Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Probability Types; Bayes Theorem; Probability Distributions: Binomial, Normal; Sampling Distribution of Means and CLT; Confidence Intervals; Hypothesis Testing

Prof Anuradha Sharma

Dean of Corporate Training, INSOFE

January 4, 2020

Content: Dr Sridhar Pappu

Probability - Types

Contingency table summarizing 2 variables, Loan Default and Age:

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability - Types

Convert it into probabilities:

(Recall Frequentist Approach of assigning probabilities)

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

$$\frac{10503}{46687} = 0.225$$

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

CSE 7215c

Probability - Types

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Note the above 2 tables are practically the same; one displays frequencies and the other probabilities. But

$$Probability = \frac{Frequency\ of\ the\ Event\ of\ Interest}{Total\ Frequency}$$

Frequencies describe what has happened and probabilities allow us to predict the future. Don't forget, they still are talking about the same data.

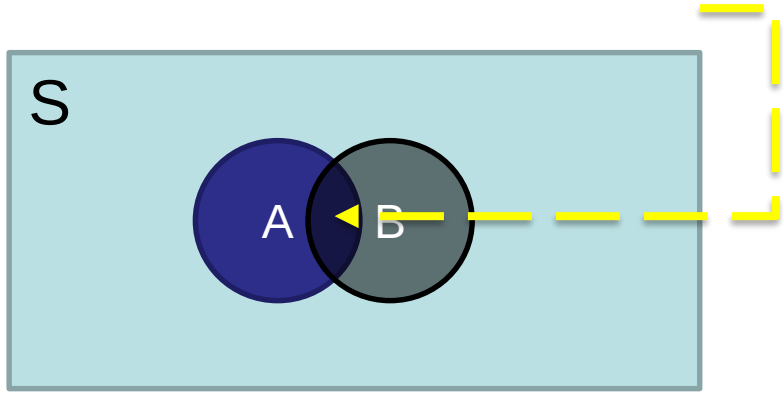
Probability - Types

Joint Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability describing a combination of attributes.

P(Yes and Young) = 0.077

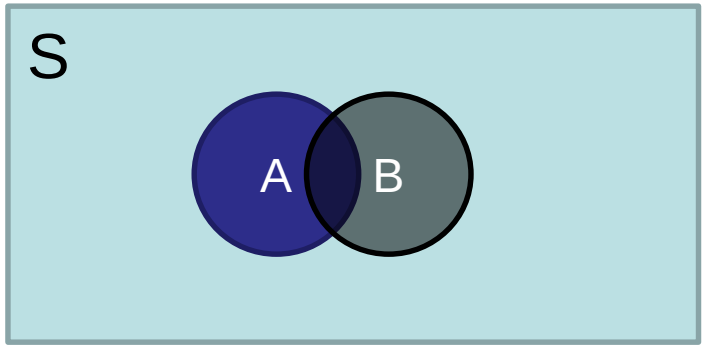


Probability - Types

Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(Yes or Young) = P(Yes) + P (Young) – P(Yes and Young)
= 0.184 + 0.302 – 0.077 = 0.409



Probability - Types

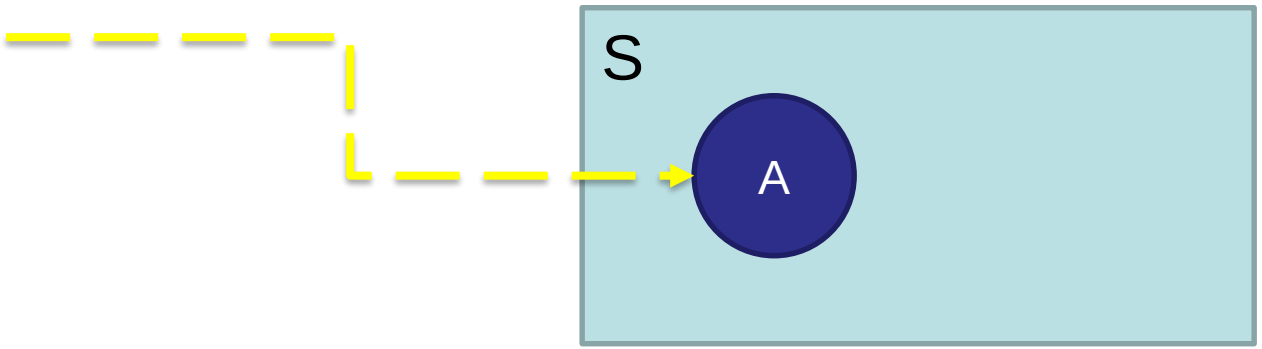
Marginal Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability describing a single attribute.

P(No) = 0.816

P(Old) = 0.008



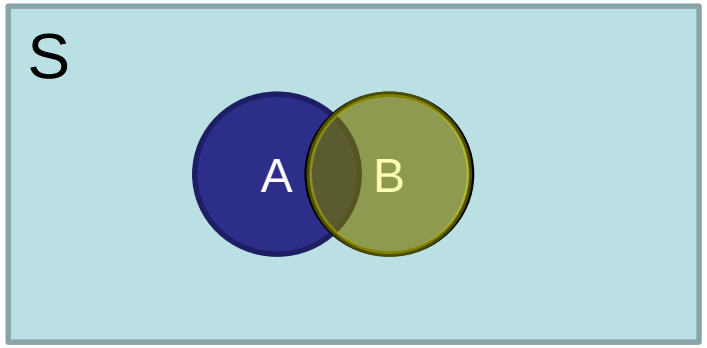
Probability - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability of A occurring **given that** B has occurred.

The sample space is restricted to a single row or column. This makes rest of the sample space irrelevant.



Probability - Types

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Conditional Probability

What is the probability that a person will not default on the loan payment given she is middle-aged?

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability - Types

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(No | Middle-Aged) = 0.586/0.690 = 0.85

Note that this is the ratio of Joint Probability to Marginal Probability,

i.e., $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$ (1)

What is the probability that a person is middle-aged given she has not defaulted on the loan payment?

P(Middle-Aged | No) = 0.586/0.816 = 0.72 (Order Matters)

i.e., $P(B|A) = \frac{P(A\ and\ B)}{P(A)}$ (2)

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} \Rightarrow P(A\ and\ B) = P(B) * P(A|B)$$

Similarly

What happens when A and B are INDEPENDENT?

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)} \Rightarrow P(A\ and\ B) = P(A) * P(B|A)$$

Equating, we get

$$P(A|B) * P(B) = P(A) * P(B|A)$$

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Conditional Probability -> Bayes' Theorem

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

We saw earlier

P(No | Middle-Aged) = 0.586/0.690 = 0.85

$P(A|B) = \frac{P(A)*P(B|A)}{P(B)}$ Here, A is **No** and B is **Middle-Aged**

Let us expand the denominator, which is P(Middle-Aged).

We can see from the table that

P(Middle-Aged) = P(Middle-Aged **and** No) + P(Middle-Aged **and** Yes)

We also have seen that

Joint Probability = Conditional Probability * Marginal Probability

Conditional Probability -> Bayes' Theorem

We saw earlier

P(No | Middle-Aged) = 0.586/0.690 = **0.85**

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

$$P(No|Middle-Aged) = \frac{P(Middle-Aged\ and\ No)}{P(Middle-Aged)}$$

$$= \frac{P(Middle-Aged|No) * P(No)}{P(Middle-Aged\ and\ No) + P(Middle-Aged\ and\ Yes)}$$

$$= \frac{P(Middle-Aged|No) * P(No)}{P(Middle-Aged|No) * P(No) + P(Middle-Aged|Yes) * P(Yes)}$$

$$= \frac{\frac{0.586}{0.816} * 0.816}{\frac{0.586}{0.816} * 0.816 + \frac{0.104}{0.184} * 0.184} = \frac{0.586}{0.586 + 0.104} = \frac{0.586}{0.690} = \mathbf{0.85}$$

Why do we need a more complex looking form of the equation when we can get the answer directly from the table?

Because many a times we will not get all that information and in the detailed form as in the table.

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|not\ A) * P(not\ A)}$$

Bayes' Theorem in Cancer Detection

- Bayes' Theorem lets us look at the skewed test results and correct for errors, recreating the original population and finding the real chance of a true positive result.
- Bayes' Theorem allows you to **find reverse probabilities**, and to **revise original probabilities** based on new information.

1. **Correct for measurement error** if you know the real probability and the chance of a false positive or negative
2. **Relate the actual probability to the measured test probability**-Given test results and knowledge of error rates, you can predict actual chance of having the disease given a positive test

Case – Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is **0.005**. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of **0.85** for detecting cancer correctly. In women without breast cancer, it has a chance of **0.925** for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

Source: https://www.cancer.gov/about-cancer/understanding/statistics

Source: https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

Bayes' Theorem in Cancer Detection

Case – Clinical trials

P(Cancer) = 0.005 (aka Prior Probability)
P(Test positive | Cancer) = 0.85 (aka Likelihood)
P(Test negative | No cancer) = 0.925
The positive test result could be a true positive or a false positive?
P(Cancer | Test positive) = ? (aka Posterior or Revised Probability)
P(Test Positive) aka Evidence

$$Posterior\ Probability = \frac{Prior\ Probability * Likelihood}{Evidence}$$

Bayes' Theorem in Cancer Detection

Case – Clinical trials

P(Cancer) = 0.005 (aka Prior Probability)
P(Test positive | Cancer) = 0.85 (aka Likelihood)
P(Test negative | No cancer) = 0.925
P(Cancer | Test positive) = ? (aka Posterior or Revised Probability)
P(Test Positive) aka Evidence

$$P(Cancer|Test+) = \frac{P(Cancer) * P(Test+|Cancer)}{P(Test+|Cancer) * P(Cancer) + P(Test+|No\ cancer) * P(No\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995} = \frac{0.00425}{0.078875} = 0.054$$

A positive mammogram only means you have a 5.4% chance of cancer, rather than 85% (the supposed accuracy of the test). There means there will be many false positives in a given population. For a rare disease, most of the positive test results will be wrong.

Should the person be worried she has cancer? NO!

Bayes' Theorem in Spam Detection

Case – Spam filtering

Apache SpamAssassin™

Latest News

2015-04-30: SpamAssassin 3.4.1 has been released! Highlights include:

- improved automation to help combat spammers that are abusing new top level dc
- tweaks to the SPF support to block more spoofed emails;
- increased character set normalization to make rules easier to develop and stop sp
- continued refinement to the native IPv6 support; and
- improved Bayesian classification with better debugging and attachment hashing.

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 20% of the mails marked as spam, i.e., P(Free | Spam) = 0.20. Assuming 0.1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Bayes' Theorem in Spam Detection

Case – Spam filtering

P(Spam) = 0.50
P(Free | Spam) = 0.20
P(Free | No spam) = 0.001
P(Spam | Free) = ?

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

All messages with word 'free' are usually spam!

Analyzing attributes

PROBABILITY DISTRIBUTIONS

Sample Software Output – Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05
Residual	25	443.7745253	17.75098101		
Total	26	913.4318519			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962

Sample Software Output – Logistic Regression

```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95015  -0.32016  -0.05335   0.26538   1.72940

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

Random Variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

Points scored per game	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Frequency, f	1	4	6	12	5	1	1

Points scored per game	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Probability *Recall the Frequentist (empirical) approach of assigning probabilities*	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{12}{30}$	$\frac{5}{30}$	$\frac{1}{30}$	$\frac{1}{30}$

Leads to Descriptive Stats

Leads to Inferential Stats

CSE 7215c

Probability Distribution of Income

Income (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

Frequency Distribution

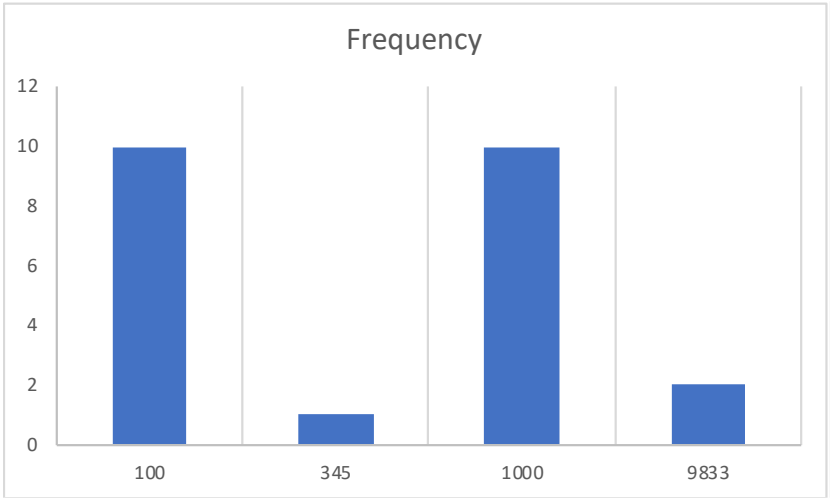
Frequency converted to probability using the Frequentist (or the Empirical) approach of assigning probabilities

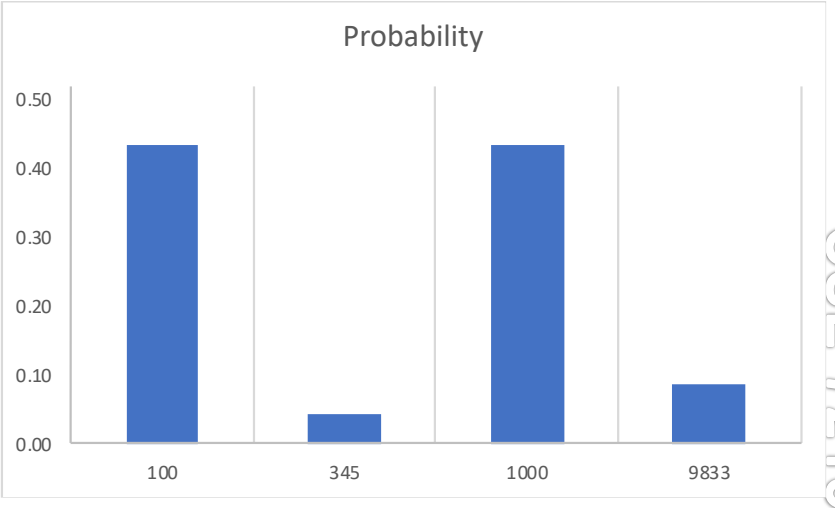
Income (BHD)	100	345	1000	9833
Probability	0.43	0.04	0.43	0.09

Probability Distribution

Why do you need a probability distribution?

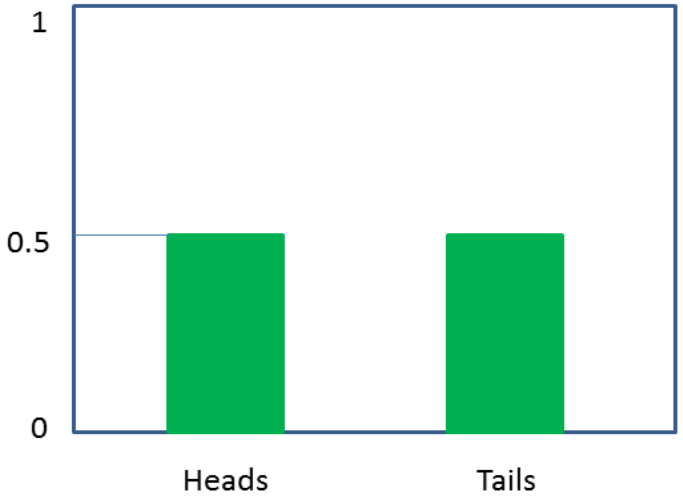
Once a distribution is calculated, it can be used to determine the EXPECTED outcome(most likely outcomes) and the probability of any event of interest to us.

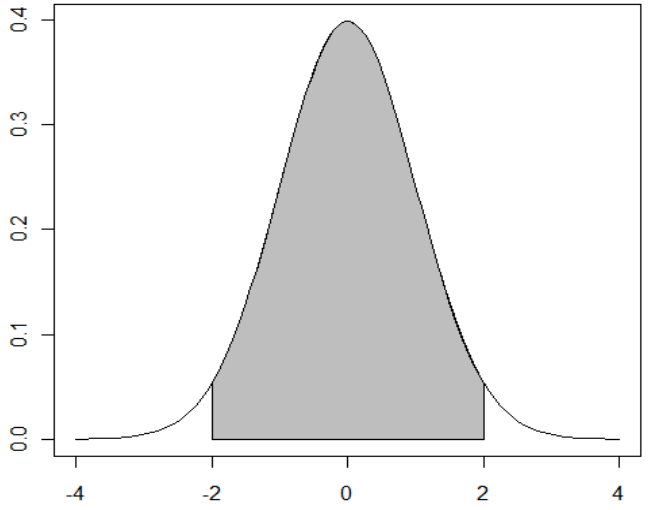




CSE 7215c

Discrete and Continuous





Countable

Measurable

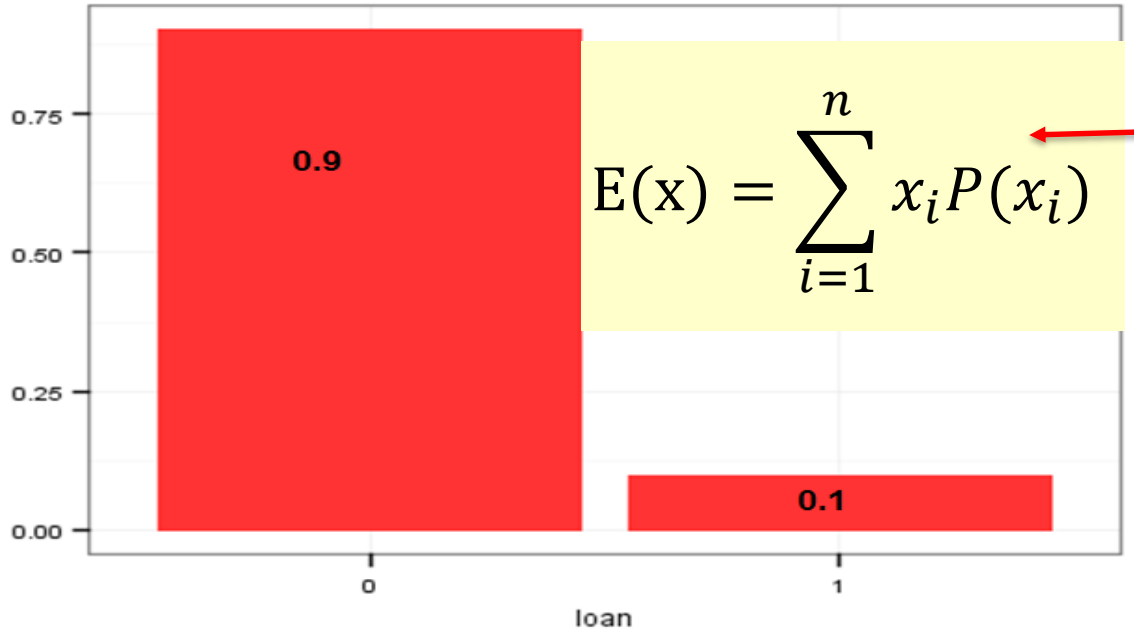
CSE 7215c

Can any function be a probability distribution?

Discrete Distributions	Continuous Distributions
Probability that X can take a specific value x is $P(X = x) = p(x)$.	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$.
It is non-negative for all real x.	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx = 1$
Probability Mass Function (PMF)	**Probability Density Function (PDF)**

Mathematical functions that generate the data are the PMF and PDF-we use them to get the probabilities of interest to us

Expectation: Discrete



$$\mathrm{E}(\mathrm{x}) = \sum_{i=1}^{n} x_i P(x_i)$$

Recall anything like this?

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2
Probability	0.43	0.04	0.43	0.09

Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$

Expectation, $E(X) = 100 * 0.43 + 345 * 0.04 + 1000 * 0.43 + 9833 * 0.09 = 1348$

Variance

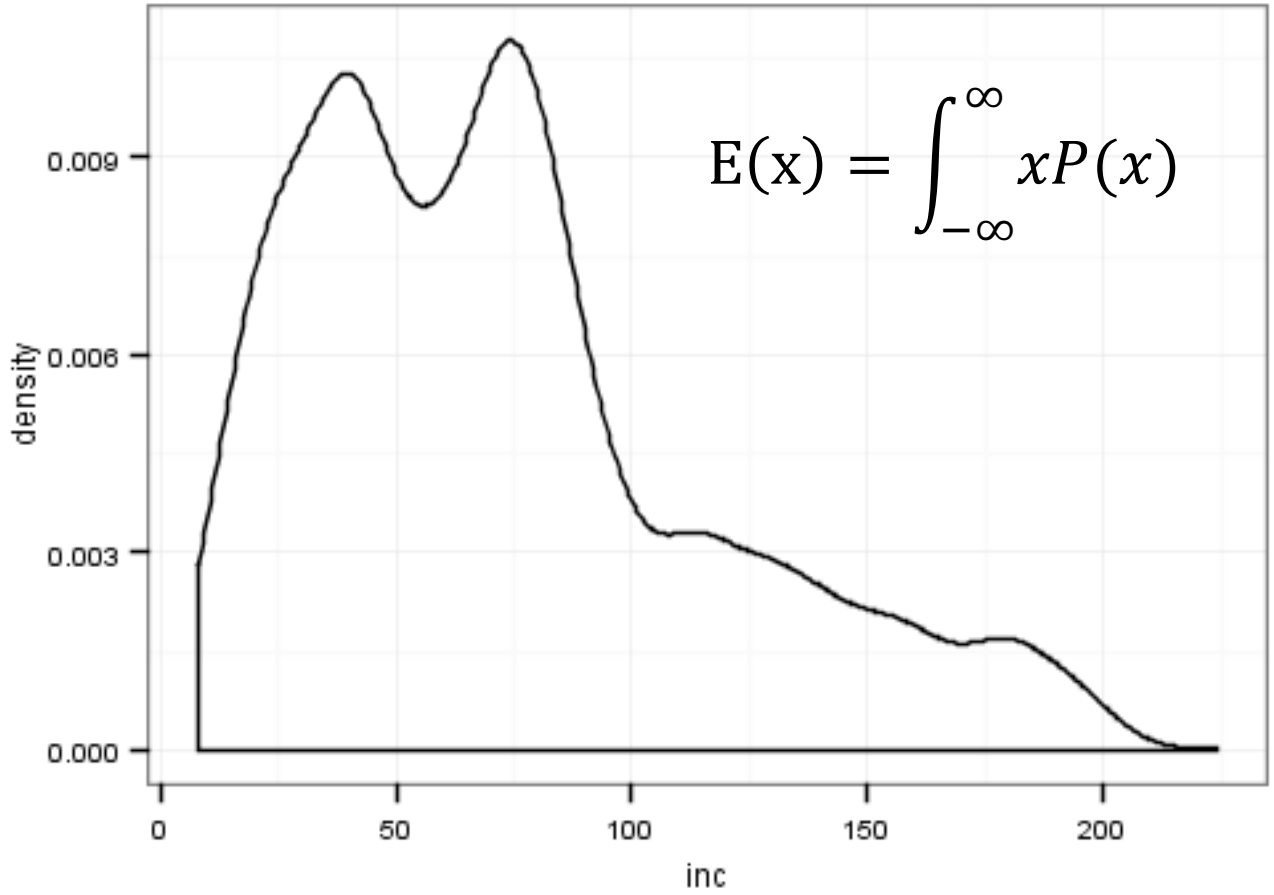
EXPECTATION, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

VARIANCE, $Var(X) = Mean\ (Expectation)\ of\ the\ Squared\ Deviations, i.e.,$

$$E(X - \mu)^2 = \Sigma(x - \mu)^2 P(X = x)$$

$$\sigma = \sqrt{Var(X)}$$

Expectation: Continuous

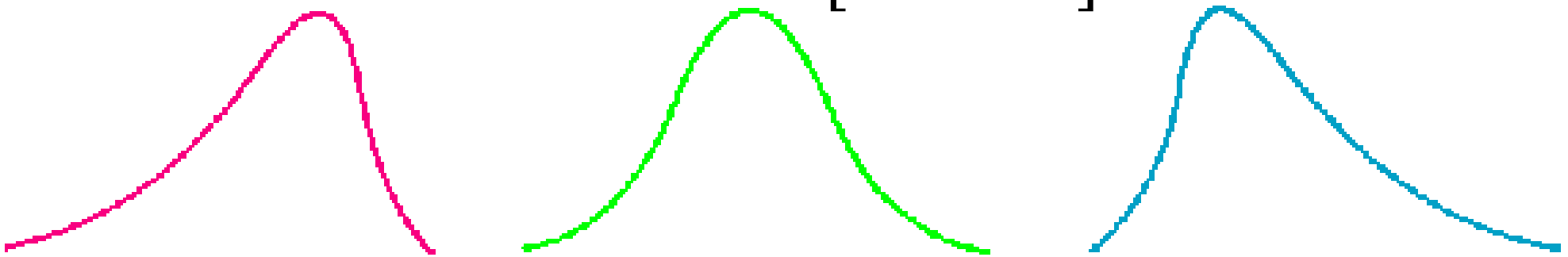


$$\mathrm{E(x)} = \int_{-\infty}^{\infty} xP(x)$$

Further Understanding the Shape of a Distribution - Skewness

A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

$$skew(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$



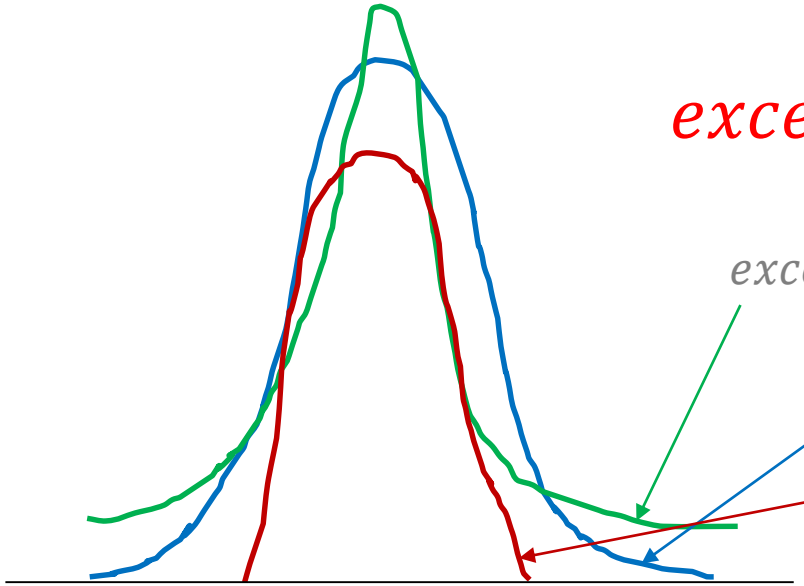
Negatively (left) skewed distribution

Normal distribution

Positively (right) skewed distribution

Further Understanding the Shape of a Distribution - Kurtosis

A measure of the ‘tailed’ness of the data distribution as compared to a normal distribution. Negative kurtosis means a distribution with light tails (fewer extreme deviations from mean (or outliers) than in normal distribution). Positive kurtosis means a distribution with heavy tails (more outliers than in normal distribution).



$$excess\ kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

$excess$ Kurtosis > 0

$excess$ Kurtosis = 0 for Normal

$excess$ Kurtosis < 0

Kurtosis for a normal distribution is 3

Rules of Thumb – Skewness and Kurtosis

Skewness

- Highly skewed: < -1 or > +1
- Moderately skewed: -1 to -0.5 or 0.5 to 1
- Symmetrical: -0.5 to 0.5

Excess Kurtosis

- High: < -1 or > +1
- Medium: -1 to -0.5 or 0.5 to 1
- Small: -0.5 to 0.5

Describing a Distribution – Summary

Measure	Formula	Description
Mean (μ)	$E(X)$	Measures the centre of the distribution of X
Variance (σ^2)	$E[(X-\mu)^2]$	Measures the spread of the distribution of X about the mean
Skewness	$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$	Measures asymmetry of the distribution of X
Kurtosis (excess)	$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]-3$	Measures 'tailed'ness of the distribution of X and useful in outlier identification

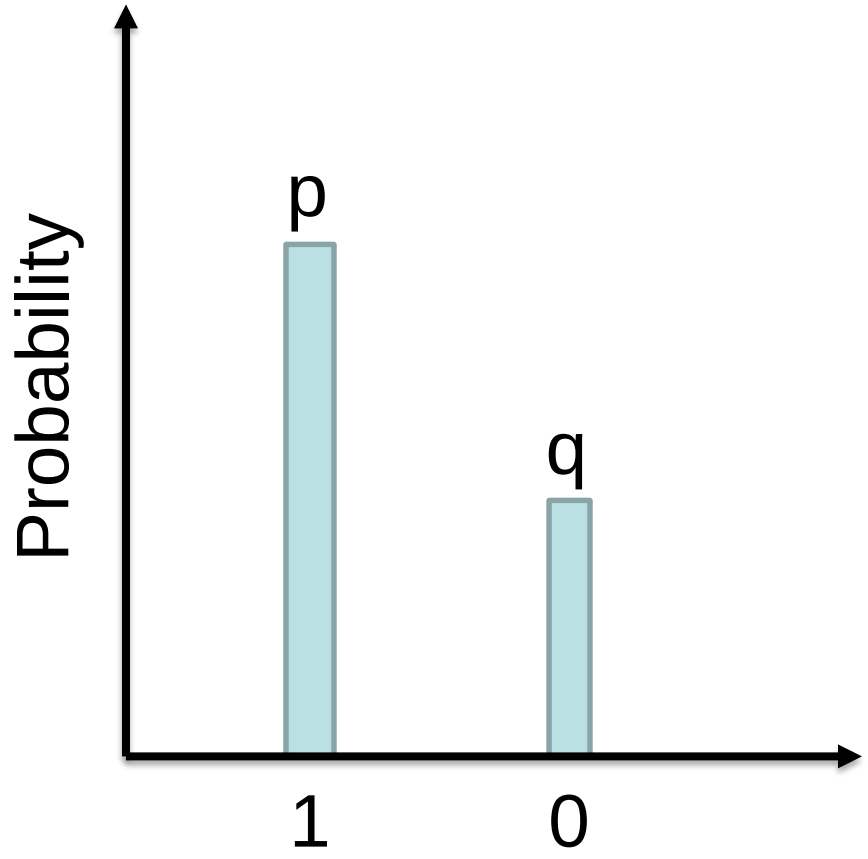
SOME COMMON DISTRIBUTIONS

Bernoulli

There are two possibilities (loan taker or non-taker) with probability *p* of success and *1-p* of failure

- Expectation: *p*
- Variance: *p(1-p)* or *pq*, where *q=1-p*

Bernoulli



$$Expectation, E(X) = \sum x_i P(x_i)$$

$$= 1 * p + 0 * q = p$$

$$Variance, Var = \sum (x_i - \mu)^2 P(x_i)$$

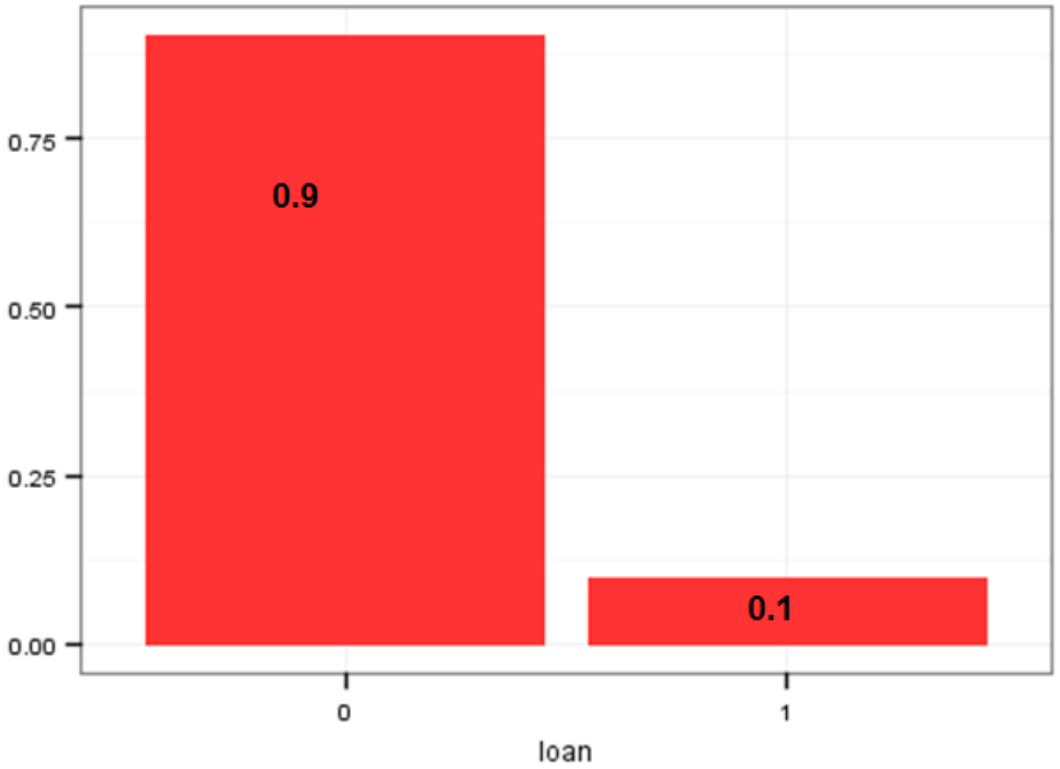
$$= (1-p)^2 * p + (0-p)^2 * (1-p)$$

$$= p(1-p)$$

Binomial Distribution

If I randomly pick 10 people, what is the probability that I will get exactly

- 0 loan takers = 0.9^{10}
- 1 loan taker = $10 * 0.1^1 * 0.9^9$
- 2 loan takers = $C_2^{10} * 0.1^2 * 0.9^8$



Binomial Distribution

If there are two possibilities with probability *p* for success and *q* for failure, and if we perform *n* trials, the probability that we see *r* successes is

PMF, $\mathrm{P}(\mathrm{X} = \mathrm{r}) = C_r^n p^r q^{n-r}$

CDF, $\mathrm{P}(\mathrm{X} \leq \mathrm{r}) = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Binomial Distribution

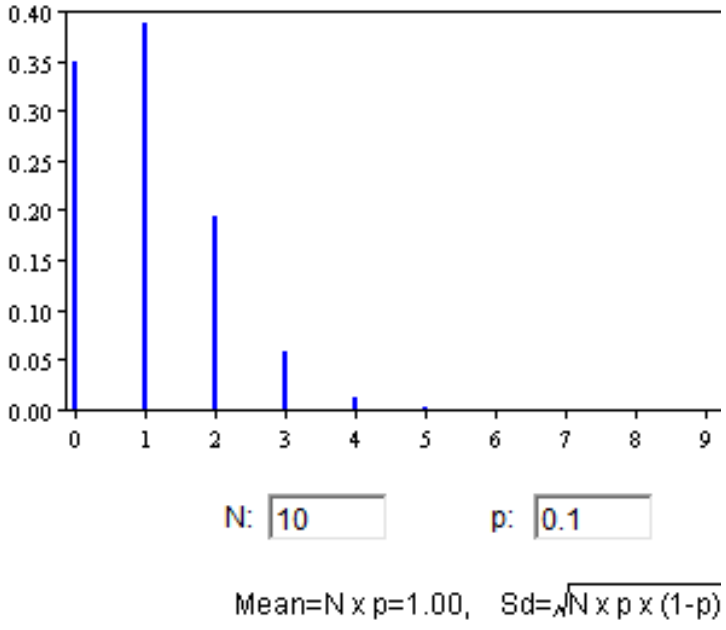
$$E(X) = np \qquad Var(X) = npq$$

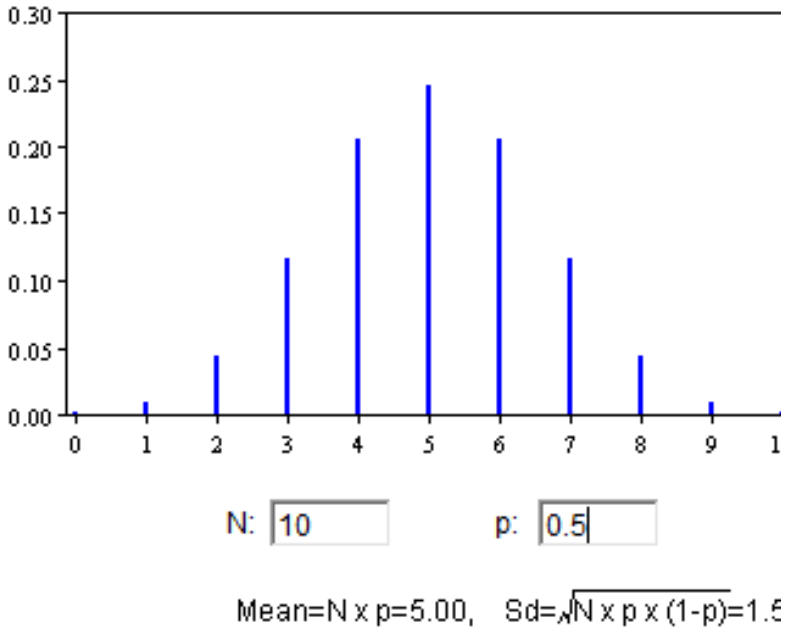
When to use?

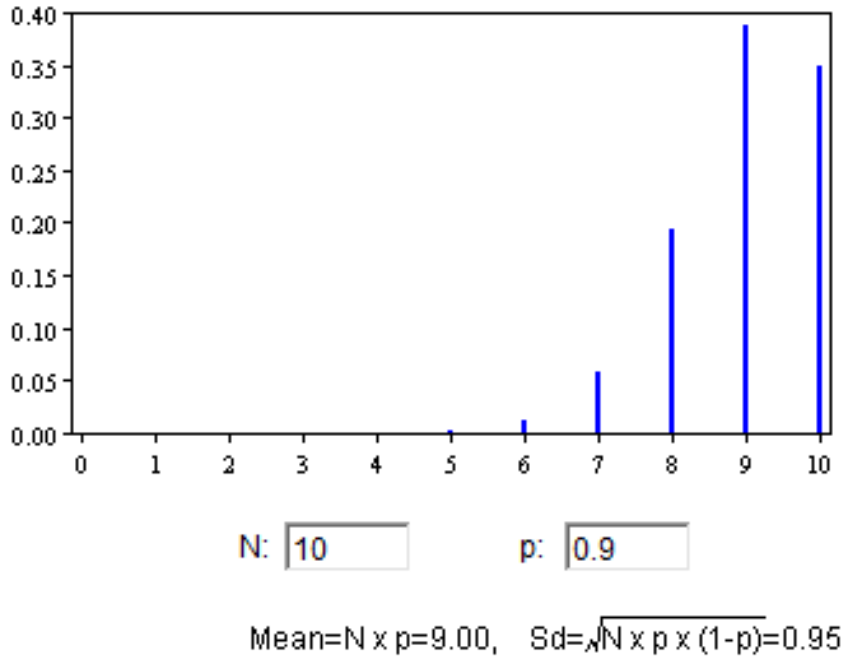
- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same (identical) for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

X~B(n,p)

$$P(X = r) = C_r^n p^r q^{n-r}$$







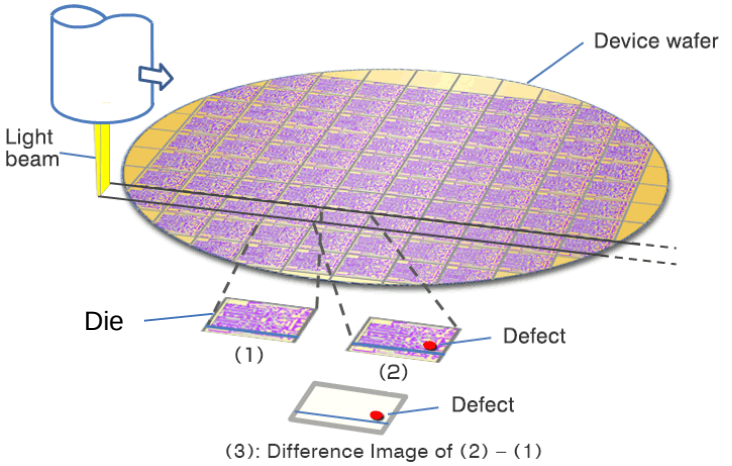
Ref: http://onlinestatbook.com/2/probability/binomial_demonstration.html
Last accessed: June 14, 2019 on Chrome with CheerpJ addin

Binomial Distribution in Manufacturing Industry

A company makes semiconductor wafers. The probability of a defective die on the wafer is 0.001. What is the probability that a random sample of 500 dies will contain exactly 5 defective dies?

n = 500, p = 0.001, r = 5

$^{500}C_5*(0.001)^5*(1-0.001)^{495} = 0.00156$

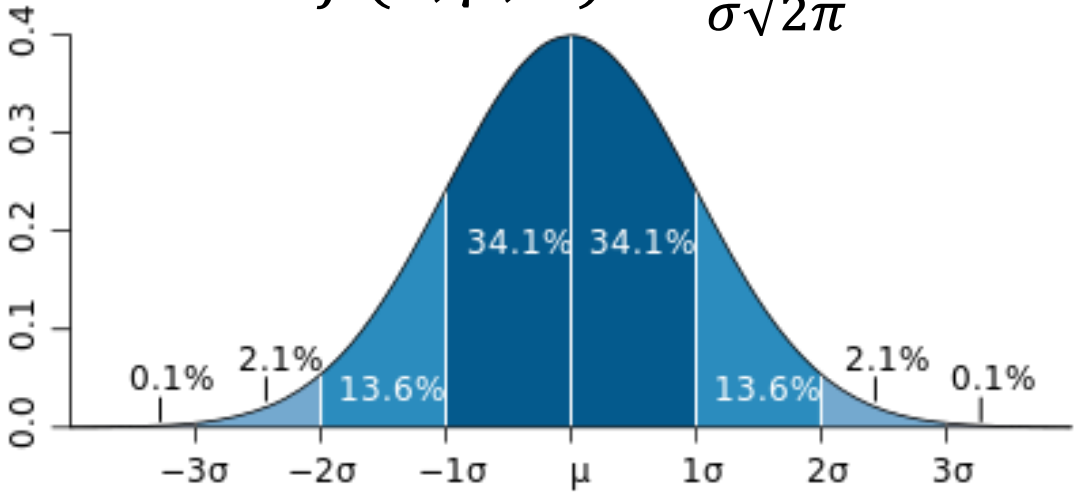


NORMAL DISTRIBUTION

Normal (Gaussian) Distribution

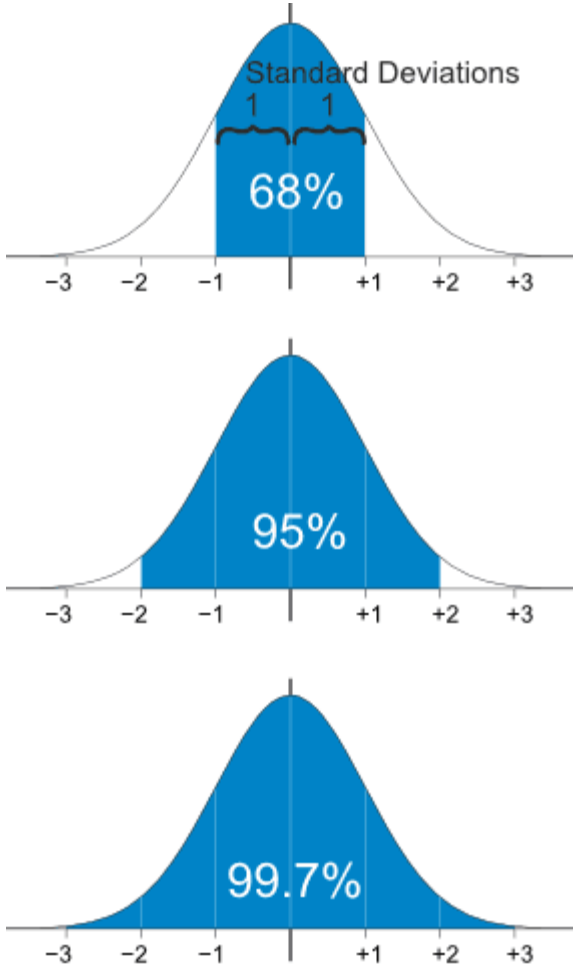
- Mean = Median = Mode
- 68-95-99.7 empirical rule
- Zero Skew and Kurtosis
- X ~ N(μ, σ^2)
- Shaded area gives the probability that X is between the corresponding values

PDF: $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Normal Distribution

You know the 68-95-99.7 rule.



A company produces a valve that is specified to weigh 1500g, but there are imperfections in the process. While the mean weight is 1500g, the standard deviation is 300g.

Q1. What is the range of weights within which 95% of the valves will fall?

Q2. Approximately 16% of the weights will be more than what value?

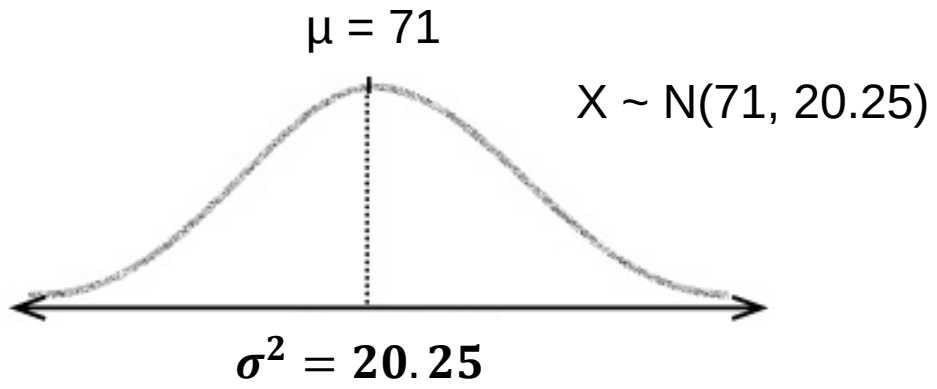
Q3. Approximately 0.15% of the weights will be less than what value?

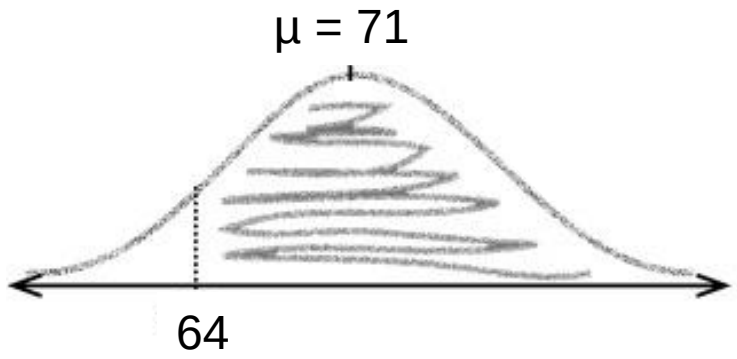
Image source: http://www.mathsisfun.com/data/standard-normal-distribution.html
Last accessed: December 15, 2017

Calculating Normal Probabilities

Step 1: Determine the distribution

Julie wants to marry a person taller than her and is going on blind dates. The mean height of the 'available' guys is 71" and the variance is 20.25 inch^2 (yuck!).



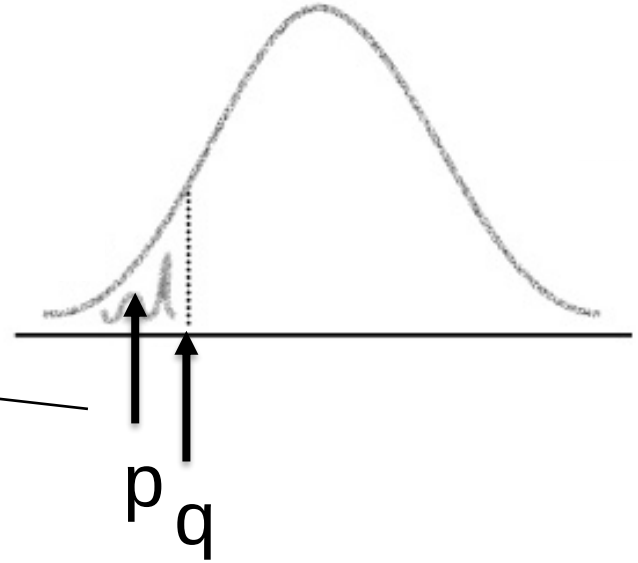


Calculating Normal Probabilities

Step 2: Calculate the probability

Probability (Cumulative Distribution Function, CDF) in a *Normal Distribution*

R: pnorm
Python: scipy.stats.norm.pdf



Quantile (Inverse CDF) or Percentile Point Function in a *Normal Distribution* – The value corresponding to the desired probability.

R: qnorm
Python: scipy.stats.norm.ppf

Calculating Normal Probabilities

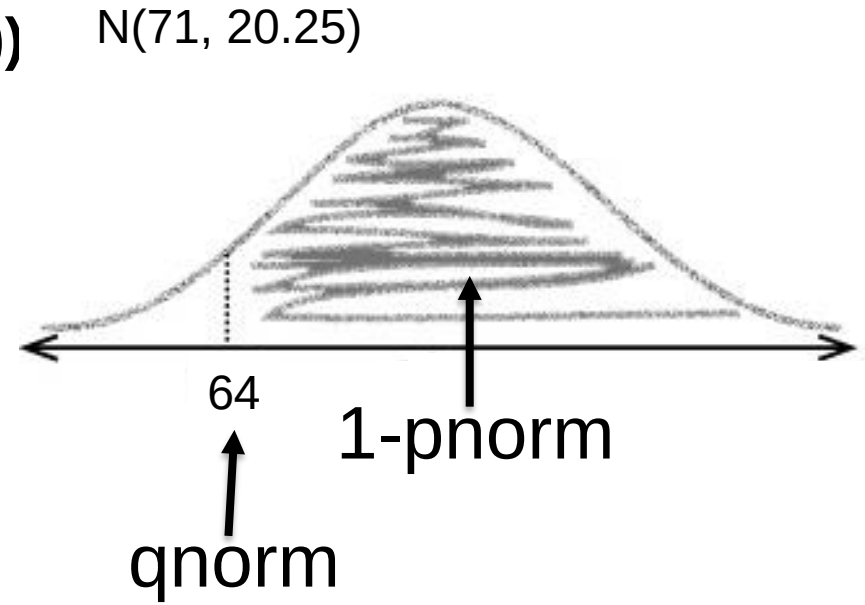
Step 2: Calculate the probability

1-pnorm(64, mean=71, sd=sqrt(20.25))
or
1-pnorm(64, 71, 4.5)

Answer: 1-0.0599 = 94.01%

qnorm(0.0599, 71, 4.5)

Answer: 64



Calculating Normal Probabilities

Q. Julie just realized that she wants her date to be taller when she is wearing her heels, which are 5” high. Find the new probability that her date will be taller.

A. 1-pnorm(69, 71, 4.5). This gives P(X>69) = 67%

Calculating Normal Probabilities

Q. Julie wants to have at least 80% probability of finding the right guy. What is the maximum size of heels she can wear?

A. qnorm(0.20, 71, 4.5). This gives a value of 67.2”. As Julie is 64” tall, the maximum heel size she should wear is about 3”.

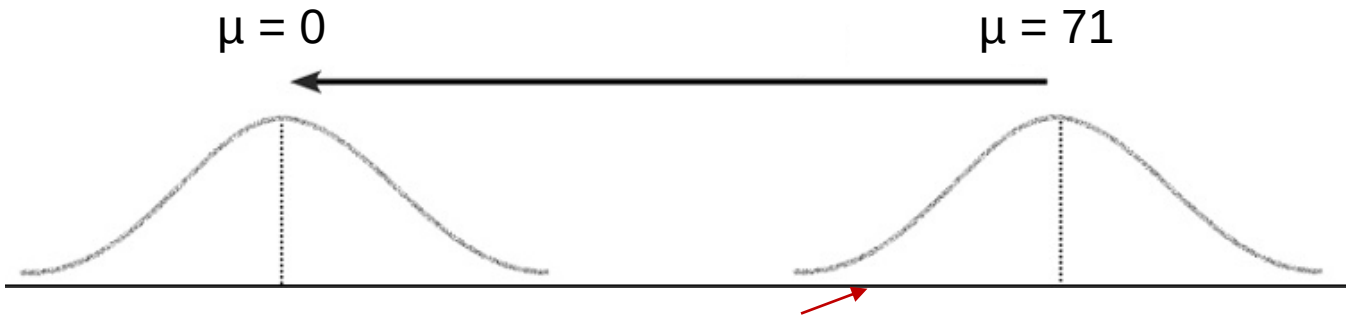
Standard Normal Distribution – z Distribution

Standardize X ~ N(71,20.25) to Z ~ N(0,1)

1. Move the mean

   This gives a new distribution

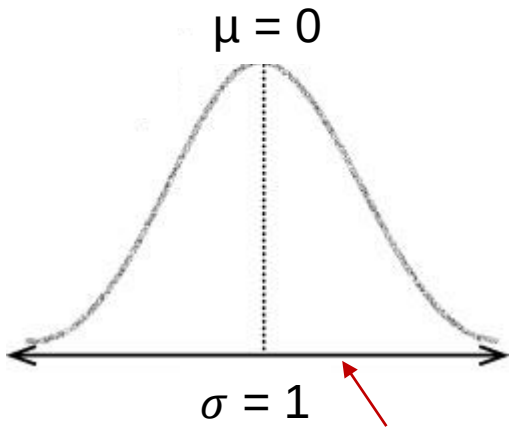
   X-71 ~ N(0,20.25)



Random variable is x, the actual heights of available guys

2. Squash the width by dividing by the standard deviation

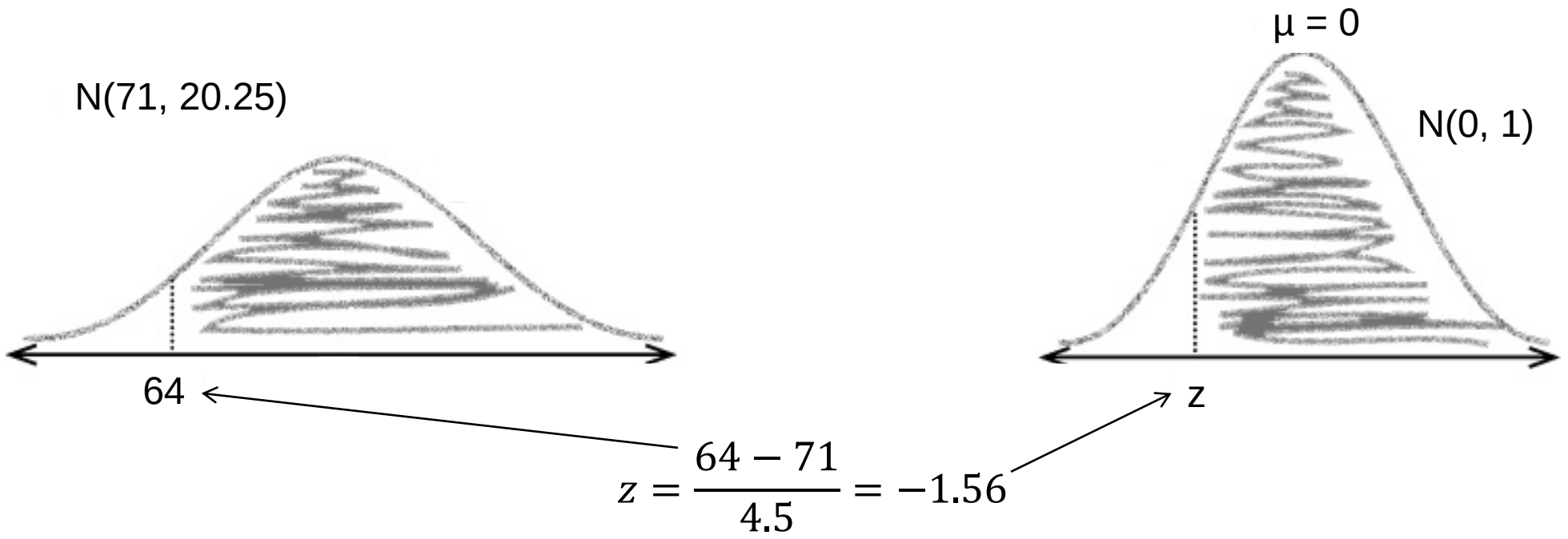
   This gives us $\frac{X-71}{4.5} \sim N(0,1)$



$Z = \frac{X-\mu}{\sigma}$ is called the Standard Score or the z-score.

Random variable is z, the the standardized heights of available guys

Standard Normal Distribution – z Distribution



Julie is 64” tall, i.e., she is 1.56 standard deviations shorter than the average height of the available guys.

1-pnorm(-1.56,0,1) = 1-pnorm(64,71,4.5) = 94.01%

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

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

$\bar{X} \sim N(\mu, \sigma^2/n)$

Sampling Distribution and CLT in Airlines Industry

Fuel Efficiency and Cost of Weight (COW)

The cost of weight is widely used to **evaluate the impact that adding or removing weight has on fuel consumption.**

A first example is **reducing the OEW (operating empty weight)** by removing galley equipment or reducing the quantity of potable water. Reducing equipment weight in your aircraft can result in a significant impact on fuel consumption.

At the Aircraft Commerce Conference in October 2018 in Bangkok, Arief Rachman, Senior Manager, Head of Scheduling Department at Citilink Indonesia, explained how removing one oven from the cabin resulted in saving 20kg of fuel per flight due to weight reduction. He also explained how they reduced potable water quantity brought onboard on shorter flights and consequently saved fuel. Based on the fleet size and the number of flights, **it represents 2 tons of fuel per year.**

Airline fuel efficiency is all about multiplying small actions by big numbers. For example, United Airlines decided to use lighter paper on inflight magazine and asserts that this slight weight reduction is saving 643,000 kg of fuel a year.

Source: https://blog.openairlines.com/how-to-use-the-cost-of-weight-to-be-more-fuel-efficient

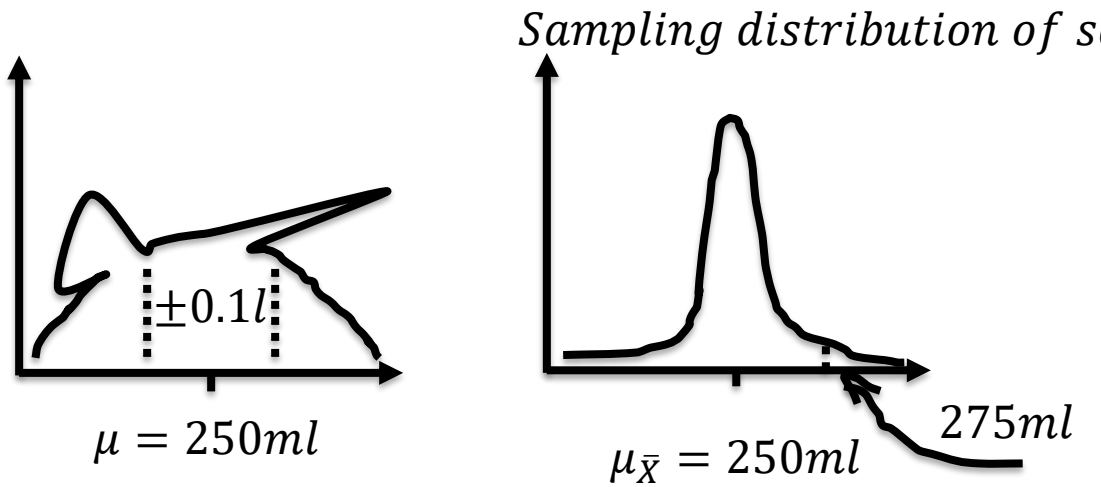
Last accessed: March 23, 2019

Sampling Distribution and CLT in Airlines Industry

The average passenger drinks 250ml of water when flying a short distance with a standard deviation of 0.1l. The flight has 80 passengers and it decided to carry 22l of water keeping in mind COW. What is the probability that they will run out of water?

$\mu = 250, \sigma = 100$

$P(run\ out) \Rightarrow P(use > 22l) \Rightarrow P(average\ water\ use\ per\ passenger > 275ml)$



$$1 - pnorm\left(275, 250, \frac{100}{\sqrt{80}}\right)$$

$$= 0.013, i.e., 1.3\%$$

INFERENTIAL STATISTICS

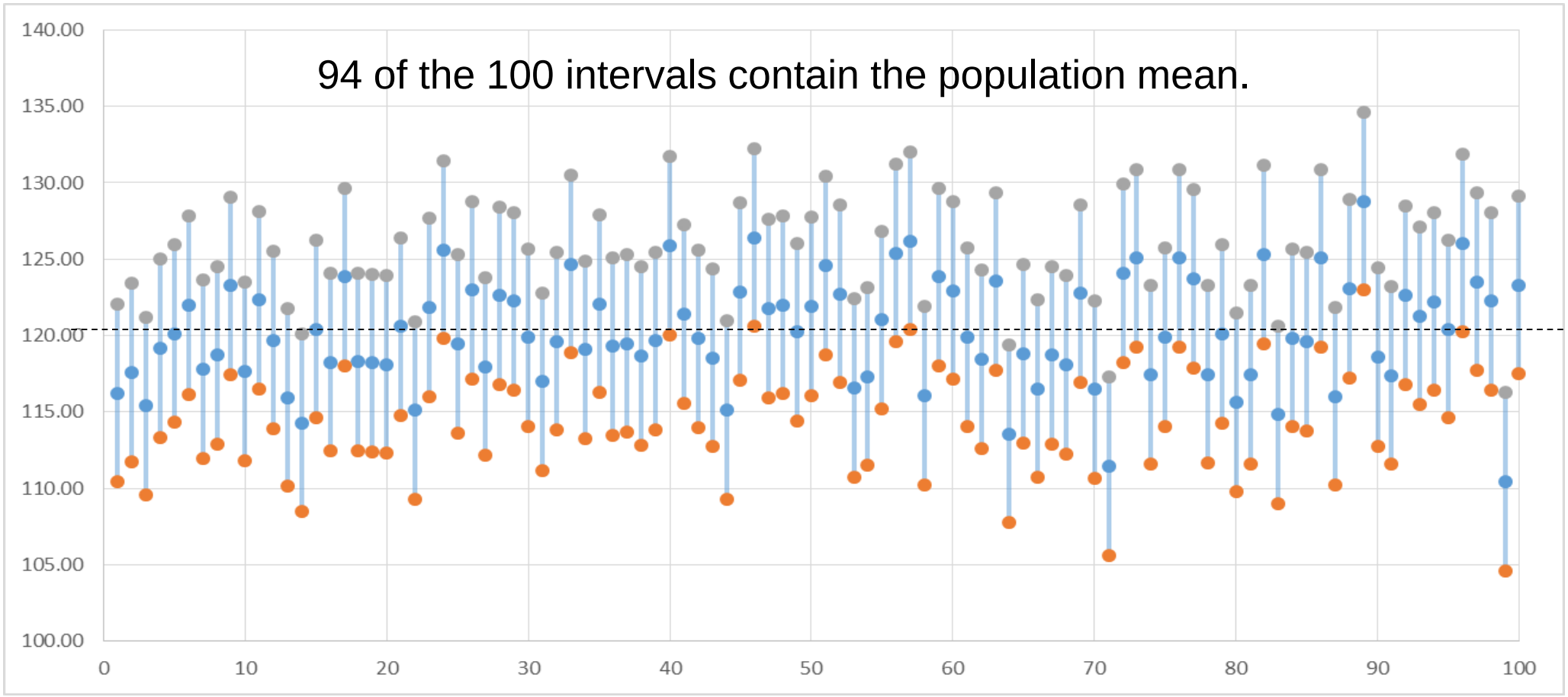
CONFIDENCE LEVELS AND CONFIDENCE INTERVALS

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

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

Uttar Pradesh-80	BJP+	SP-BSP-RLD	Cong	
Times Now-VMR	58	20	2	
India Today-Axis	62-68	10-16	1-2	
Republic-C-Voter	38	40	2	
News18-Ipsos	60-62	17-19	2	
Today's Chanakya-News 24	65	13	2	
ABP-Nielsen	22	56	2	
Maharashtra-48	**NDA**	**UPA**		
Times Now-VMR	38	10		
India Today-Axis	38-42	6-10		
Republic-C-Voter	34	14		
News18-Ipsos	42-44	4-6		
Today's Chanakya-News 24	38	10		
ABP-Nielsen	34	14		
West Bengal-42	**TMC**	**BJP**	**Cong**	
Times Now-VMR	29	11	2	
India Today-Axis	19-22	19-23	0-1	
Republic-C-Voter	29	11	2	
News18-Ipsos	36-38	3-5	0-1	
Today's Chanakya-News 24	23	18	1	
ABP-Nielsen	24	16	2	

Confidence Level and Interval - Excel



A confidence interval is a range of values that is likely to contain the unknown population parameter!

If you draw a random sample many times, a certain percentage of the confidence intervals will contain the population mean. This percentage is the **confidence level**. It is not a probability.

Source : https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-confidence-intervals-and-confidence-levels

Confidence Intervals and Margin of Error in Meteorology

Monsoon to be 'near normal' at 96%: IMD

Vishwa.Mohan@timesgroup.com

New Delhi: India is likely to have 'near normal' monsoon this year, India Meteorological Department said on Monday in what could be an encouraging signal to farmers and the economy ahead of six phases of the Lok Sabha polls.

SEE FLAP OPPOSITE

▶ Rain forecast set to push industrial demand, page 13

Though there is only 39% probability of "near normal" monsoon in the first stage forecast, and a 32% chance of "below normal" rains, the prediction of "well distributed rainfall" will bring cheer to agriculturalists as skewed distribution — a pattern some times linked to climate change — leaves one or the other region in a deficient rainfall (drought) situation.

IMD said the monsoon is likely to be 96% of average, differing from private forecaster Skymet's prediction of "below normal" rains. Interestingly, IMD for the first time used 'near normal' as a category in place of 'normal'.

▶ Drought prediction, P 6

Forecast predicts 17% probability of drought in '19

▶ Continued from page 1

The forecast suggests 2019's southwest monsoon rainfall is likely to be near normal. Quantitatively, the monsoon seasonal (June to September) rainfall is likely to be 96% of the long period average (LPA) with a model error of ±5%," said M Rajeevan, ministry of earth sciences (MoES) secretary.

IMD chief K J Ramesh explained that both term (normal and near normal) were technically the same. The forecast percentage of 96% sits on the line separating normal from below normal.

The LPA of the seasonal rainfall over the country as a whole for the period 1951-2000 is 89 cm. The monsoon is categorised as "normal" when the June-September period has rainfall in the range of 96-104% of the LPA. On the other hand, the seasonal rainfall is considered "below normal" if it falls in the range of 90-96% of the LPA

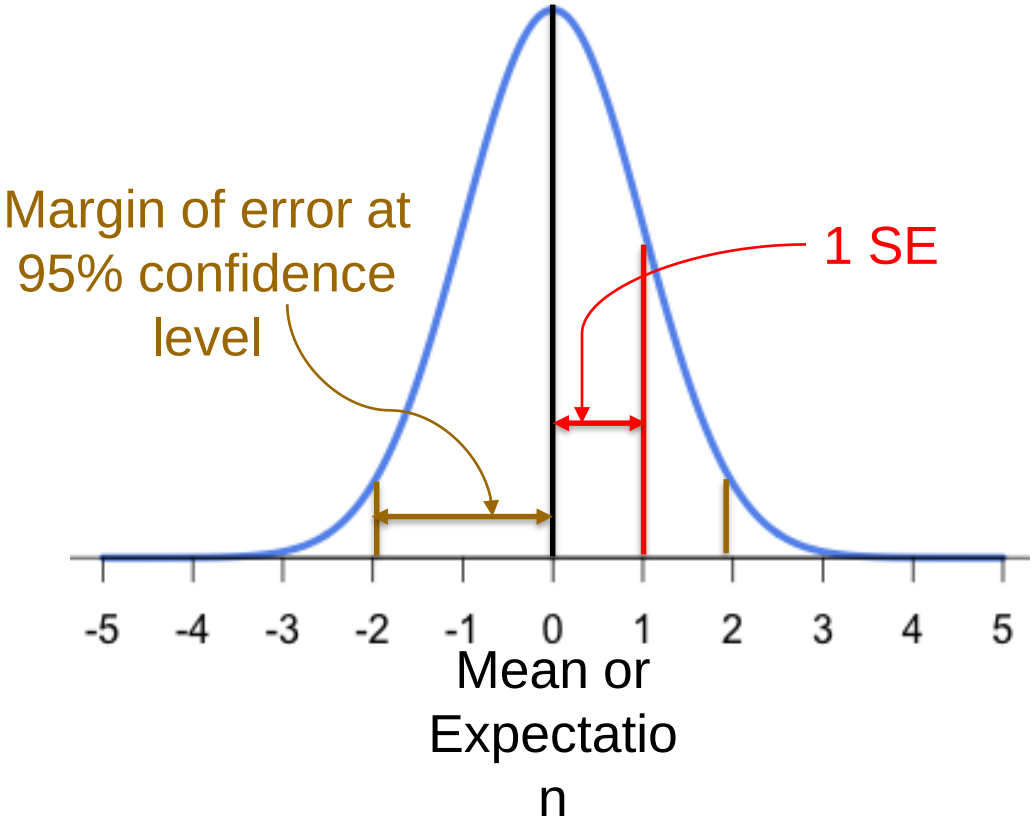
CSE 7215c

SE, Margin of Error, Confidence Interval and Sample Size

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

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

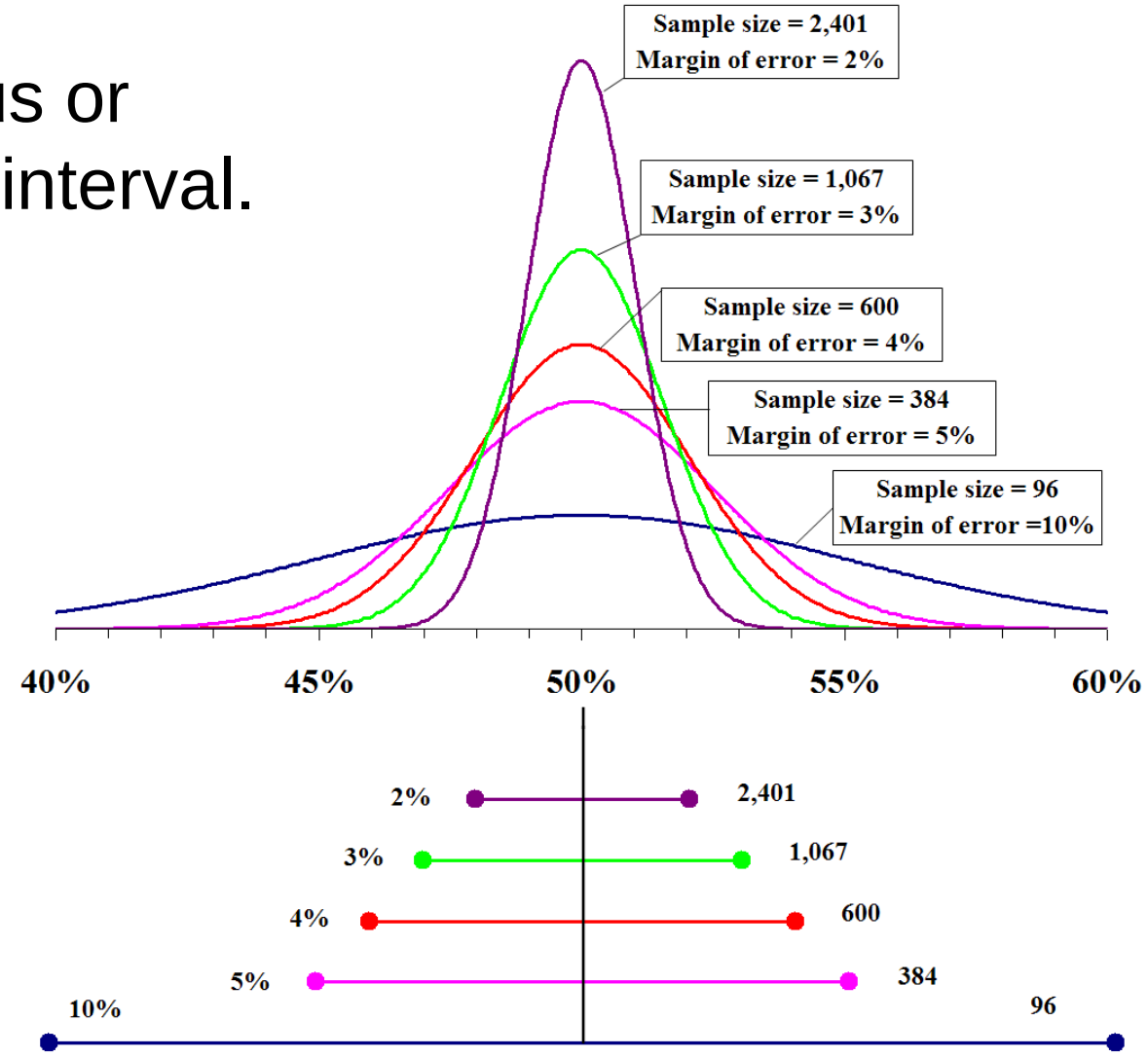
Margin of error is the **maximum expected difference** between the true population parameter and a sample estimate of that parameter.



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

Confidence Intervals - Summary

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 $1 - p_s$	$(p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}})$

Confidence Intervals (for Proportions) in Healthcare Industry

You studied a sample of 400 pregnant women and found that in the sample, the proportion of anemic women is 0.53. Construct a 99% confidence interval for the proportion of anemic women in the population of pregnant women.

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

$$\left(p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}}\right)$$

$$0.53 - 2.58 * \sqrt{\frac{0.53 * 0.47}{400}} < \boldsymbol{p} < 0.53 + 2.58 * \sqrt{\frac{0.53 * 0.47}{400}} = \mathbf{0.466 < p < 0.594}$$

Half of all pregnant women are anaemic

▶Continued from P 1

The WHO defines wasting as low weight for height, stunting as low height for age, and underweight as low weight for age.

The survey also found that just over half of all pregnant women were anaemic. This would automatically translate into their newborn being weak. Overall, 53% of women and 23% of men in the 15-49 age group were anaemic.

There is wide variation among states. The data for UP has not been released in view of the ongoing polls, according to Balram Paswan, professor at Mumbai-based International Institute for Population Sciences which was the nodal agency for the survey done for the health ministry. But poorer states like Bihar, Madhya Pradesh, Jharkhand, Assam, Rajasthan and Chhattisgarh have higher than national average rates on all markers.

More advanced states like those in the south, Haryana and Gujarat have slightly better numbers but are still at unacceptable levels. In Tamil Nadu, 51% children are anaemic while in Kerala it is over one-third. In many states, stunting has declined but the share of severely wasted children has increased. These are clear signs of an endemic crisis of hunger in the country that policy makers don't appear to be addressing.

HUNGER CRISIS: Overall, 53% of women and 23% of men in the 15-49 age group were anaemic

Confidence Intervals for Means in Healthcare Industry

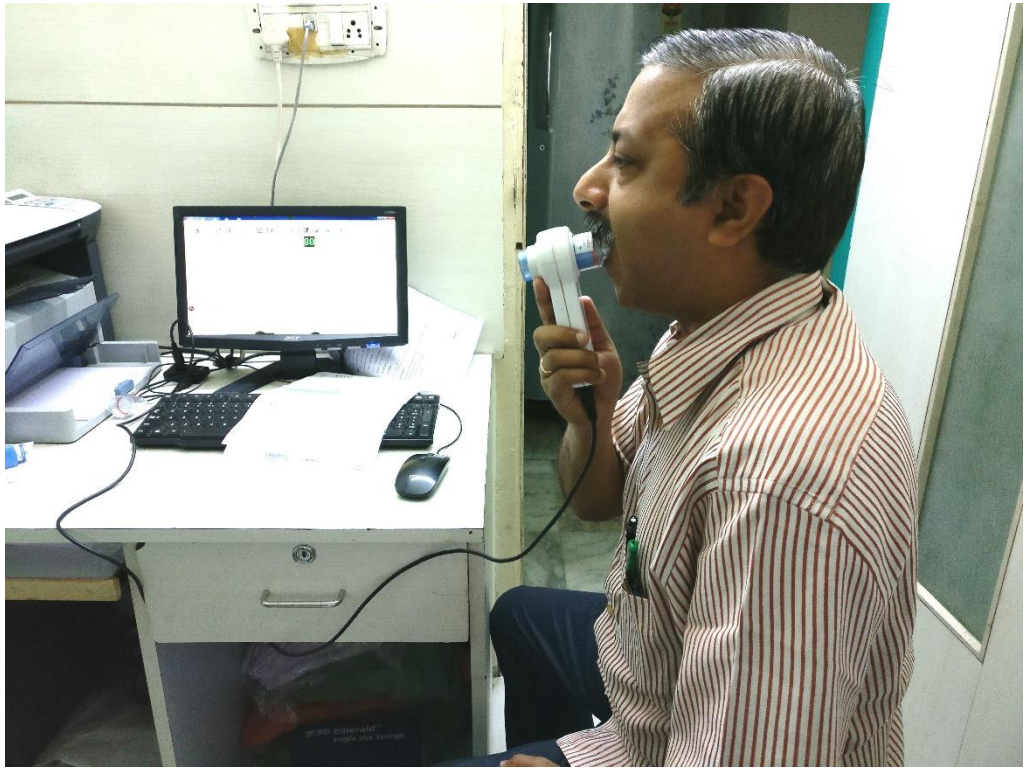
The lung function in 57 people is tested using FEV1 (Forced Expiratory Volume in 1 Second) measurements. The mean FEV1 value for this sample is 4.062 litres and standard deviation, s is 0.67 litres. Construct the 95% Confidence Interval.

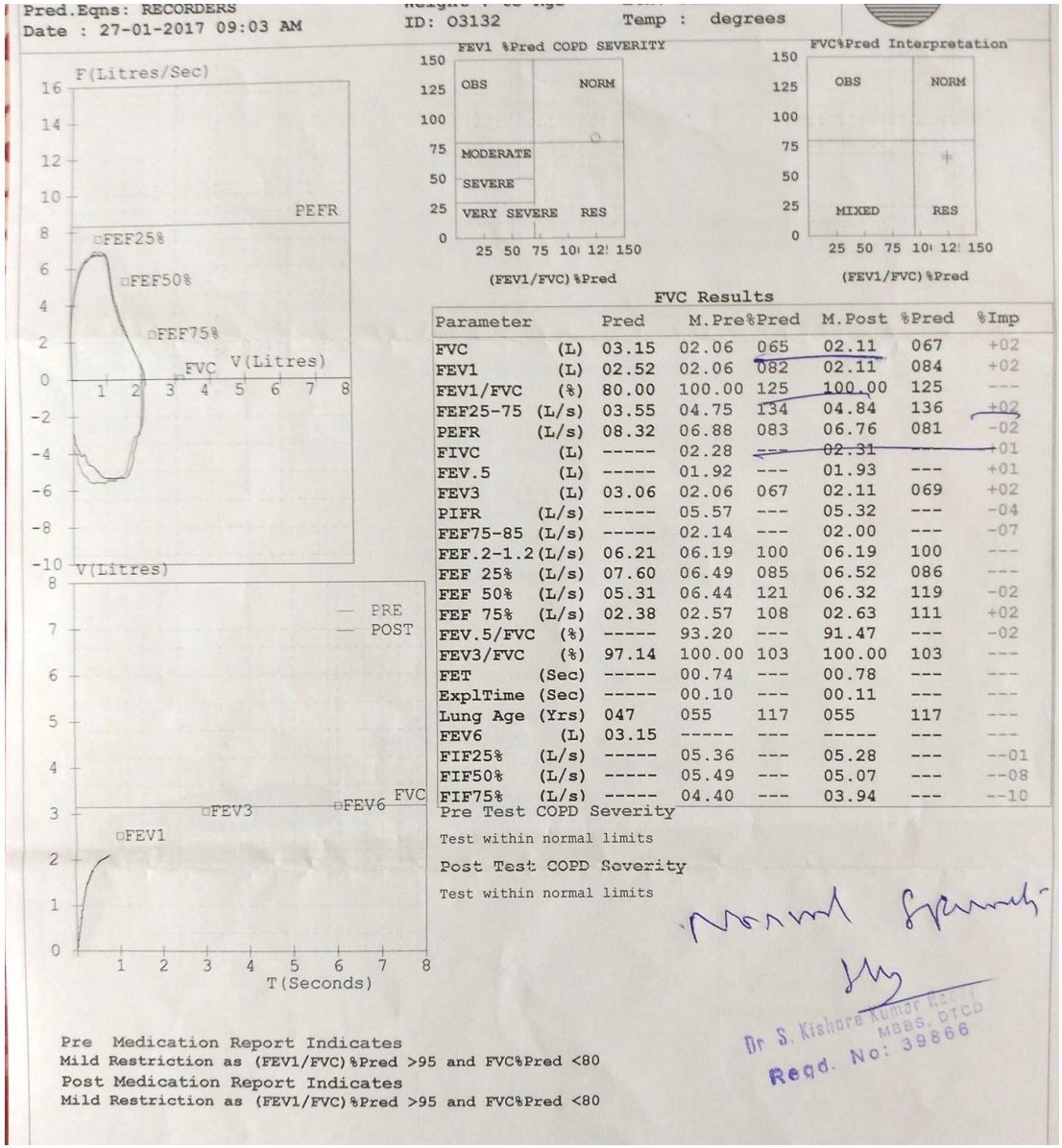
FEV1 values of 57 male medical students

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			

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

Confidence Intervals in Healthcare Industry



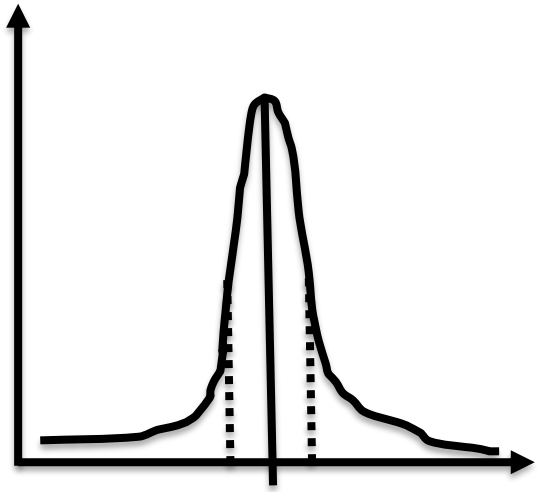


Confidence Intervals in Healthcare Industry

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

$$95\%\ CI: \left(4.062 - 1.96 * \frac{0.67}{\sqrt{57}}, 4.062 + 1.96 * \frac{0.67}{\sqrt{57}}\right)$$
$$= (3.89, 4.23)$$

Attention Check



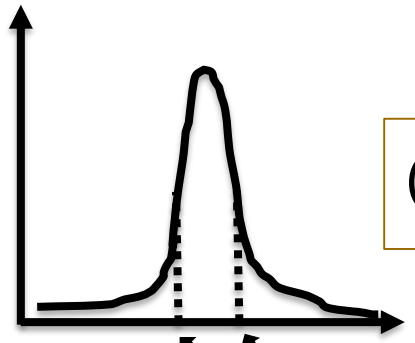
What happens to confidence interval as confidence level increases?

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

What happens to the confidence interval as sample size increases?

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

Connecting the Dots



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)

Depends on the confidence level

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

If you draw a random sample many times, a certain percentage of the confidence intervals will contain the population mean.

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

Source : https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-confidence-intervals-and-confidence-levels

Interview Question – Google US

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?

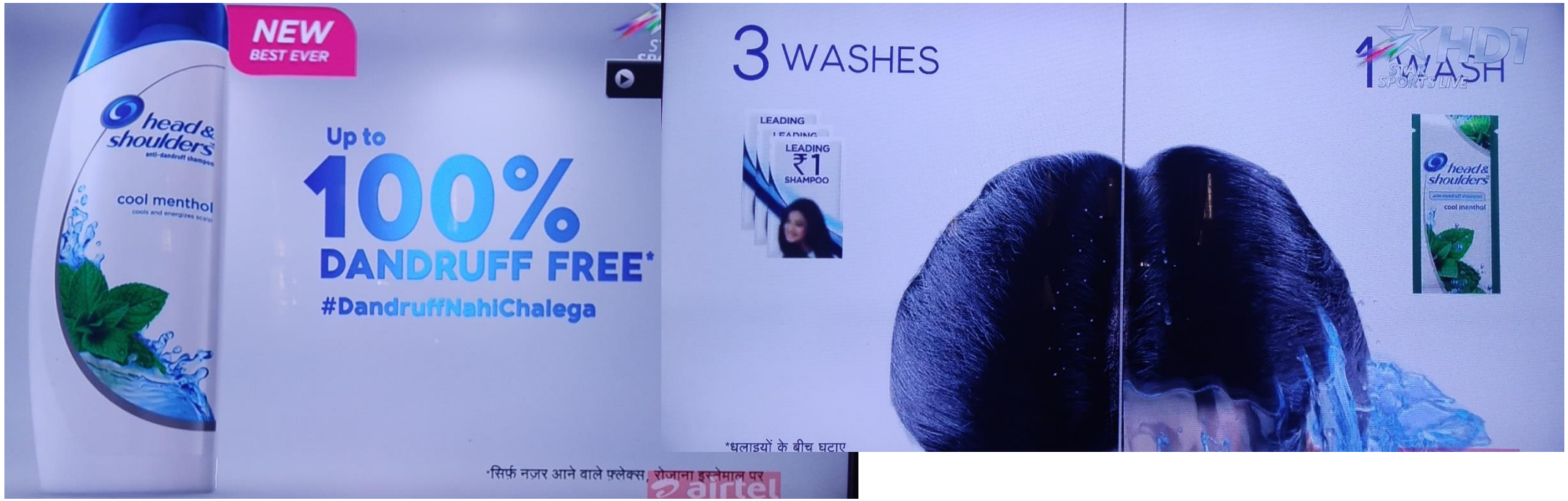
A. $P(X \geq 15) = P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20) = 0.021$

pbinom(14, 20, 0.5, lower.tail = FALSE, log.p = FALSE)

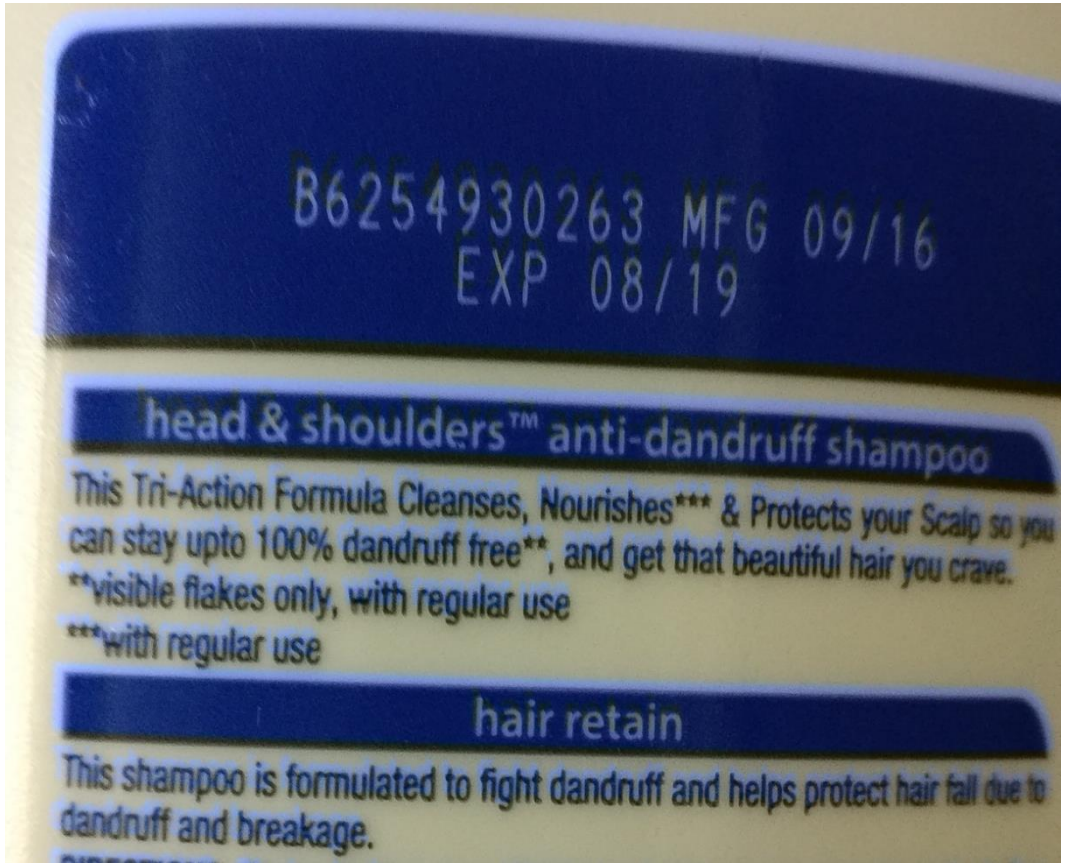
INFERENTIAL STATISTICS

HYPOTHESIS TESTS

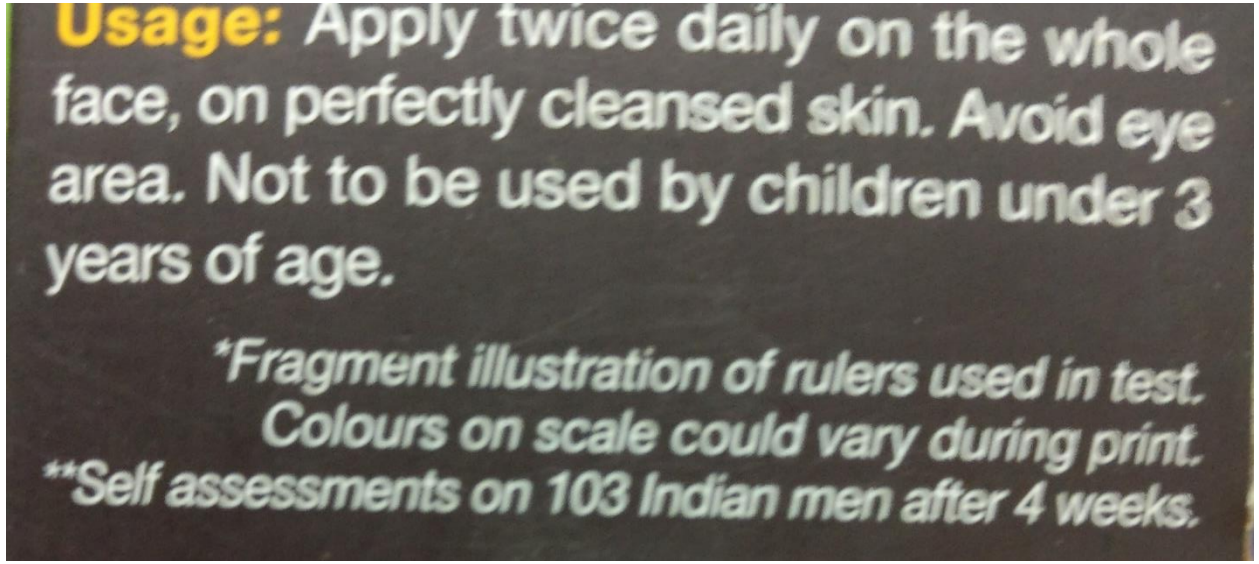
Hypothesis tests give a way of using samples to test whether statistical claims are likely to be true or not.



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

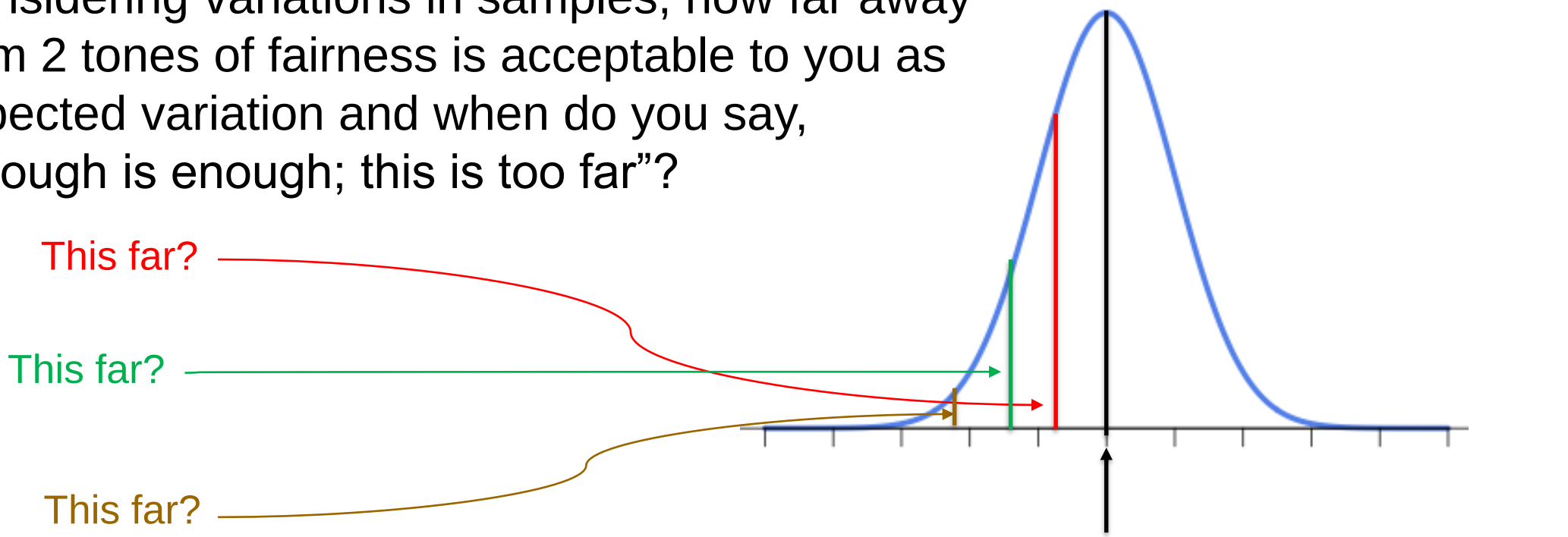


Hypothesis tests give a way of using samples to test whether statistical claims are likely to be true or not.



Hypothesis Testing Process

Considering variations in samples, how far away from 2 tones of fairness 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

Garnier Men PowerWhite improves fairness by 2 within 4 weeks.

This is called Null Hypothesis and is represented by H_0.

In this case, H_0: Tone = 2

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.

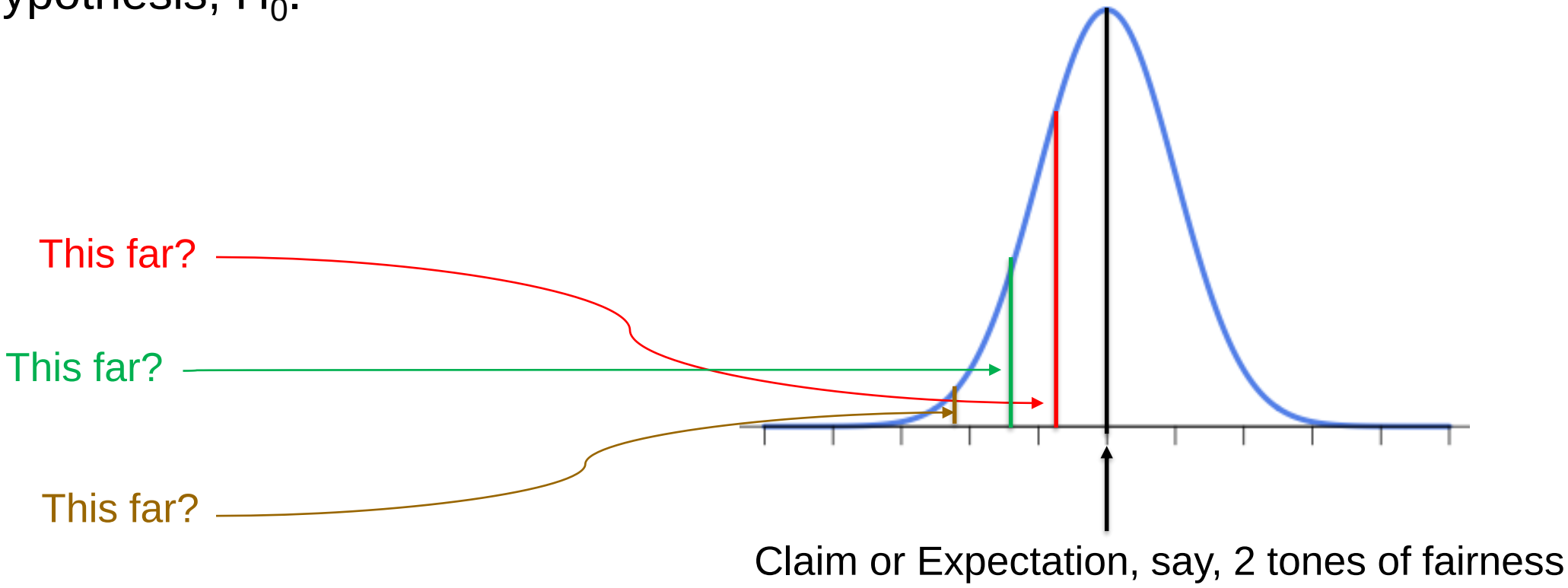
In this case, H_1: Tone < 2

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: Specify the critical region

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 2: Specify the critical region

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

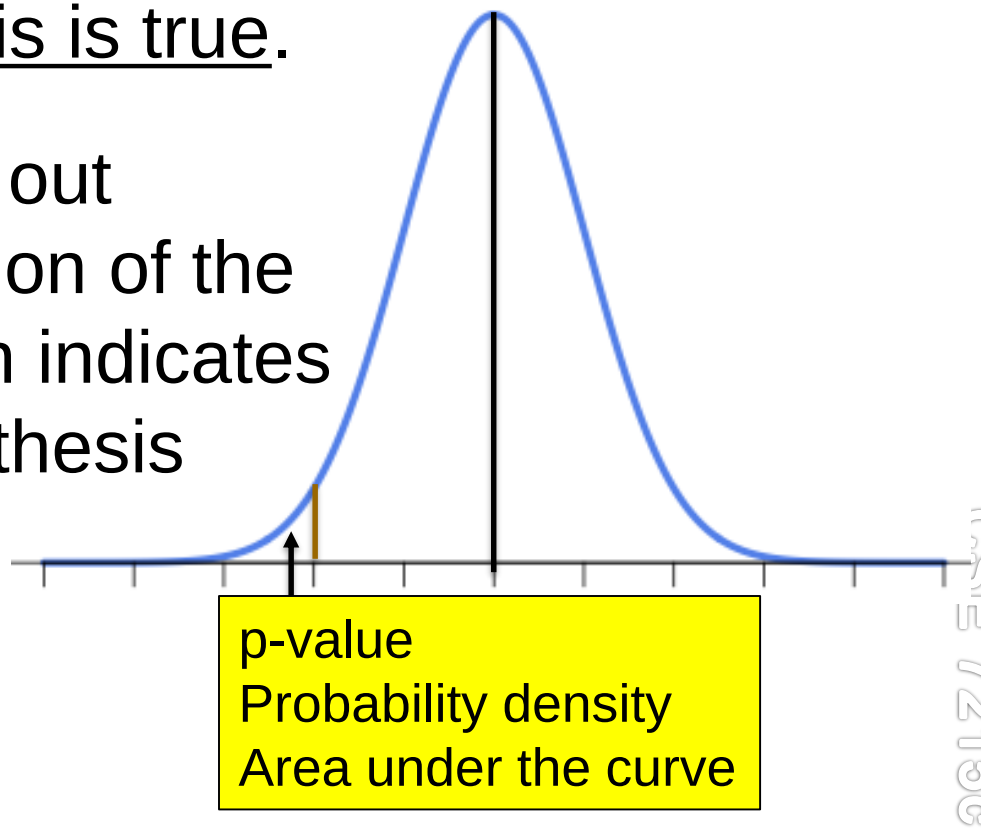
Recall that in a 95% CI, a 5% of the CI's of the samples will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 2 tones of fairness increase, is rejected.

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

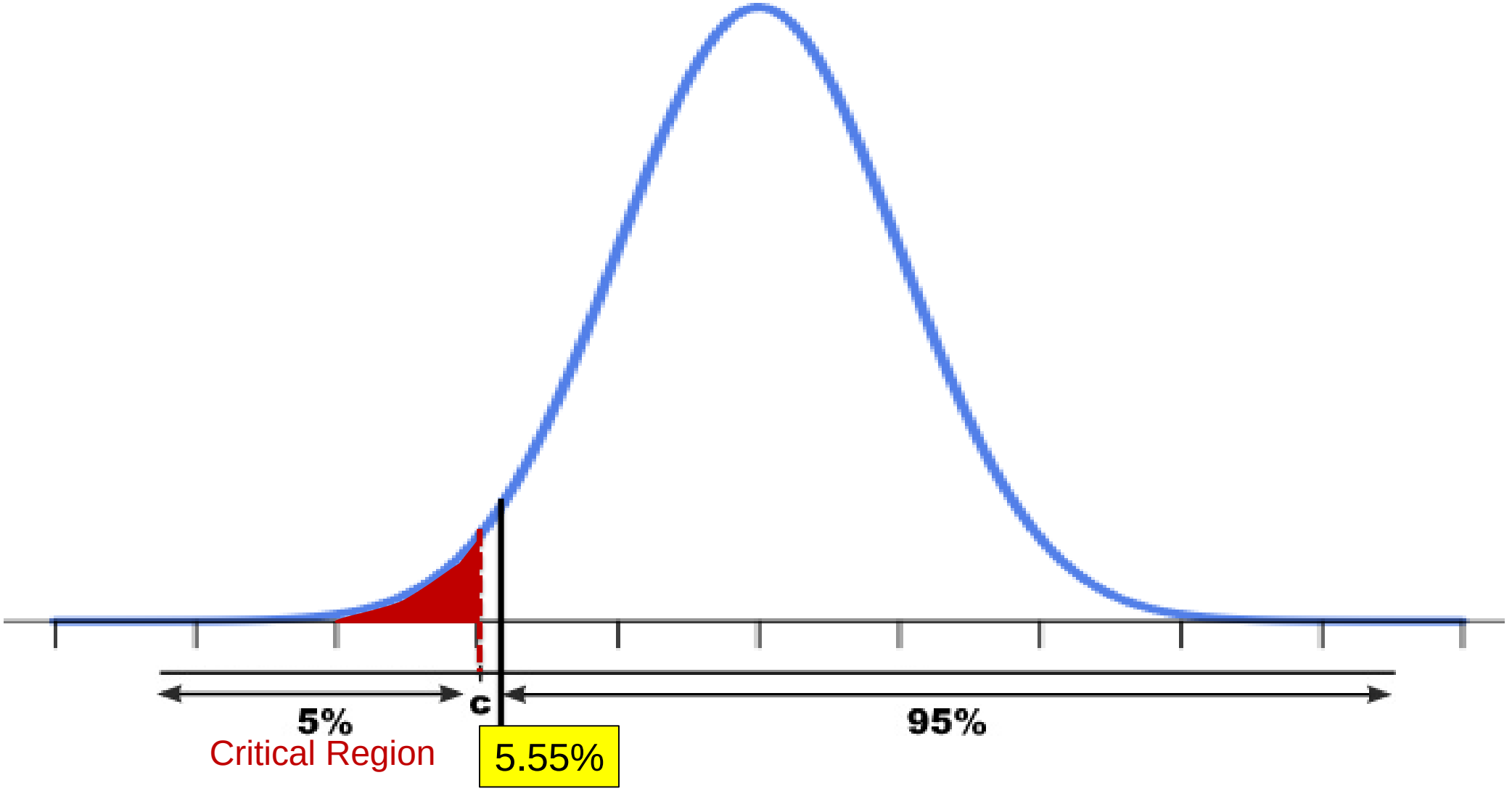
Step 3: Find the *p*-value

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 4: Is the sample result in the critical region?



Step 5: Make your decision

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

Attention Check

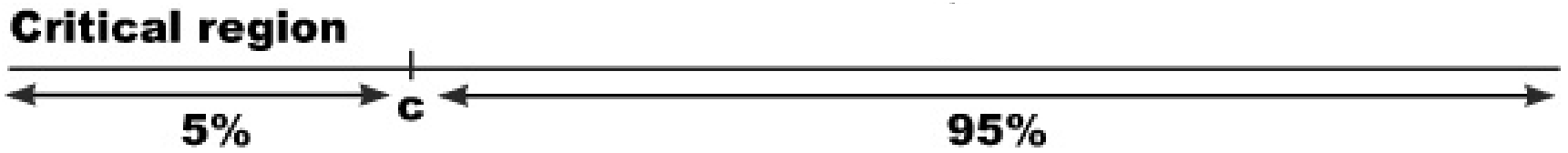
In hypothesis testing, do you assume the null hypothesis to be true or false?

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.

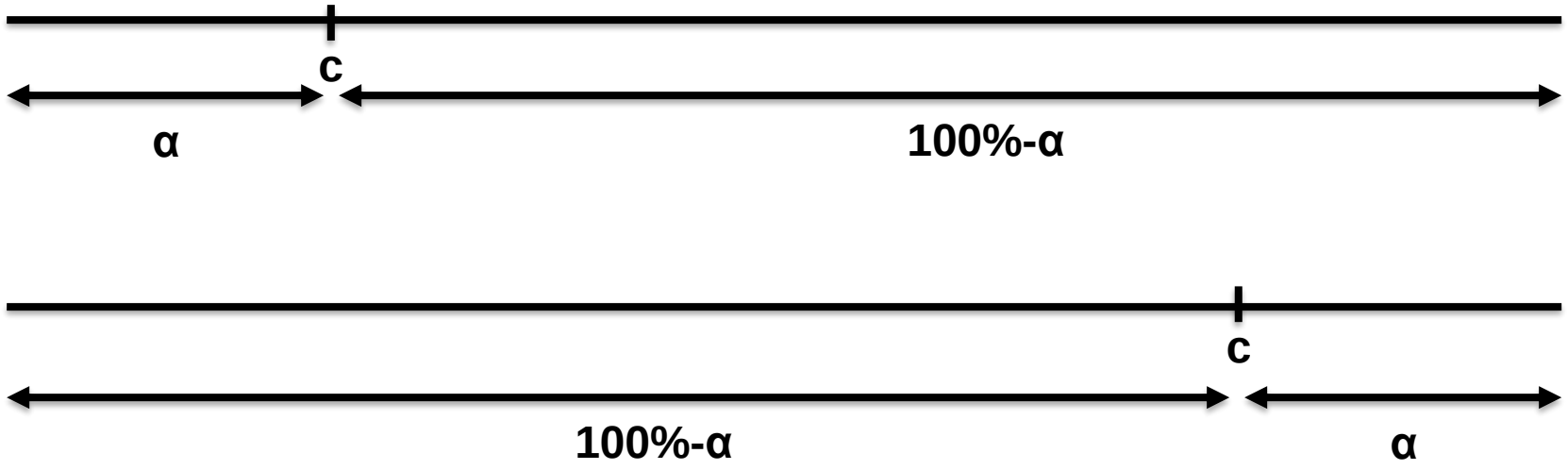
Do you need weaker evidence or stronger to reject the null hypothesis if you were testing at the 1% significance level instead of the 5% significance level?

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.

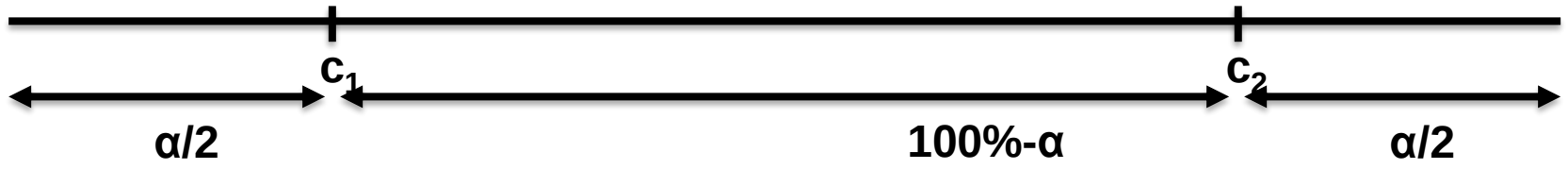


Critical Region Up Close

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.

- Garnier Men PowerWhite hypothesis test as discussed till now.
  One-tailed/Lower-tailed

- If we were checking whether significantly more or significantly less than 2 tones fairer result was achieved, i.e., H_1: Fairness tone improvement $\neq 2$.
  Two-tailed test

- The coin is biased.
  Two-tailed test

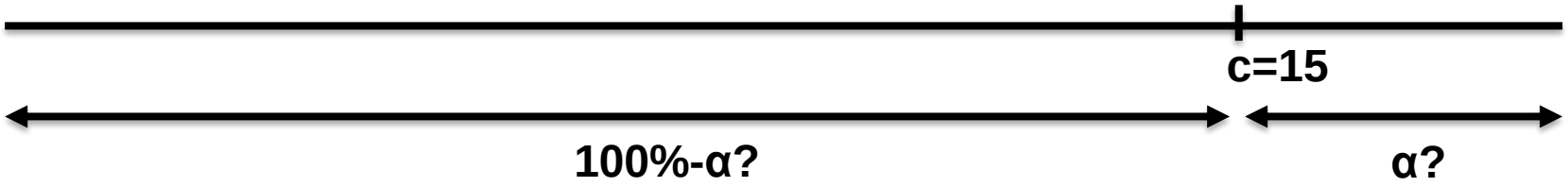
- The coin is biased towards heads with probability 0.8.
  One-tailed/Upper-tailed

The Missing Link in the Google 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?

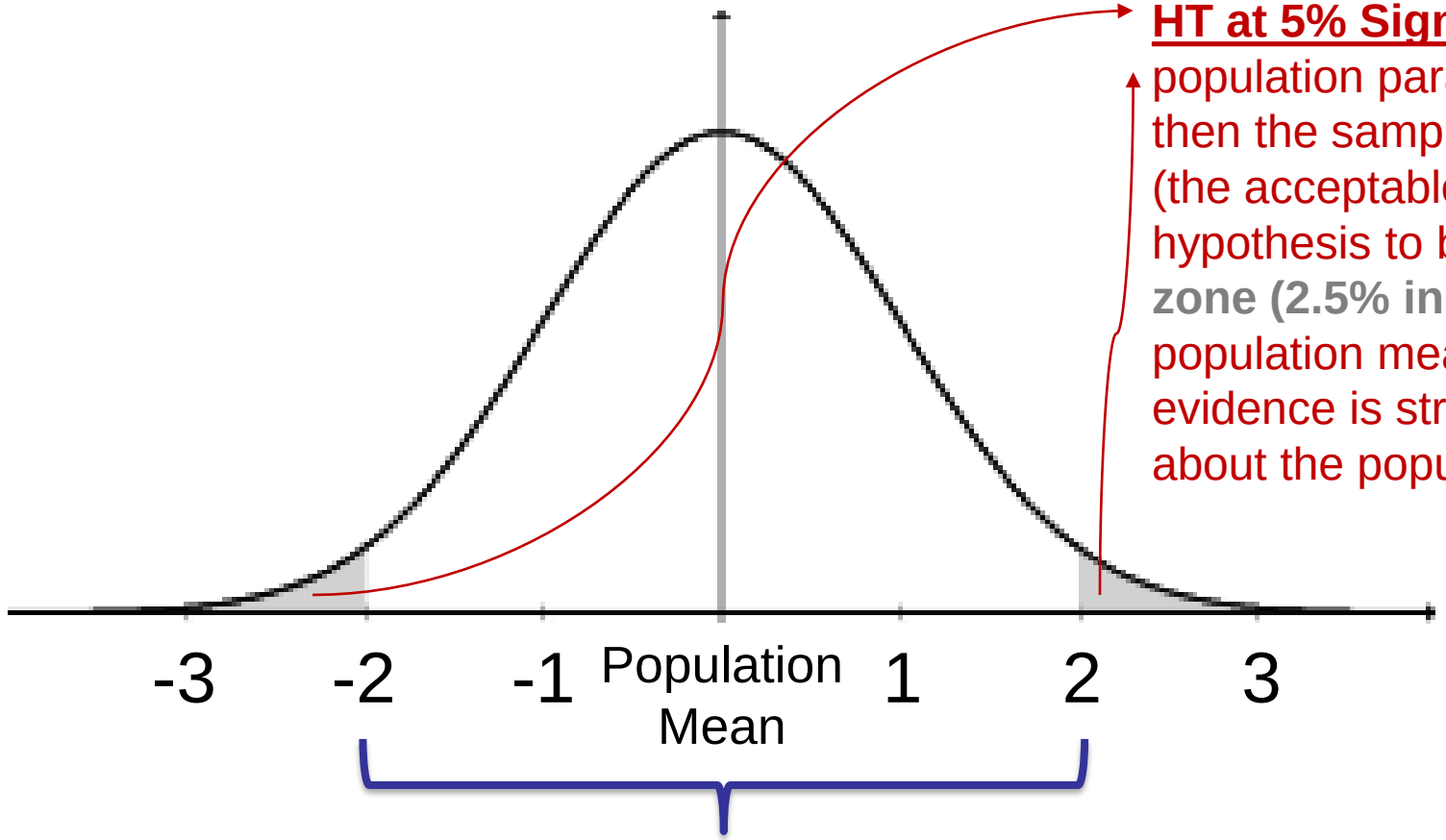


At 5 % level of significance, we reject the NULL hypothesis (that the coin is fair), At 1%, we fail to reject the NULL hypothesis(that the coin is fair)

The hypothesis test doesn’t answer the question whether the company is telling the truth or not or if Garnier Men PowerWhite really works or not or if 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.

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

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)

PUNE

Kirloskar - Pune
S. L. Kirloskar Center for Executive Education,
Kirloskar Corporate Office, 8th Floor,
Cello Platina, Model Colony, Shivaji Nagar – 411005

MUMBAI

Kanakia Wall Street, 4th Floor, Andheri-Kurla Road
Chakala, Andheri East, Mumbai - 400093

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