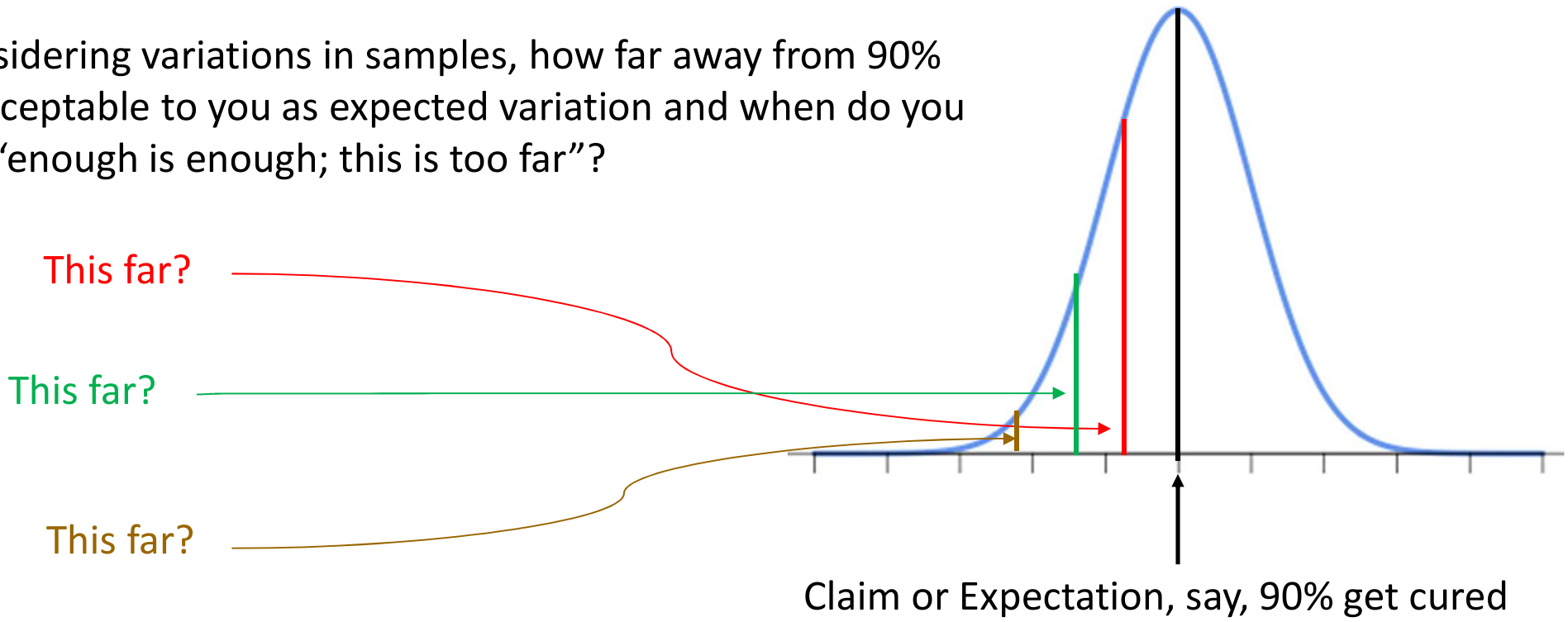
Hypothesis tests give a way of using samples to test whether or not statistical claims are likely to be true or not.

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

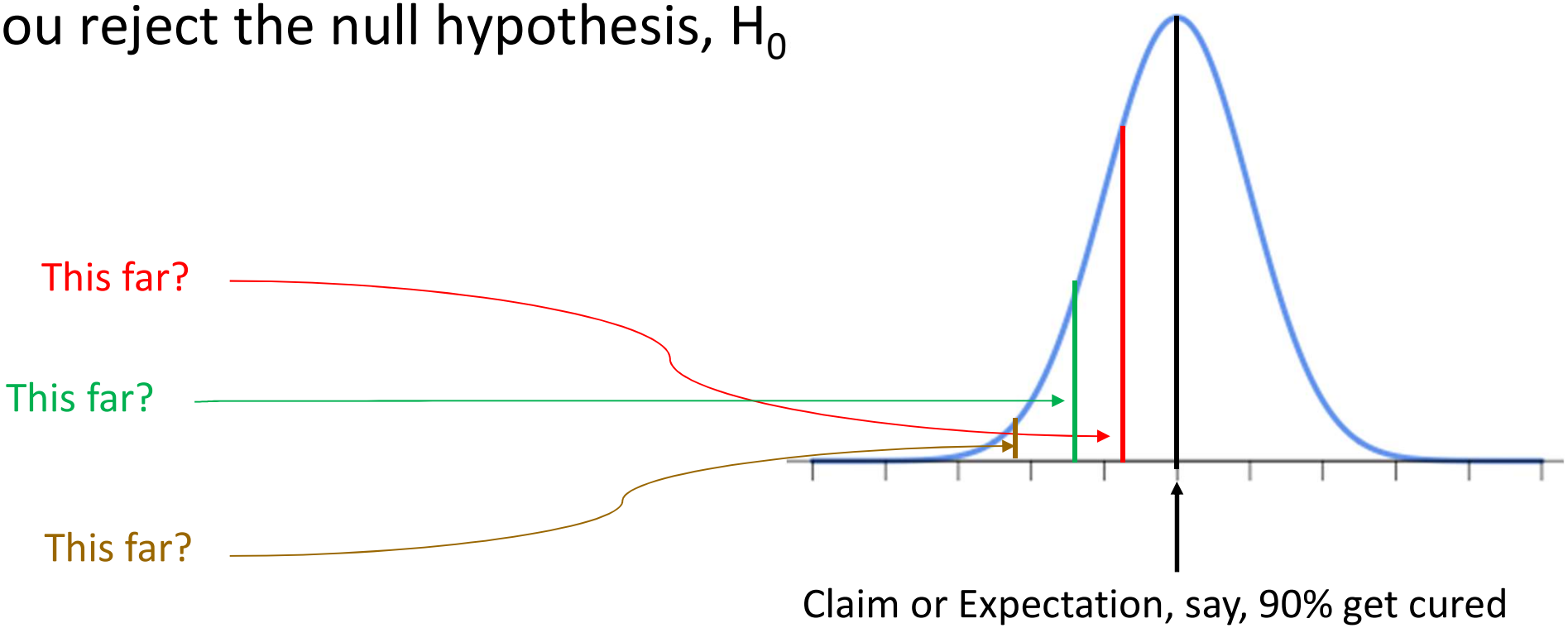
- Two hypotheses in competition:
  - H_0: The NULL hypothesis, usually the most conservative.
  - H_1 or H_A: The ALTERNATIVE hypothesis, the one we are actually interested in.
- Examples of NULL Hypothesis:
  - The coin is fair
  - The new drug is no better (or worse) than the placebo
- Examples of 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 $\alpha = 5\%$.

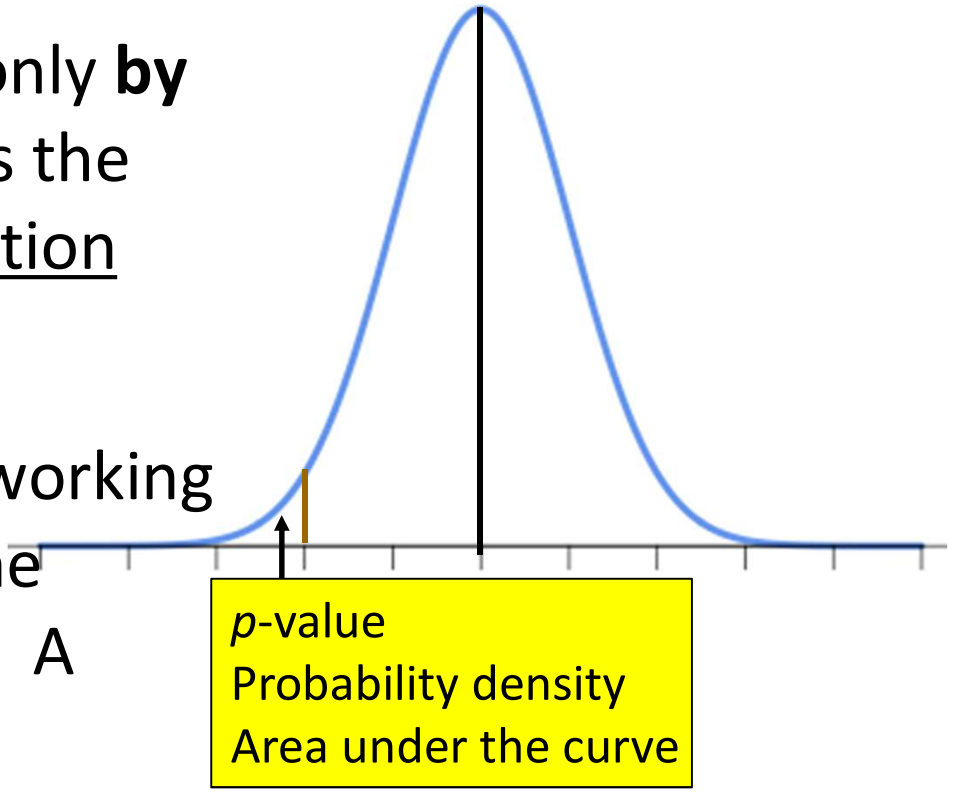
Recall that in a 95% CI, there is a 5% chance that the sample will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 90% snorers are cured, is rejected.

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

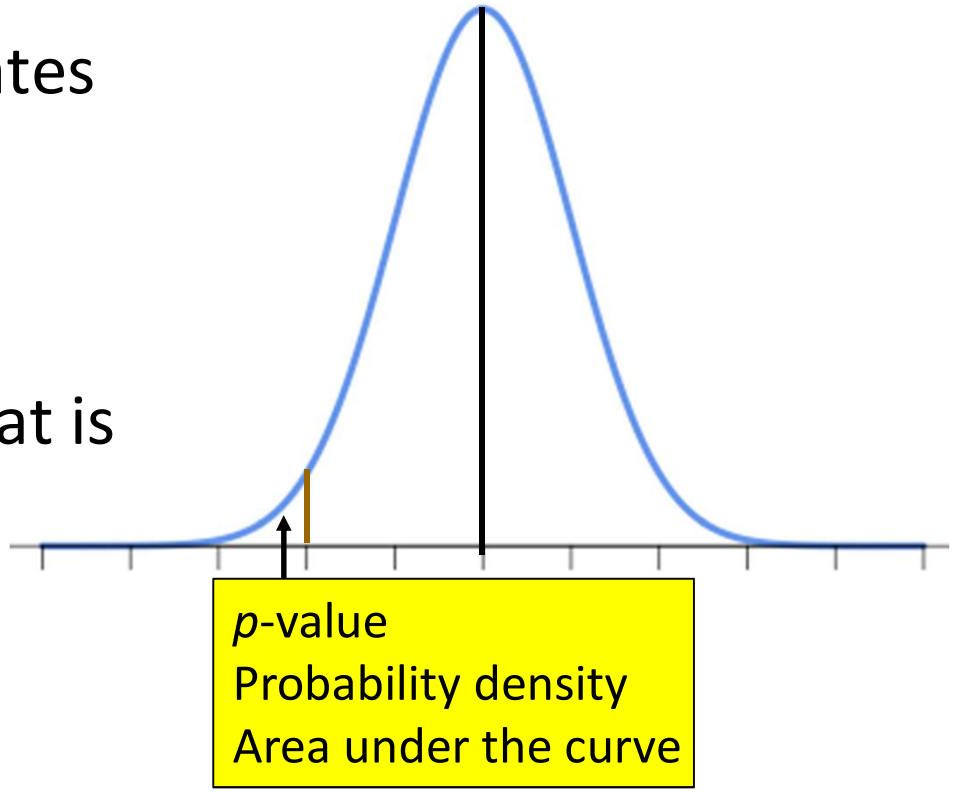
p-value is the probability of getting only **by chance** a value <u>at least</u> as extreme as the one in the sample <u>under the assumption that the null hypothesis is true</u>.

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.

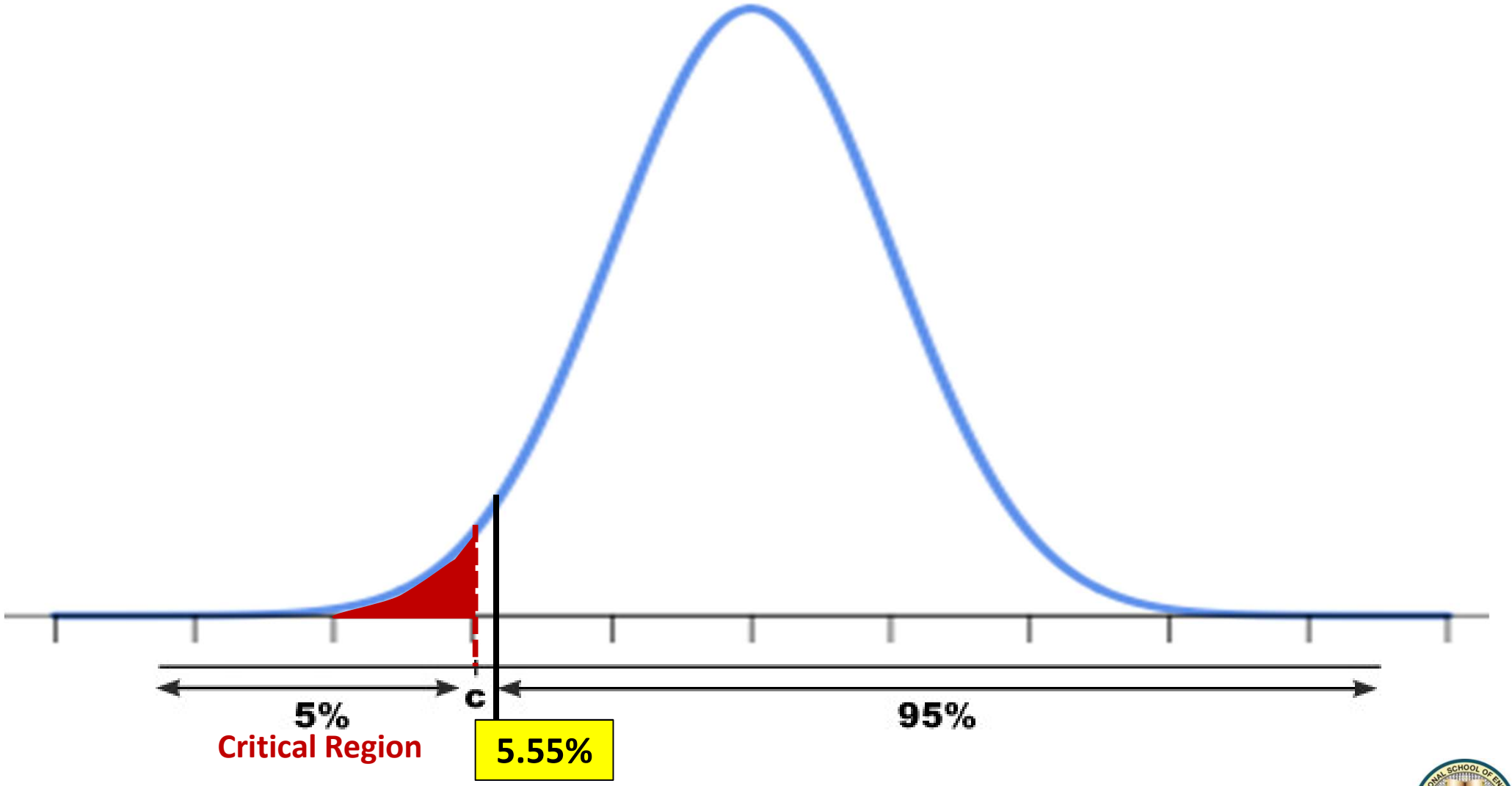
Step 5: Find the *p*-value

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). *Also, can be approximated to X ~ N(90,9). Why?*

What is the probability of 80 or fewer getting cured?

p = 0.002

Rcode: pbinom(80, 100, 0.9, lower.tail = TRUE, log.p = FALSE)

What is your decision?

Since the *p*-value (0.002) is less than the Significance Level of 0.05, the null hypothesis can be rejected.

Attention Check

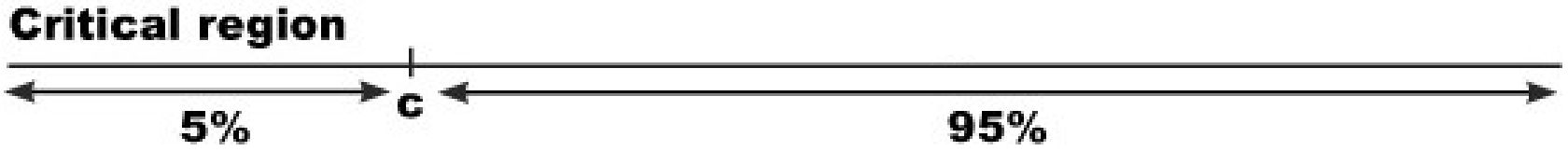
In hypothesis testing, do you assume the null hypothesis to be true or false?

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.

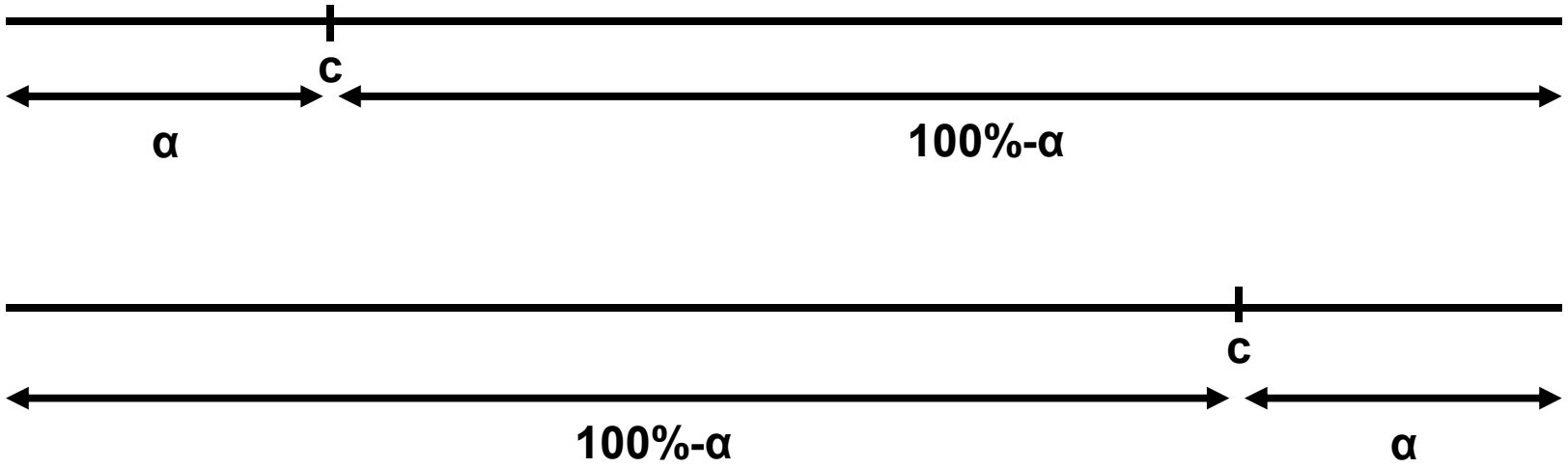
Critical Region Up Close

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.

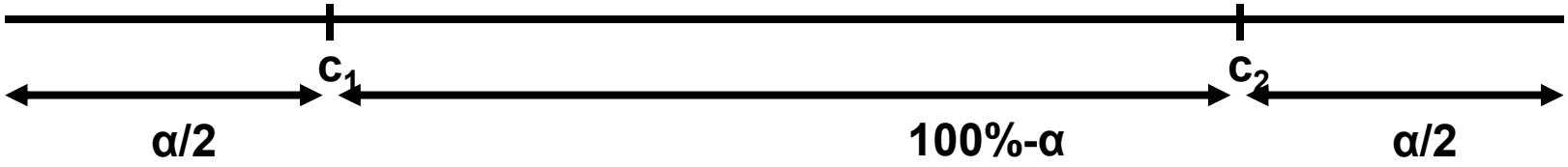


Critical Region Up Close

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

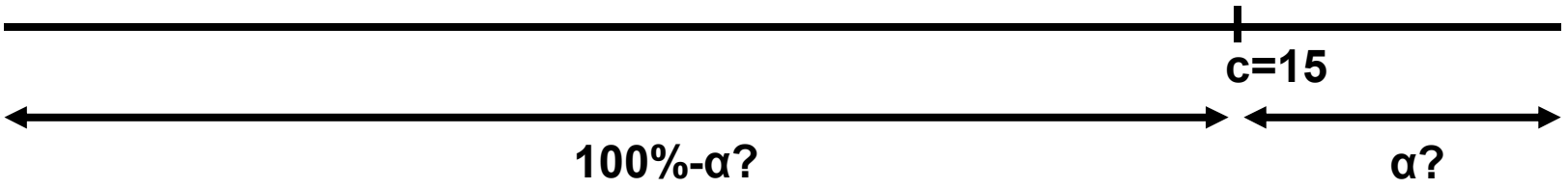
- The coin is biased towards heads with probability 0.8.
  One-tailed/Upper-tailed

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?



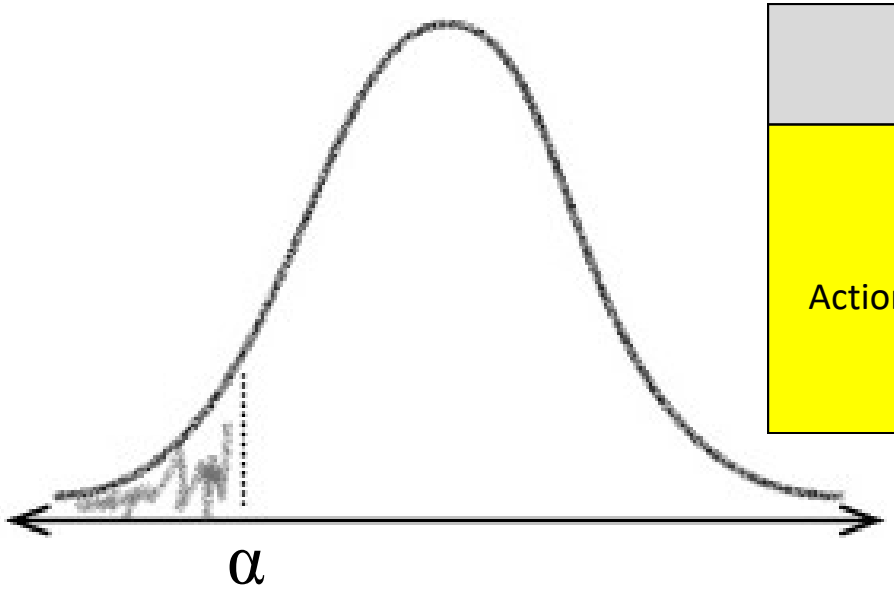
The hypothesis test doesn't answer the question whether the coin is biased or not; it only states whether the evidence is enough to reject the null hypothesis or not ***at the chosen significance level***.

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

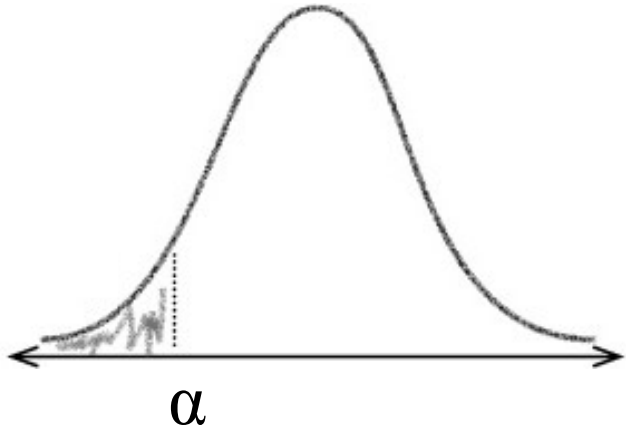
Probability of Getting Type II Error

P(Type II error) = β

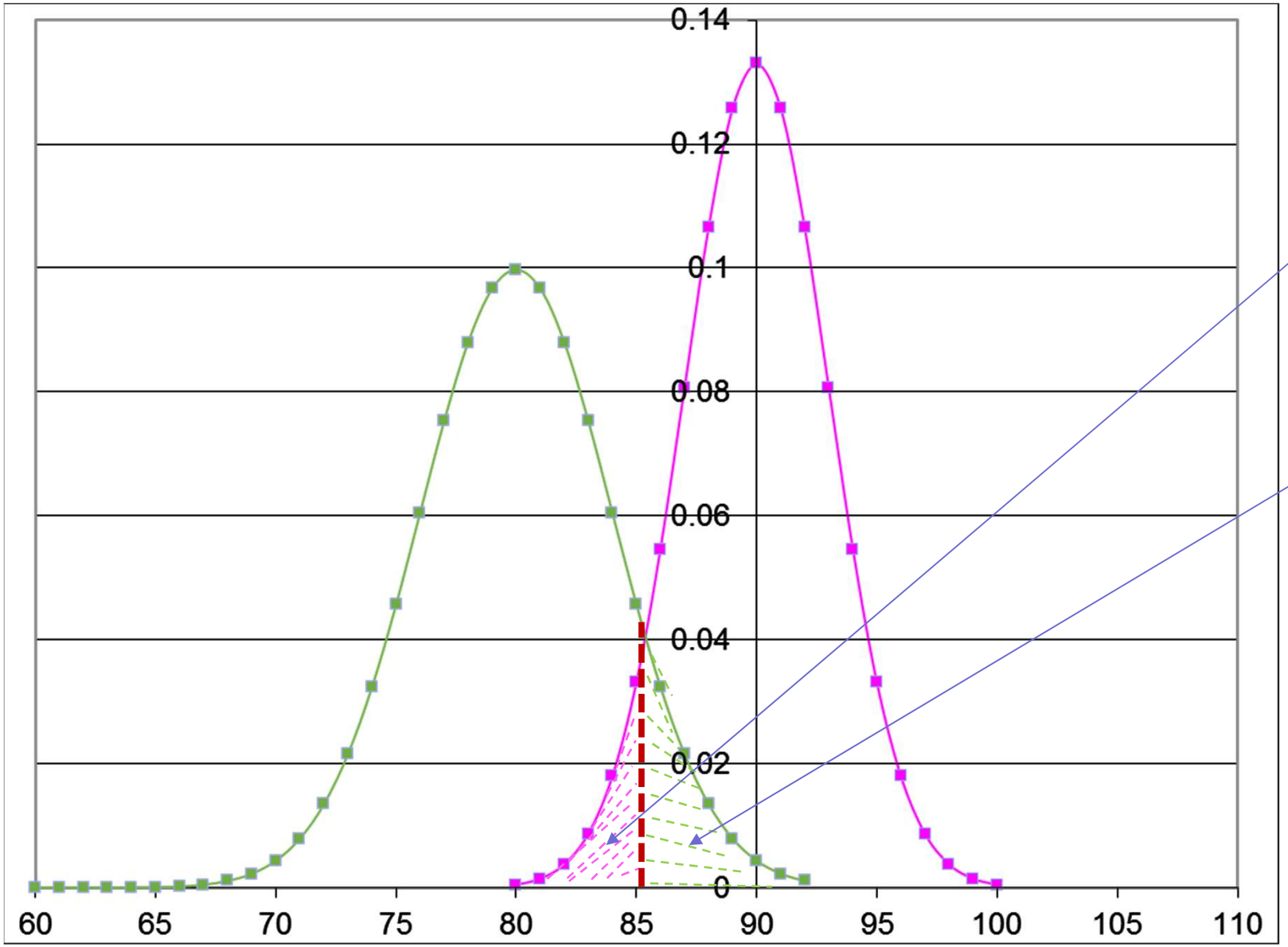
To find β (the difficult way to remember),

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

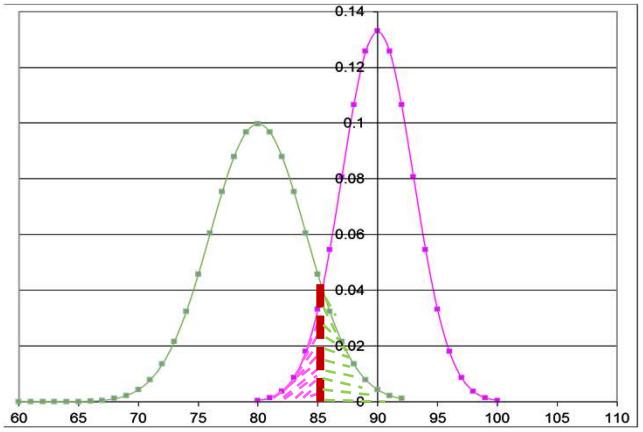
P(Z < z) = 0.05 for 5% Significance Value. From probability tables, z = -1.64.

Using R, qnorm(0.05,0,1) = -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%.

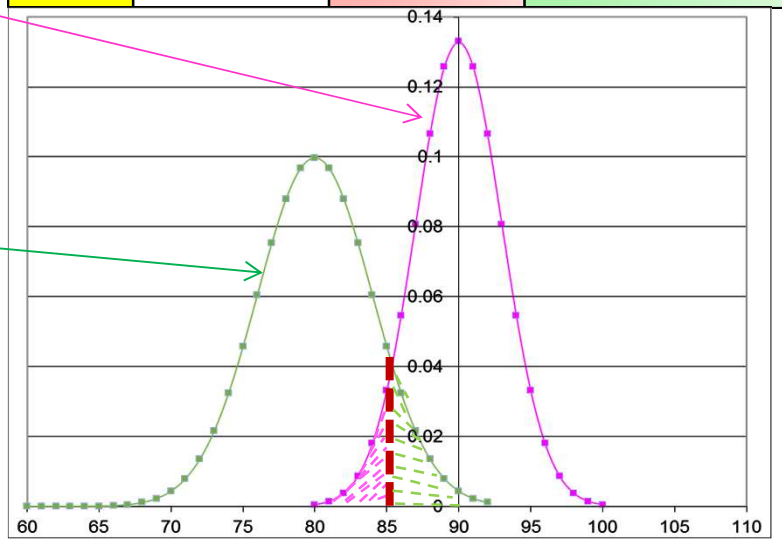
Probability of Getting Type II Error

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

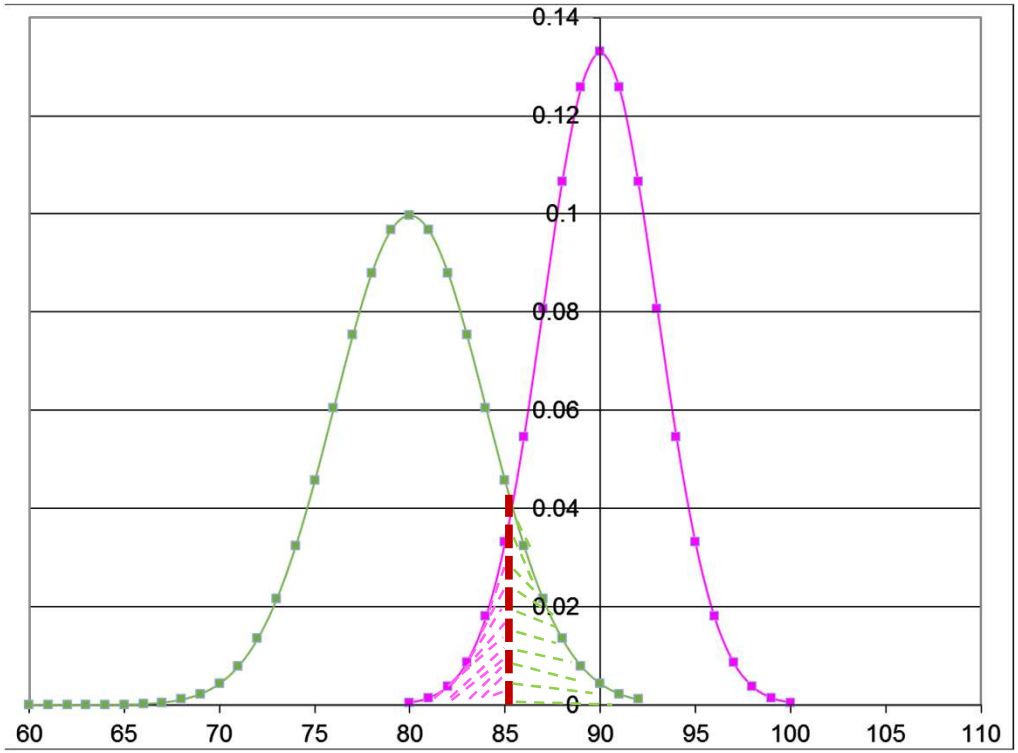


Power of a Hypothesis Test

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.

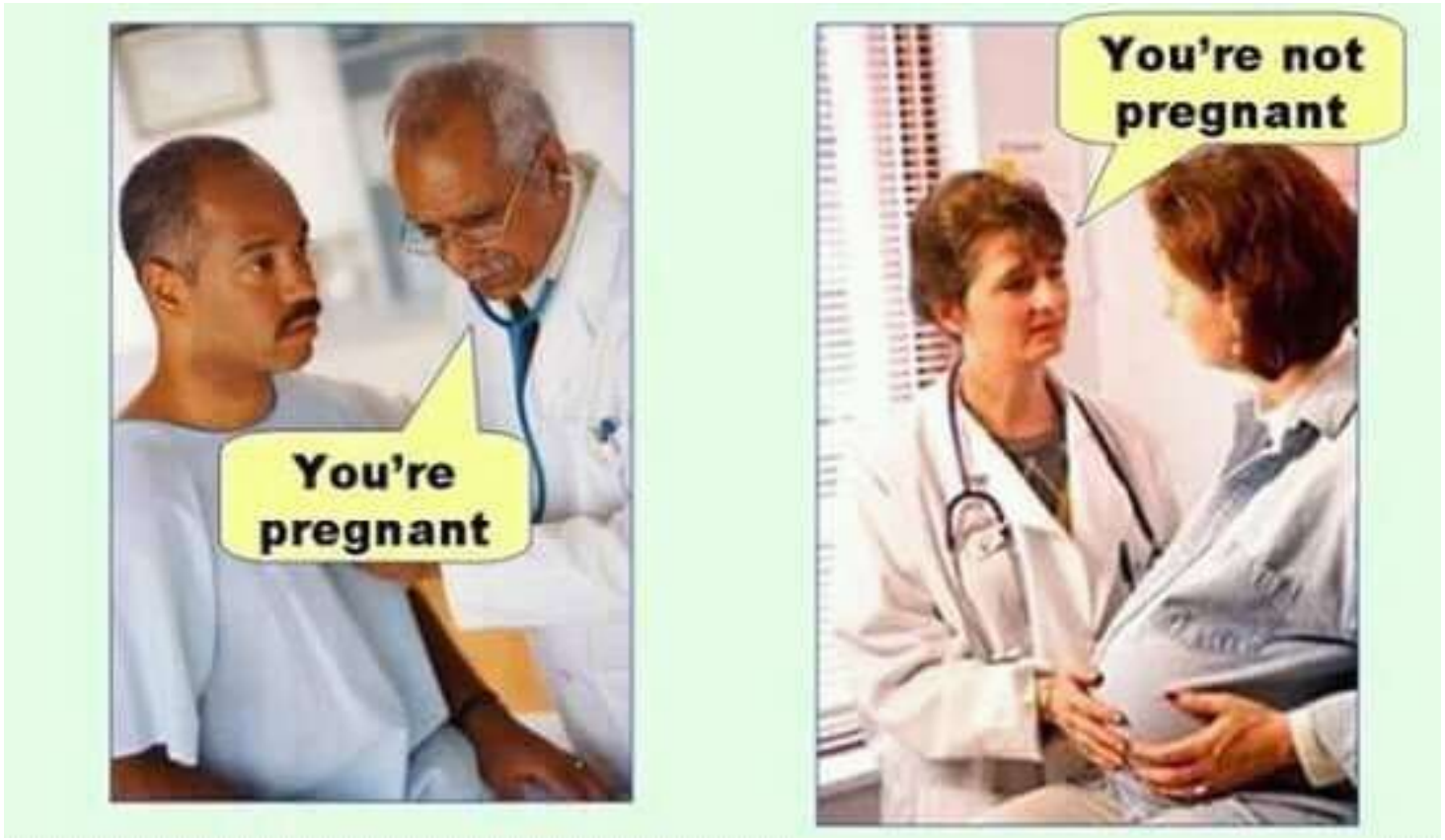
What is the Power of the test?

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.

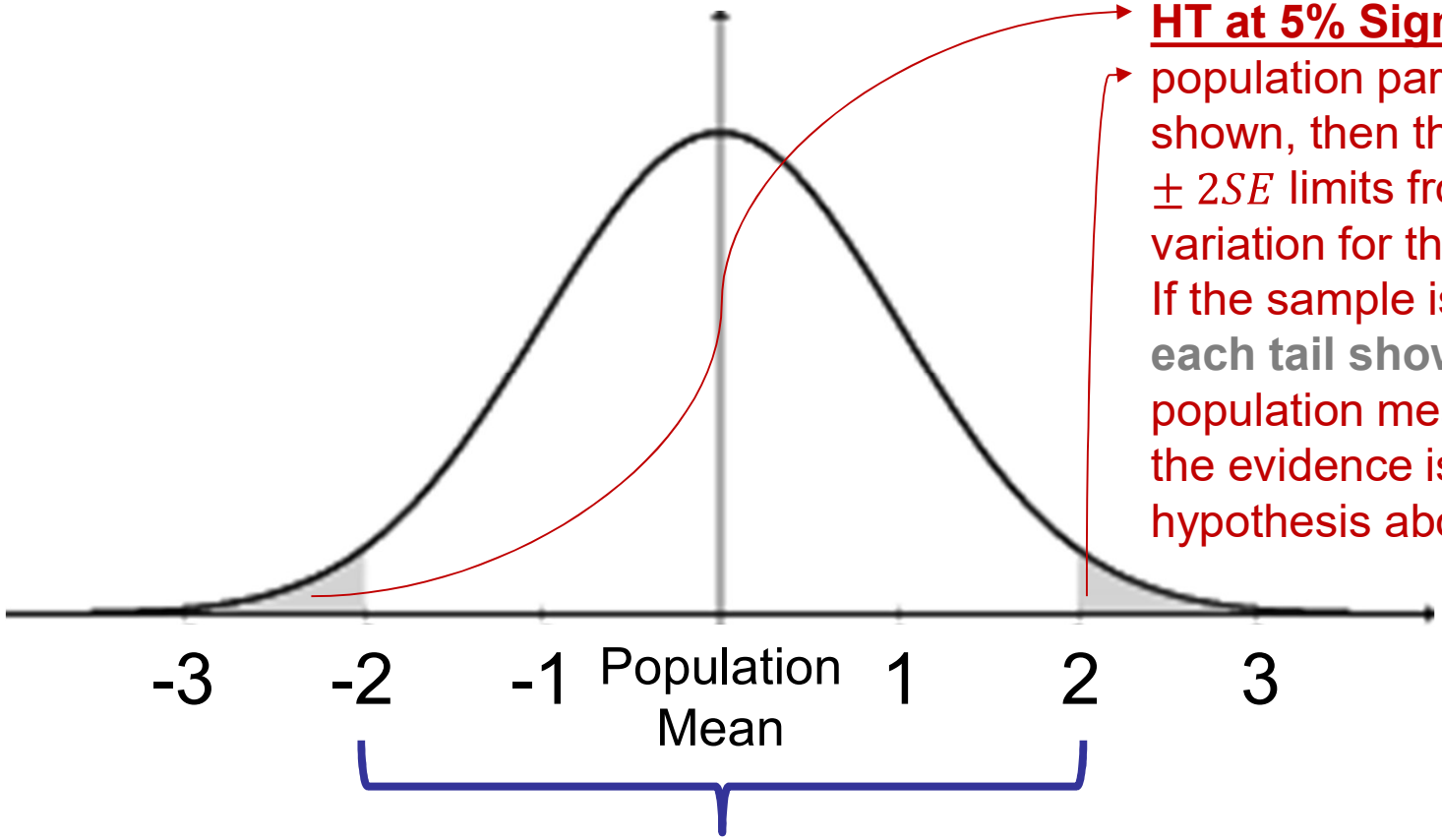
Hypothesis Testing

TYPE I ERROR
(False +ve)

TYPE II ERROR
(False -ve)



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 $\bar{x} \pm 2SE$.)

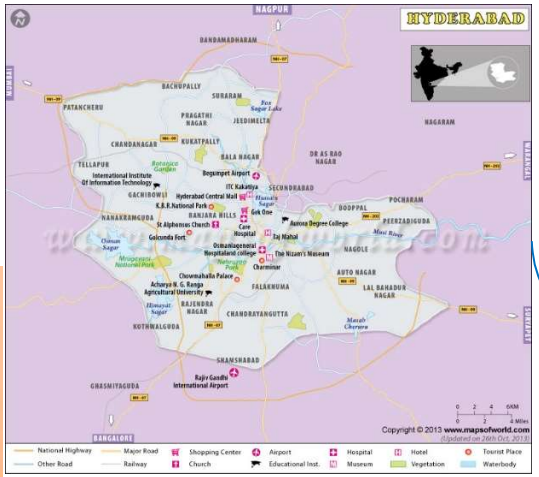
Common Test Statistics for Inferential Techniques

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

- z
- t
- χ^2 (Chi-squared)
- F

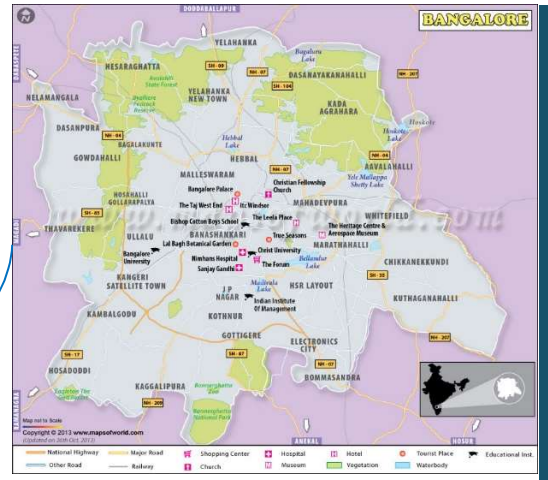
Closely related to Sampling Distribution of **Means**

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



INSOFE

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

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.