Why Study Statistics?

Statistics are part of our daily life and are all around us…

…They are also used and misused heavily…

…and so, we must study and understand Statistics.

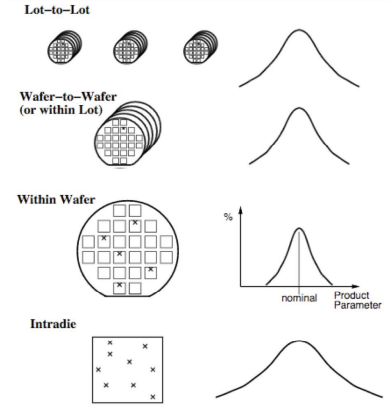
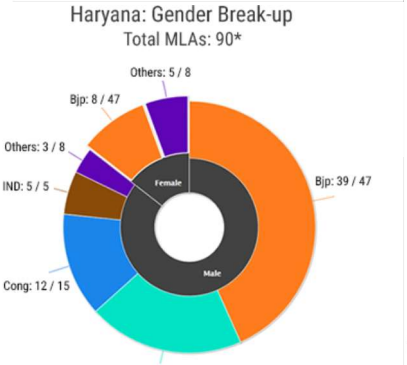


Figure 1. Spatial and temporal variation scales.



INDIA BEATS CHINA, GROWS 7.5% IN Q4

■ **Economists doubtful; India Inc sees downside risks**

PAWAN BALI | DC
NEW DELHI, MAY 29

India's GDP accelerated in January to March 2015 period (fourth quarter of 2014-15) to 7.5 per cent, overtaking China as the fastest growing major economy in the world.

This took the overall GDP growth for FY 2014-15 to 7.3 per cent against 6.9 per cent in the 2013-14.

Finance minister Arun Jaitley said that the economy is clearly on "a recovery path". He said an economy growing at fastest pace in the world cannot be 'fragile' as alleged by the former Prime Minister Manmohan Singh.

Finance ministry said that those sectors within control of policy — manufacturing and services — improved substantially while those dependent on factors beyond the policy control such as agriculture (dependent on weather) and exports (on foreign demand), "did less well."

However, the high GDP numbers have come on the back of new methodology which the Central Statistical Organisation (CSO) adopted earlier this year to calculate the GDP.

In January, CSO revised India's growth to 6.9 per cent in the 2013-14 from the earlier estimate of 4.7 per cent.

Before this data revision, it was a dominant view that India had seen worst growth in 2013-14.

Some economists said that the latest GDP numbers released on Friday does not reflect the ground reality and people will now have to adjust that 7 per cent plus growth under the new methodology is not the same what it was based on older formula.

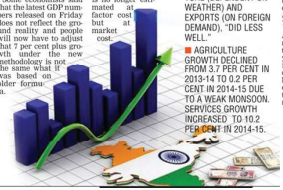
While the GDP has grown at a fast pace the corporate earnings are dismal, industrial activity is weak, private investment is yet to start and bank credit uptake is slow. Rating agency, India ratings, principal economist and director (public finance) Sunil Kumar Sinha said that the people need to recognise that methodology to calculate GDP has undergone a change and it is no longer estimated at factor cost but at market cost.

HIGHLIGHTS

■ INDIA'S GDP ACCELERATED TO 7.5 PER CENT. (VIVEK JAKINKI CHINA AS THE FASTEST GROWING MAJOR ECONOMY IN THE WORLD.

■ FACTORS BEYOND THE POLICY CONTROL SUCH AS AGRICULTURE (DEPENDENT ON WEATHER) AND EXPORTS (ON FOREIGN DEMAND), "DID LESS WELL."

■ AGRICULTURE GROWTH DECLINED FROM 3.7 PER CENT IN 2013-14 TO 0.2 PER CENT IN 2014-15 DUE TO A WEAK MONSOON. SERVICES GROWTH INCREASED TO 10.2 PER CENT IN 2014-15.



GOOD NEWS

Cholesterol not a threat

■ **US to remove high-cholesterol food from 'naughty' list**

Washington, May 26: United States officials have finally given the green light for a U-turn on previous warnings on cholesterol, which has been on the "naughty" list of nutrients for nearly 40 years. Health officials have been warning people to stay away from high-cholesterol foods since the 1970s to avoid heart disease and clogged arteries.

However, after a study, eggs, butter, full-fat dairy products, nuts, coconut oil and meat have now been classified as "safe" and have been officially removed from the "nutrients of concern" list, reported the *Washington Business Times*.

The US Department of Agriculture, which is responsible for updating the guidelines every five years, stated in its findings for 2015: "Previously, the Dietary Guidelines for Americans recommended that cholesterol intake be limited to no more than 300 mg/day. The 2015 DGAC will not bring forward this recommendation because available evidence shows no appreciable relationship between consumption of dietary cholesterol and serum (blood) cholesterol, consistent with the AHA/ACC (American Heart Association / American College of Cardiology) report. Cholesterol is not a nutrient of concern for overconsumption."

The Dietary Guidelines Advisory Committee will, in response, no longer warn people against eating high-cholesterol foods and will instead focus on sugar as the main substance of dietary concern.

The 70s, 80s and 90s were the 'non fat' years, with the US government warning people to limit the amount of high-cholesterol foods in their diets to avoid heart disease and strokes.

But nutritionists and scientists have long been campaigning for the U-turn, which started with introducing "good cholesterols" back into the "safe zone".

US cardiologist Dr Steven Nissen said: "It's the right decision. We got the dietary guidelines wrong. They've been wrong for decades."

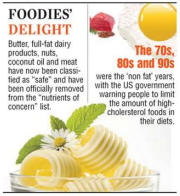
Dr Chris Masterjohn added: "When we eat more foods rich in this compound, our bodies make less. If we deprive ourselves of foods high in cholesterol — such as eggs, and butter — our body revs up its cholesterol synthesis." — *Agencies*

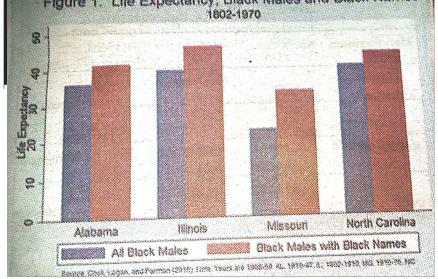
FOODIES' DELIGHT

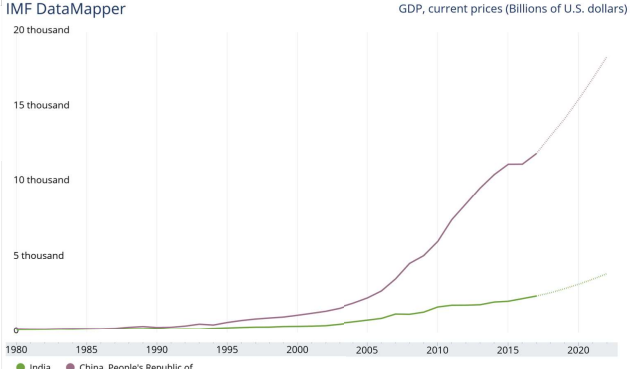
Butter, full-fat dairy products, nuts, coconut oil and meat have now been classified as 'safe' and have been officially removed from the "nutrients of concern" list.

The 70s, 80s and 90s

were the 'non fat' years, with the US government warning people to limit the amount of high-cholesterol foods in their diets.







Monsoon to be weak

■ **PMO in a crisis mode; asks MHA, Cab-secy for solutions**

DC CORRESPONDENT
NEW DELHI, JUNE 2

Already facing an agrarian crisis, there is bad news in store for India with a "deficient" monsoon predicted this year, raising fears of a drought in parts of the country.

The Prime Minister's Office has gone into overdrive, drawing up contingency plans till district level to deal with a drought, worried at a spiralling effect on all sectors.

The Cabinet Secretary and the MHA disaster management division have been tasked with putting in place detailed plans and out-of-the-box solutions to tackle crop losses.

Earth sciences minister Harsh Vardhan, giving the revised forecast for the 2015 monsoons, said on Tuesday: "I have to say with a heavy heart that as per our revised forecast, India will receive 88 per cent of rainfall of the Long Period Average, plus or minus four per cent."

The southwest monsoon is also delayed, and may now hit the Kerala coast on June 5.

Though the normal date for onset of monsoon over Kerala is June 1, the Indian Meteorological Department (IMD) had predicted that it will hit the southern state on May 30 this year.

June 1 also marks the official onset of rains in the country.

"Conditions are becoming favourable for the onset of southwest monsoon over Kerala around June 5," said an IMD statement.

Last year too, monsoon was delayed and hit the Kerala coast on June 6.

■ Another report inside

I have to say with a heavy heart that as per our revised forecast, India will receive 88% of rainfall of the Long Period Average, plus or minus four per cent. **HARSH VARDHAN, EARTH SCIENCES MINISTER**

CLIMATE HAS A NOSE FOR SHAPES

THE SHAPE OF OUR NOSES WAS FORMED BY A LONG PROCESS OF ADAPTATION TO CLIMATE, A STUDY SAYS

- **WIDER** noses are more common in warm-humid climates
- **NARROWER** noses are more common in cold-dry climates
- **NARROWER** nasal passages help to increase the moisture content of air and warm it — a bonus for those in higher latitudes

EVOLUTION OF THE NOSE

The width of the nostrils and the

ADAPTING MECHANISM

Narrower nostrils alter the airflow so the mucous-covered inside of the nose can humidify and warm the air

'Cancer may strike due to bad luck, not lifestyle'

Random Changes In DNA During Cell Division Cause Nearly Two-Third Of All Cancers In Humans, Finds Study

Subodh.Varma@timesgroup.com

Why does cancer strike some people and not others? New research shows that random changes or 'mistakes' in DNA when cells are dividing cause nearly two-third of all cancers in humans. These changes are neither caused by external factors like smoking or exposure to harmful chemicals, nor by hereditary factors. They are chance events occurring at the molecular level. In other words, cancer can strike anybody.

This upends prevailing wisdom that cancer is mostly a lifestyle related disease caused by external or environmental factors like smoking, harmful chemicals and conditions like obesity. While all these are valid and important risk increasing factors, random chance may be the real driver, if one goes by this new research.

Different types of cancers have different origins, according to the study. For example, in pancreatic cancers, 77% are due to random DNA copying errors, 18% to environmental factors, such as smoking, and the remaining 5% to heredity. In other cancer types, such as those of the prostate, brain or bone, more than 90% of the mutations are due to random copying errors.

Lung cancer is most likely to be caused by environmental factors, mostly smoking. About 65% of all the mutations are due to smoke and 35% due to DNA copying errors. Inherited factors have negligible role.The study involved a statistical analysis of cancer data from 69 countries including India, representing 4.8 billion people, more than half of the world's population. It was done by scientists from Johns Hopkins Kimmel Cancer Center at Baltimore, US, and published in the peer reviewed journal Science on March 24.

A MATTER OF CHANCE?

Human bodies grow by constant division of cells, starting from the first cell formed by fusion of the male sperm with the female egg. Every time a cell divides into two, the genetic code carrying DNA is copied. What the scientists are saying is that mistakes occur in this copying process that accumulate over time and ultimately cause cancer. "These copying mistakes are a potent source of cancer mutations that historically have been scientifically undervalued, and this new work provides the first estimate of the fraction of mutations caused by these mistakes," said the paper's lead author Cristian Tomasetti.

The researchers studied all 32 cancer types and estimated that 66% of cancer mutations result from copying errors, 29% can be attributed to lifestyle or environmental factors, and the remaining 5% are inherited. They found no strong correlation between cancer incidence and normal cell divisions across different cancer types, regardless of the countries' environment or stage of economic development. This means that lifestyle factors like smoking or exposure to toxic chemicals are also very important factors causing nearly a third of cancers.

"We need to continue to encourage people to avoid environmental agents and lifestyles that increase their risk of developing cancer mutations," co-author Bert Vogelstein emphasised.

Pvt research plays down risk from mobiles: Study

DurgeshNandan.Jha
@timesgroup.com

New Delhi: Is radiation from mobile phones harmful? Multiple studies globally have not conclusively reached an answer. But an analysis by AIIMS of all research on the subject has found an interesting pattern — government funded studies show increased risk of brain tumour on long-term exposure to mobile phone radiation while industry-funded research tends to underestimate the risk.

"We found that industry-funded studies are not of good quality and tend to underestimate the risk. Govern-

PHONE TROUBLE

- 1.33 times higher risk of brain tumour due to mobile use, according to average of all studies
- Mixed-funded (govt, industry, mobile makers) research says risk much lower, at 1.05 times
- Studies funded by phone industry and mixed sources have low quality score of 5-6
- Score of govt-funded research higher at 7-8

creased risk of brain tumour on long-term exposure," said Dr Kameshwar Prasad, head of neurology at AIIMS, who is lead author of the study.

▶Brain tumour risk, P 12

Alcohol consumption can attract mosquitoes

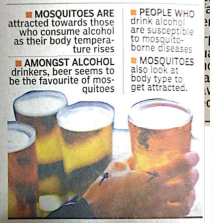
DC CORRESPONDENT
HYDERABAD, JAN. 9

Alcohol consumption can attract mosquitoes as the body odour is strong, according to a study published by the American Mosquito Control Association.

The study found that those who had taken alcohol had a larger number of mosquitoes attracted to them. It was found that those who have been attracts more mosquitoes because of the smell emitted from the body is stronger. It was found that drinking beer increased ethanol content in the sweat which lures mosquitoes. The other reason was intake of alcohol increased the body temperature and the heat makes the mosquito rush towards these people.

Entomologist Mr K. Venkatesh said, "Studies have shown that those who eat a lot of sugar and sweets have more and more mosquitoes attracted to them. It is the body odour which attracts the mosquitoes. The alcohol is mostly fermented, hence the body odour is strong."

- MOSQUITOES ARE attracted towards those who consume alcohol as their body temperature rises
- AMONGST ALCOHOL drinkers, beer seems to be the favourite of mosquitoes
- PEOPLE WHO drink alcohol are susceptible to mosquito-borne diseases
- MOSQUITOES also look at body type to get attracted.



Researchers stated that their experiments showed that those who consumed alcohol had mosquitoes going towards them. This means that the alcohol drinkers will be the first target of mosquitoes.

The increasing burden of mosquito-borne diseases like malaria, dengue and zika is making researchers study body odour which attracts mosquitoes.

Dr G H Kishan, senior general physician, added, "The cities have become breeding ground for mosquito. To add to this, there are a good number of human bodies that create the right odour to attract them. Due to this reason, the spread is large and wide in urban areas."

How hand sanitisers can harm children

New York: Scientists have warned that hand sanitisers might do more harm than good. They have found these alcohol-based, scented products might tempt youngkids to swallow the substance — leading to stomach pain, nausea, apnea and even coma.

Researchers from US Centres for Disease Control (CDC) and Prevention have identified serious consequences, including apnea, acidosis and coma in young children who swallowed alcohol-based hand sanitisers.

To characterise paediatric alcohol hand sanitiser exposures, researchers from US Centres for Disease Control and Prevention analysed data reported by poison centres among children aged 12 years during 2011 to 2014.

The study found majority of intentional exposures to alcohol hand sanitisers occurred in children aged 6-12 years. During 2011-2014, 79,669 hand sanitiser exposures in children aged 12 years were reported, of which 65,293 (82%) were alcohol exposures and 5,376 (6%) were non alcohol exposures. These data also indicate that, among older children, exposures occur less frequently during the summer months. PTI

Half of all pregnant women are anaemic

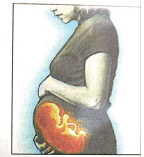
▶Continued from P 1

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

Women earned 25% less than men in 2016: Study

TIMES NEWS NETWORK

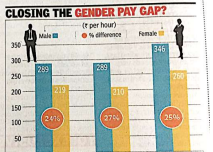
Mumbai: The average gender pay gap in India may have narrowed by two percentage points since last year, but it still stands at a yawning 25%. According to Monster Salary Index (MSI) data from 2016, men earned a median gross hourly salary of Rs 346, while women earned only Rs 260 in 2016. While it is an improvement over last year's over 27%, the pay gap is a notch higher than 24% in 2014. In the manufacturing sector, the average gender pay gap is as high as 30%.

Sanjay Modi, managing director (APAC & Middle East), Monster.com, said, "In India, the gender pay gap story holds true and the overall gap across India Inc is at 25%. This primarily is a manifestation of the underlying diversity challenges that organisations currently face."

According to the survey, which aimed at understanding the working women of India and their workplace concerns, nearly 44% women feel that gender parity need not be a top priority for their organisations. On the other hand, almost 69% expressed that even if gender parity is a priority, the management "does not walk the talk", indicating the need for India Inc to step up and implement pragmatic policies to bridge the gap.

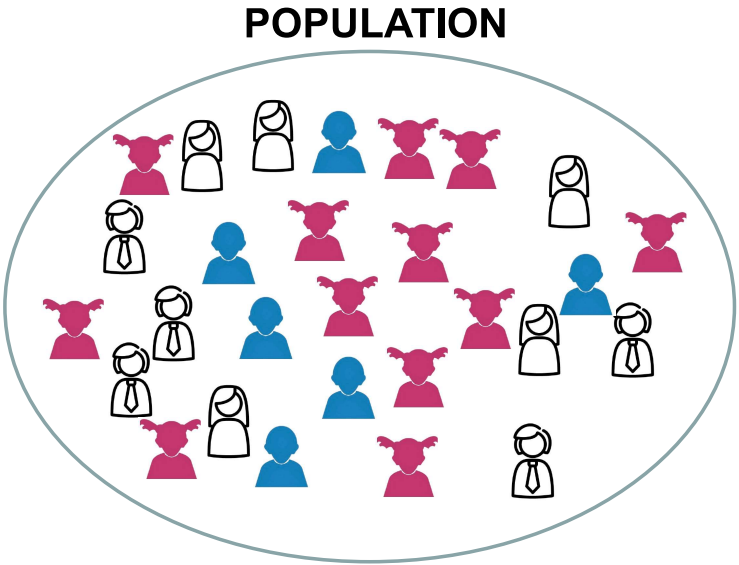
"Unexamined conventions of women's commitment to work, distractions of family commitments, societal perception of women who work for long hours, etc, are the challenges that constrain women's progress. Sensitisation towards gender diversity and a step towards examining our socialised preconceptions would help break barriers to a more fair and inclusive work environment," said Modi.

CLOSING THE GENDER PAY GAP?

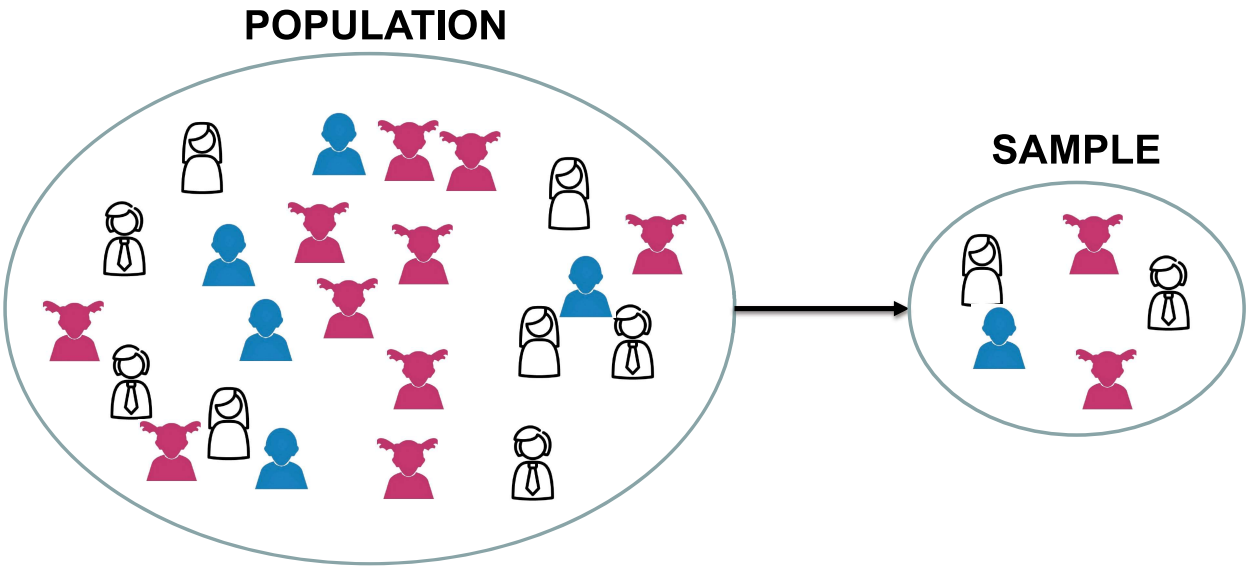


BASIC STATISTICAL TERMINOLOGY

Population and Sample



Population and Sample



Descriptive and Inferential Statistics

- Descriptive Statistics – Data gathered about a group to reach conclusions about the same group.
- Inferential Statistics – Data gathered from a sample and the statistics generated to reach conclusions about the population from which the sample is taken. Also known as Inductive Statistics.

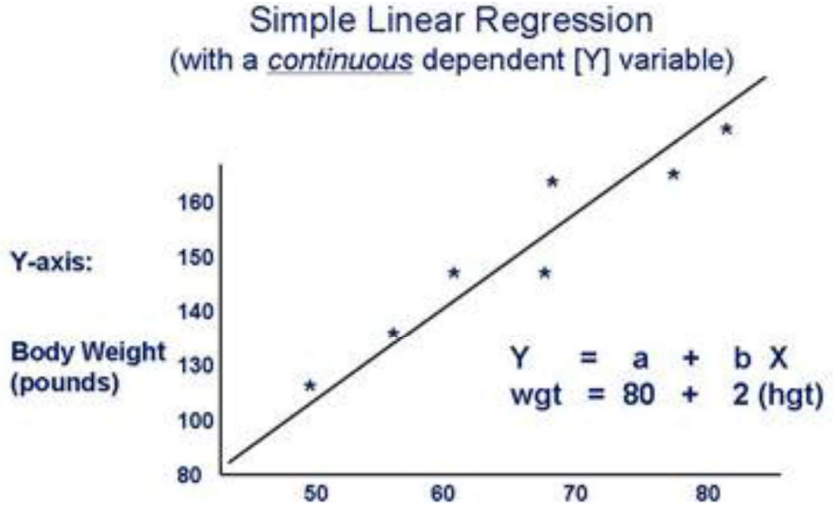
Source: http://www.arogyaworld.org/wp-content/uploads/2010/10/ArogyaWorld_IndiaDiabetes_FactSheets_CGI2013_web.pdf

Variables and Data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no
32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no
25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no
40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no

Variables – Dependent and Independent

- Dependent variables on y-axis and Independent on x-axis.
- Dependent variable also called Target variable or Class variable.



Data – Numeric and Categorical

18 kg

27 kg

Sources: http://banglanews24.com/en/files/2013August/SM/Gold-sm20130830024804.jpg,

Categorical Data (Qualitative)

Nominal

Examples

- Employee ID
- Gender
- Religion
- Ethnicity
- Pin codes
- Place of birth
- Aadhaar numbers

Ordinal

Examples

- Mutual fund risk ratings
- Fortune 50 rankings
- Movie ratings

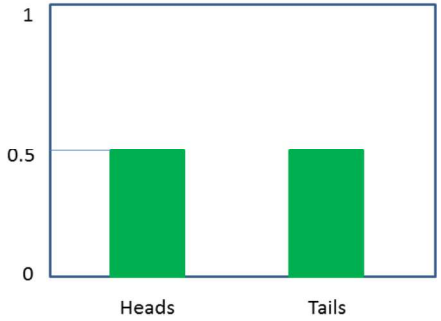
While there is an order, difference between consecutive levels are not always equal.

Numeric Data (Quantitative)

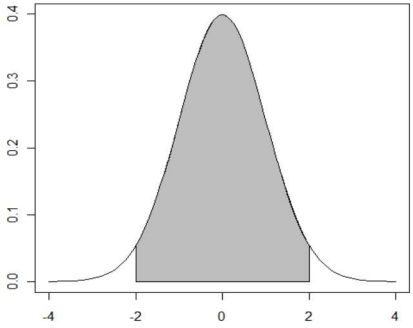
Examples

- Height
- Weight
- Time
- Volume
- Number of iPads sold
- Number of complaints received at the call centre
- Number of employees
- Percentage return on a stock
- Rupee change in stock price

Discrete and Continuous



Countable



Measurable

Discrete or Continuous?

Time between passenger arrivals at the airport	Continuous
No. of passengers arriving at the airport during a five-minute period	Discrete
Sampling 100 voters in an exit poll and determining how many voted for the winning candidate	Discrete
No. of defects in a batch of 50 items	Discrete

DESCRIBING DATA THROUGH STATISTICS

A Central Tendency - Mean

In 2015, American Airlines President Scott Kirby mentioned that 87% of their passengers fly the airline once or less per year but account for more than half of the airline's revenues.

Ticket Price ($)	90	100	110	500	600	700
Frequency, f	40	37	10	3	5	5

Flown once | Flown more than once

- **Why does American Airlines match the fares of ultra low-cost carriers such as Spirit Airlines?**

About 87 percent of American's passengers fly the airline once or less per year, and those passengers account for more than half of the airline's revenue, Kirby said. In essence, the airline has to match budget carriers like Spirit to remain competitive, he said.

"If 50 percent of our customers are up for grabs, we can't walk away from that side of the business," Kirby said.

$$\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{40*90+37*100+10*110+3*500+5*600+5*700}{40+37+10+3+5+5} = \$164$$

Source: https://www.bizjournals.com/dallas/blog/2015/10/american-airlines-ceo-doug-parker-says-carriers.html and https://www.mckinsey.com/industries/travel-transport-and-logistics/our-insights/how-airlines-can-gain-a-competitive-edge-through-pricing

Another Central Tendency - Median

Source: http://www.business-standard.com/article/companies/ambani-gets-205-times-ril-s-median-

Revenue Management in Airlines Industry

Willingness to pay. Collecting and crunching data about customers, airlines understand passengers' tastes and behavior well enough to offer them transportation options they prefer and, more important, are ready to spend money on. So, revenue managers start from measuring willingness to pay (WTP). Willingness to pay reveals *"when"* a customer is likely to pay *"a maximum price"* for a product or service, explains the data scientist. *"It's assumed that customers are ready to pay more when there is less time before departure time. And society finds this pricing fair. WTP in the airline industry, therefore, depends on the day before departure (DBD). In practice, specialists define median WTP — a price that 50 percent of customers would like to pay for a ticket on a specific DBD. Such WTP is equivalent to price elasticity* (the number of passengers that would buy a ticket if a price drops by a certain percent) *with some assumptions between market demand and supply."*

This metric is connected to *dynamic pricing* — the practice of pricing a product based on a specific customer's willingness to pay. The calculation of WTP requires selecting data correctly. Revenue management can combine similar markets and, alternatively, distinguish high and low seasons, as well as holidays and weekends.

"Approaches to this type of statistical analysis were developed nearly 10 years ago. These days, it's easier to conduct research and present its results thanks to the development of data science and visualization capabilities. Considering that each

Source: https://www.altexsoft.com/blog/datascience/7-ways-how-airlines-use-artificial-intelligence-and-data-science-to-improve-their-operations/

Another Central Tendency - Median

Median: Arrange data in increasing order and find the mid-point $\frac{(n-1)}{2} + 1$.

Ticket Price ($)	90	100	110	500	600	700
Frequency, f	40	37	10	3	5	5

There are 100 data points. So, the median is the 50.5^{th} data point or the average of the 50^{th} and 51^{st} data points. When these 100 ticket prices are arranged in ascending order, both the 50^{th} and the 51^{st} values are \$100, and hence their average is also \$100.

Median is not affected by outliers, whereas Mean gets pulled towards the outliers. However, Mean incorporates all values in its computation whereas Median is simply the value of the mid-point in sorted data.

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

Mean = Median = 10 for all 3 players.

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

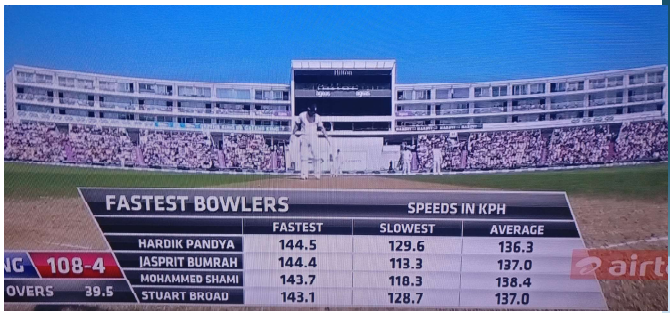
Measuring Spread and Variability

Range = Max - Min

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1



Measuring Spread and Variability

Exclude outliers scientifically – Quartiles*

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Lower quartile (25th percentile, Q1) = $\left(\frac{1*(n-1)}{4}+1\right)^{th}$ = 3.5th point = 6.5

Middle quartile (50th percentile, Q2) = Median = $\left(\frac{2*(n-1)}{4}+1\right)^{th}$ = 6th data point = 10

Upper quartile (75th percentile, Q3) = $\left(\frac{3*(n-1)}{4}+1\right)^{th}$ = 8.5th data point = 10.5

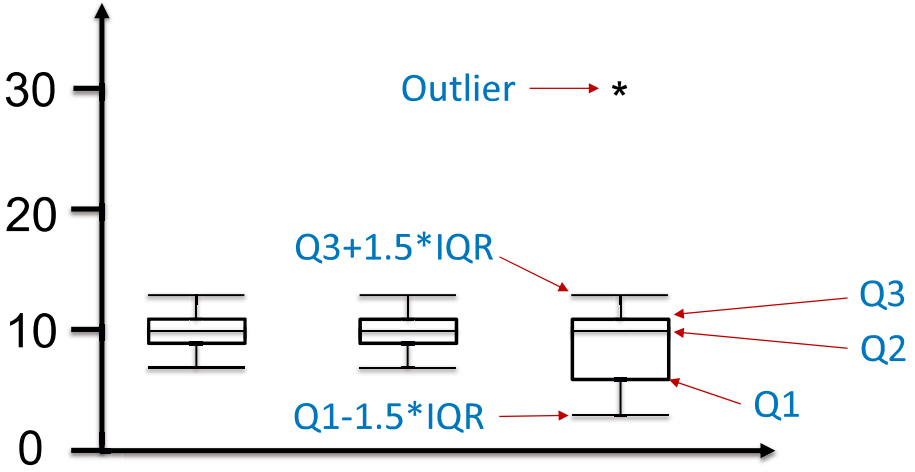
Interquartile range, IQR = Q3-Q1 (central 50% of data)

**9 different methods coded in R for Quantile calculations. No universal acceptance of a standard method. Let us only focus on what quantiles mean and how to use them.*

Measuring Spread and Variability

Exclude outliers scientifically – Quartiles

Box and whisker diagram or Box plot



3 3 6 7 7 10 10 10 11 13 30

Q1 = 6.5

Q2 = 10

Q3 = 10.5

IQR = 10.5 – 6.5 = 4

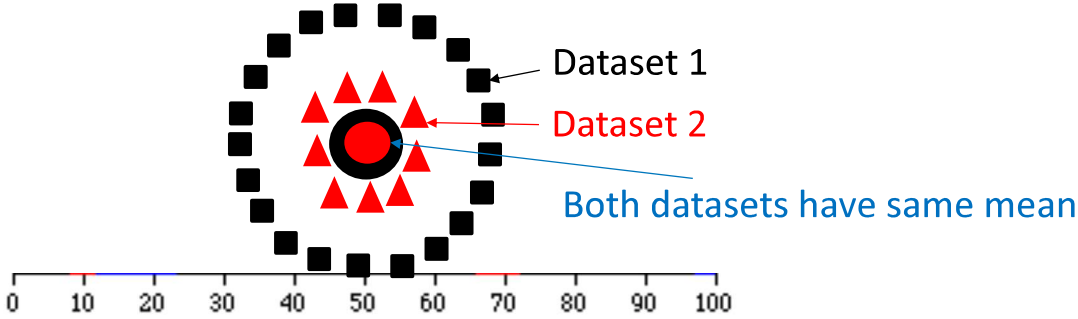
1.5*4 = 6

Upper whisker = 10.5+6 = 16.5 -> 13 (closest value **below** the calculated value)

Lower whisker = 6.5-6 = 0.5 -> 3 (closest value **above** the calculated value)

Measuring Spread and Variability

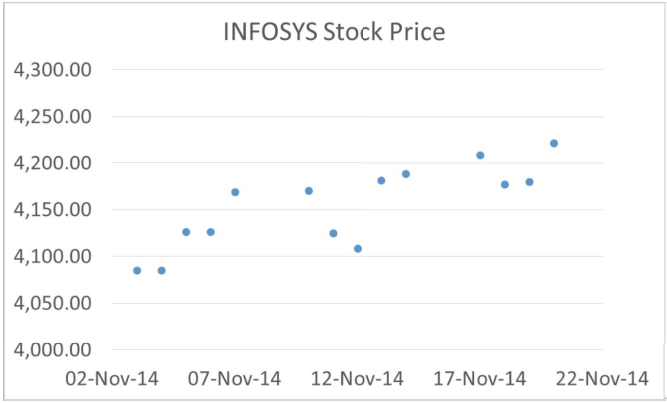
Range and IQR give the spread but still do not describe variability (consistency).

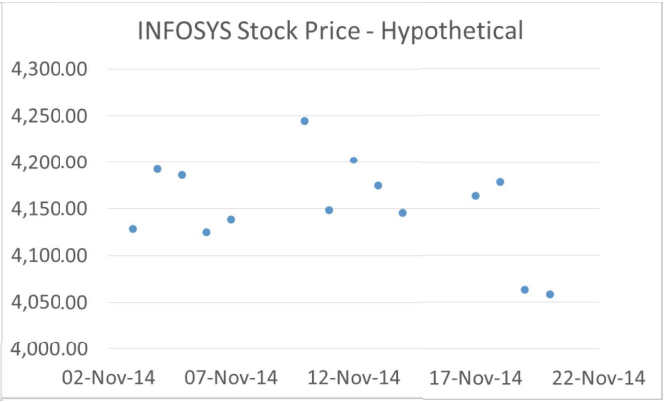


Average distance from the mean?

3 3 6 7 7 10 10 10 11 13 30

Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation - Excel





Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation - Excel

- Mean Distance in both cases = 0
- Mean Absolute Deviation in both cases = 38.17
- Std Dev is 42.54 in the first case and 48.80 in the second.

Measuring Spread and Variability

Variance = $\frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma x^2}{n} - \mu^2$ (Derive)

3 3 6 7 7 10 10 10 11 13 30

Units are squared, which is not intuitive.

Standard Deviation, $\sigma = \sqrt{Variance}$

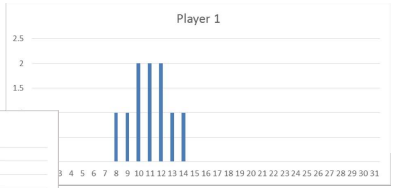
Measuring Spread and Variability

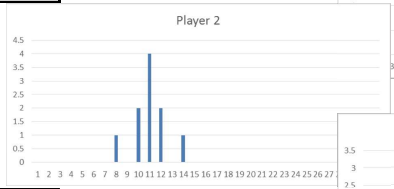
Calculate standard deviation for each player.

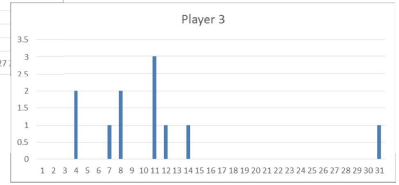
Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1







1.73, 1.48, 7.02

Player 3 is the least reliable.

ASPM : Taxi Times : Unimpeded Taxi Times

Calendar Year From 2010 To 2017 ; Airport=ELP, SFO, ORD, DFW, JFK

				Taxi Out Time				Taxi In Time			
Year	Airport	Carrier Code	Season	Unimpeded	Average	Median	10th Percentile	Unimpeded	Average	Median	10th Percentile
2010	DFW	TRS	4	11	13	12	9	5.4	7	7	4
2010	DFW	UAL	1	11.9	19	16	11	5.4	8.5	7	4
2010	DFW	UAL	2	12	17.3	15	12	5.1	7.6	7	4
2010	DFW	UAL	3	11.3	18.3	16	12	5.5	7.6	7	3
2010	DFW	UAL	4	11.3	16.5	16	12	5.5	7.5	7	4
2010	DFW	USA	1	12.6	17.1	14	10	5.7	7.9	7	5
2010	DFW	USA	2	13	15.9	14	10	5.4	7.5	7	4
2010	DFW	USA	3	13.4	17.1	15	12	5.1	7.8	7	4
2010	DFW	USA	4	12.9	15.6	14	11	5.3	7.5	7	4

Taxi time report generated at https://aspm.faa.gov/apm/sys/TaxiTimes.asp

Table 3 - Descriptive Statistics on the airlines with the largest market share on top 5,000 U.S. domestic markets (n=4002)[24]

	Minimum	Maximum	Mean	Std. Deviation
Distance	129	2724	1081.8	653.292
Daily Passenger	199	9910	771.79	854.279
Fare of Airlines with Largest Market Share	70.96	452.58	180.93	63
LCC dummy	0	1	0.57	0.496
Legacy dummy	0	1	0.24	0.427
Delta dummy	0	1	0.18	0.381
Merger dummy	0	1	0.5	0.5
Average fare	76.04	426.11	178.56	55.54
Largest market share on the route	16.64	100	60.14	19.7

Descriptive Statistics on the airlines with the lowest fare on top 5,000 U.S. domestic markets (n=4002)

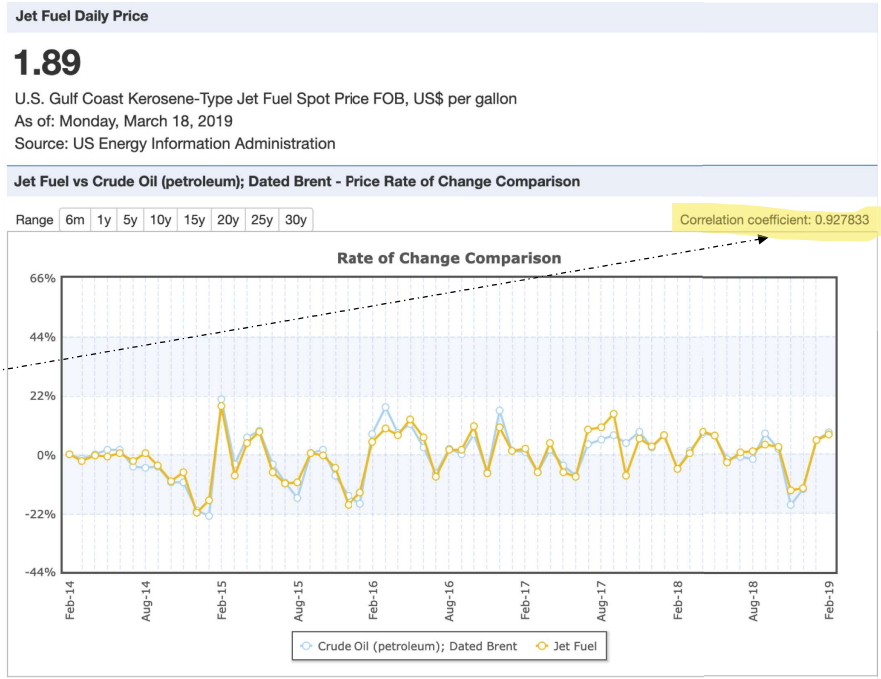
	Minimum	Maximum	Mean	Std. Deviation
Distance	129	2724	1081.8	653.3
Daily passenger	199	9910	771.79	854.28
Average fare	76.04	426.11	178.56	55.54
Market share of lowest-fare airlines	1.02	100	37.14	27.98
LCC dummy	0	11	0.74	0.47
Legacy dummy	0	1	0.15	0.356
Delta dummy	0	1	0.11	0.318
Merger dummy	0	1	0.5	0.5
Lowest fare	66.48	380.27	156.24	47

Source: http://www.scielo.br/img/revistas/jtl/v8n2/a03tab03.jpg from the article "A merger effect on different airline groups: empirical study on the Delta-Northwest merger in 2008" at http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2238-10312014000200004
Last accessed: March 21, 2019

Measuring Variability of Two Variables

When we want to build models and predict one variable using another, we need to understand the relationship between them.

Covariance and its related cousin, Correlation Coefficient help us do that.



Source: https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=60&commodity=crude-oil-brent
Last accessed: March 22, 2019

Measuring Variability of Two Variables – Excel*

$s_x{}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$, $s_y{}^2 = \frac{\sum(y-\bar{y})^2}{n-1}$, $s_{xy}^2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$, where $s_x{}^2$ is the sample variance of the *x* values, $s_y{}^2$ is the sample variance of the *y* values and s_{xy}^2 is the covariance.

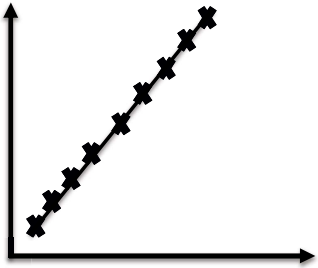
Correlation Coefficient, $r = \frac{s_{xy}^2}{s_x s_y}$.

So, correlation coefficient is simply *standardized (or scaled) covariance*.

* Height and weight data generated randomly using Excel.
Jet fuel price and Crude oil prices from https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=60&commodity=crude-oil-brent

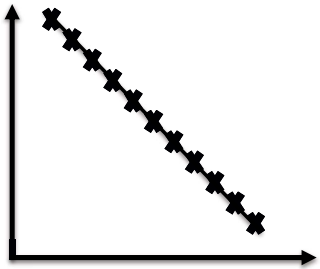
Correlation Coefficient

Correlation coefficient, *r*, is a number between -1 and 1 and tells us how well the two variables are related.



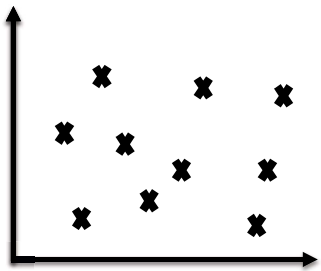
r = 1

Positive Linear Correlation



r = -1

Negative Linear Correlation



r =0

No Correlation

It gives the **strength** and **direction** of the linear relationship between two variables.

Covariance and Correlation - SUMMARY

- **Covariance**

  Tells you the direction of relationship between 2 variables

- **Correlation Coefficient**

  Tells you the direction AND strength of linear relationship between 2 variables

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

PROBABILITY BASICS

Assigning Probabilities

Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ event\ occurs}{total\ possible\ \#\ of\ outcomes}$$

Example: Tossing of a fair die

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Probability can be determined post conducting a thought experiment.

$$P(E) = \frac{\#\ of\ times\ an\ event\ occurred}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$

Example: Tossing of a weighted die…well!, even a fair die. The larger the number of experiments, the better the approximation.

This is the most used method in statistical inference.

Analyzing attributes

PROBABILITY DISTRIBUTIONS

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

Probability Distribution of Income

Ticket Price ($)	90	100	110	500	600	700
Frequency, f	40	37	10	3	5	5

Frequency Distribution

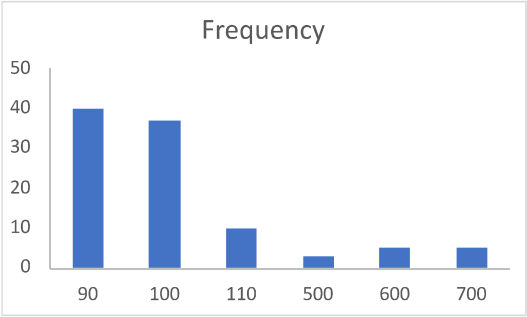
Frequency converted to probability using the Frequentist or the Empirical approach of assigning probabilities

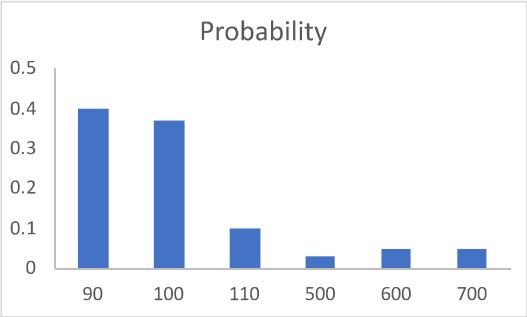
Ticket Price ($)	90	100	110	500	600	700
Probability	.40	.37	.10	.03	.05	.05

Probability Distribution

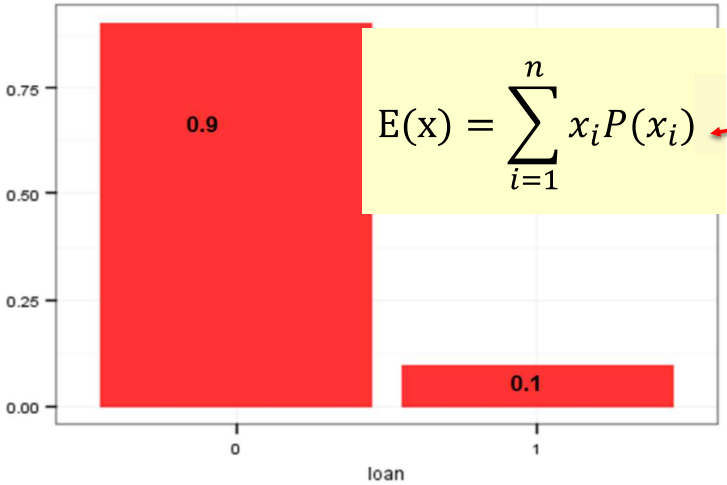
Why do you need a probability distribution?

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





Expectation: Discrete



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

Recall anything like this?

Ticket Price ($)	90	100	110	500	600	700
Frequency, f	40	37	10	3	5	5
Probability	.40	.37	.10	.03	.05	.05

Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{40*90+37*100+10*110+3*500+5*600+5*700}{40+37+10+3+5+5} = \$164$

Expectation, $E(X) = 90 * 0.40 + 100 * 0.37 + 110 * 0.10 + 500 * 0.03 + 600 * 0.05 + 700 * 0.05 = \$164$

Variance

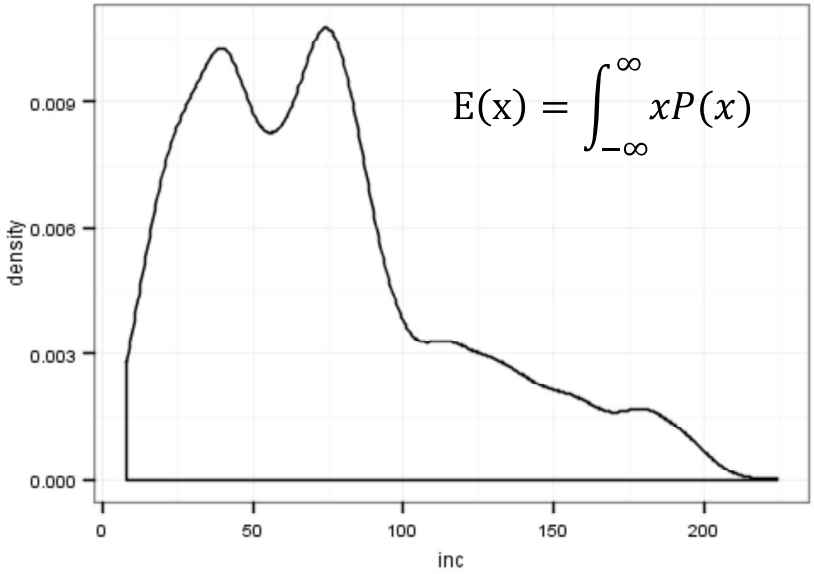
EXPECTATION, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

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



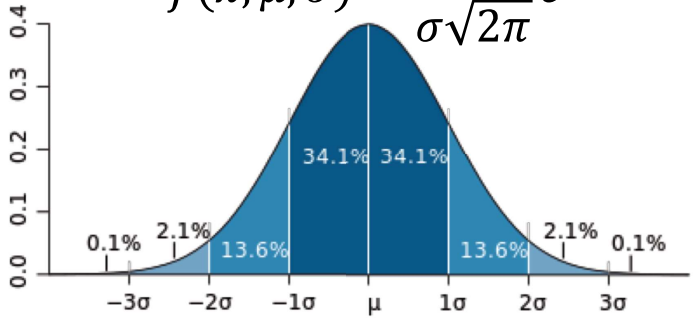
$$\mathrm{E(x)} = \int_{-\infty}^{\infty} xP(x)$$

NORMAL DISTRIBUTION

Normal (Gaussian) Distribution

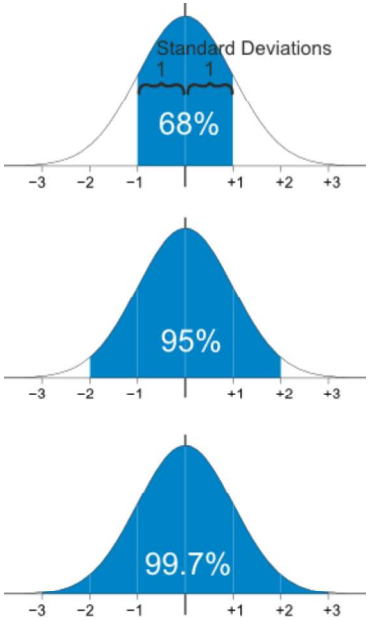
- Mean = Median = Mode
- 68-95-99.7 empirical rule
- X ~ N(μ, σ^2)

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



Shaded area gives the probability that X is between the corresponding values

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?

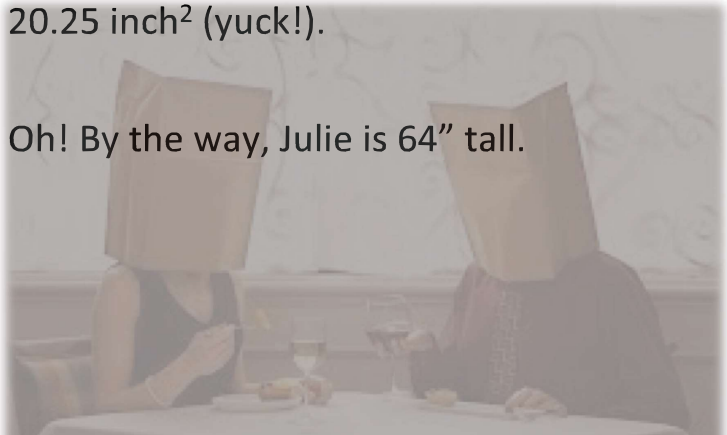
Q2. Approximately 16% of the weights will be more than what value?

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

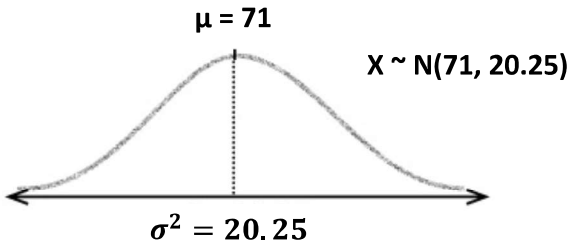
Calculating Normal Probabilities

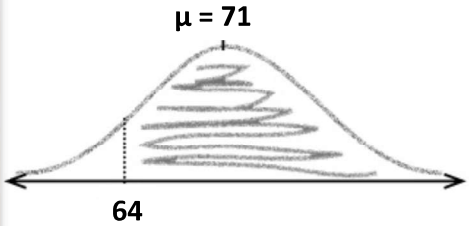
Step 1: Determine the distribution

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!).



Oh! By the way, Julie is 64" tall.





Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)

1. Move the mean

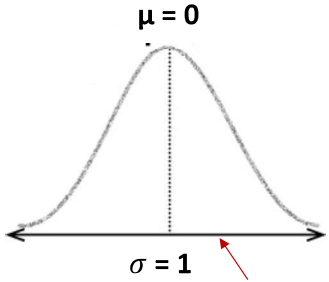
   This gives a new distribution

   X-71 ~ N(0,20.25)

2. Squash the width by dividing by the standard deviation

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

Random variable is **x**, the ***actual*** heights of available guys



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

Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)

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

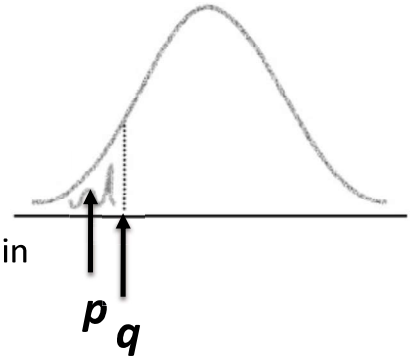
Calculating Normal Probabilities

Step 2: Calculate the probability

In R functions, the distribution is abbreviated and prefixed with an alphabet.

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

*q*norm: Quantile (Inverse CDF) in a *Normal Distribution* – The value corresponding to the desired probability.



Calculating Normal Probabilities

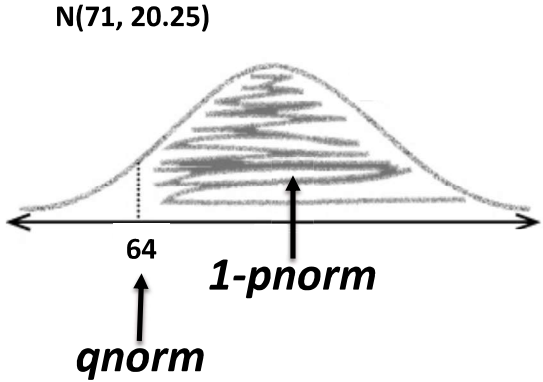
Step 2: Get the probability from R

1-pnorm(64, 71, 4.5)
or
1-pnorm(-1.56, 0, 1)

Answer: 1-0.0599 = 94.01%

qnorm(0.0599, 71, 4.5)

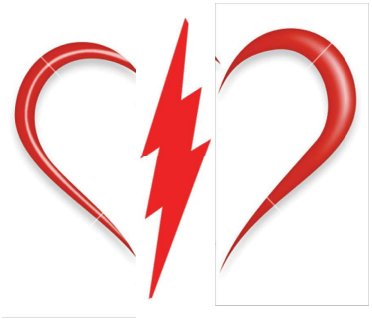
Answer: 64



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".

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)

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

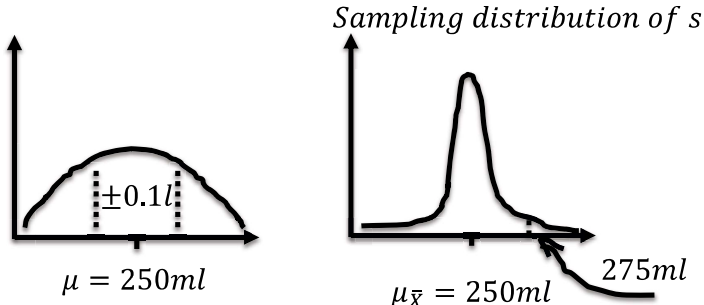
Last accessed: March 23, 2019

Sampling Distribution

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?

$\mu = 250, \sigma = 100$

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



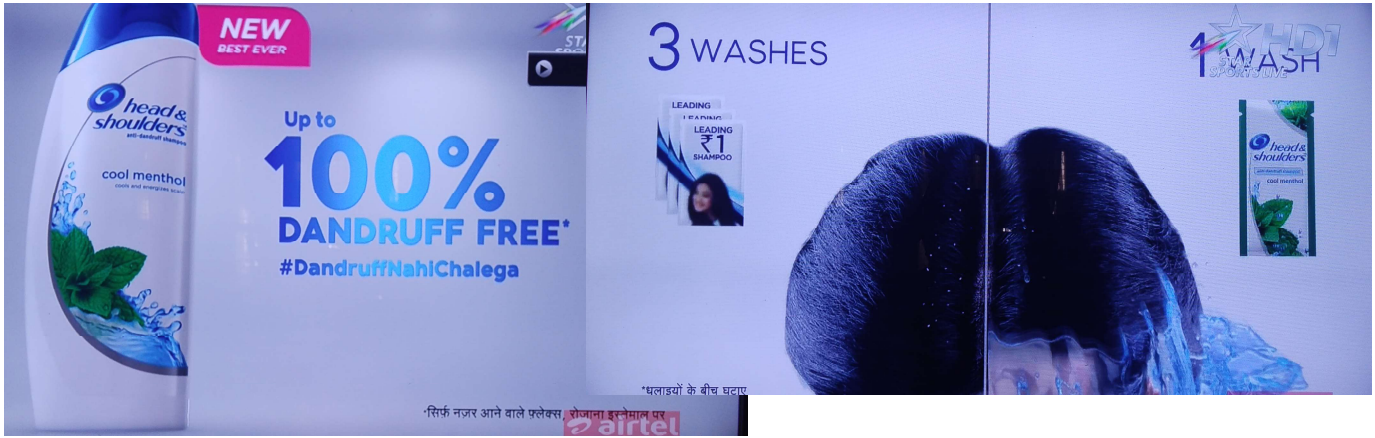
$$1 - pnorm\left(275, 250, \frac{100}{\sqrt{80}}\right)$$

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

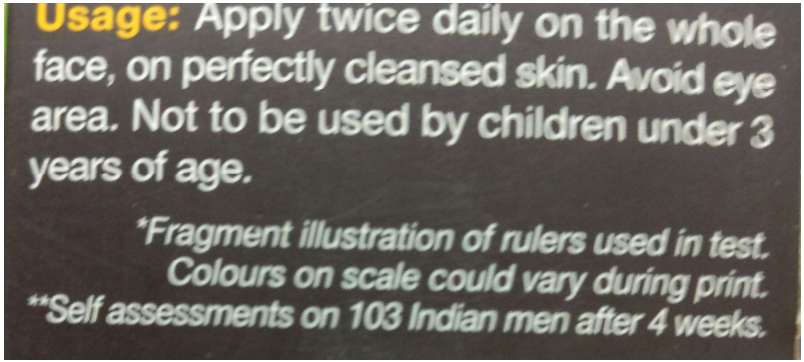
INFERENTIAL STATISTICS

HYPOTHESIS TESTS

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

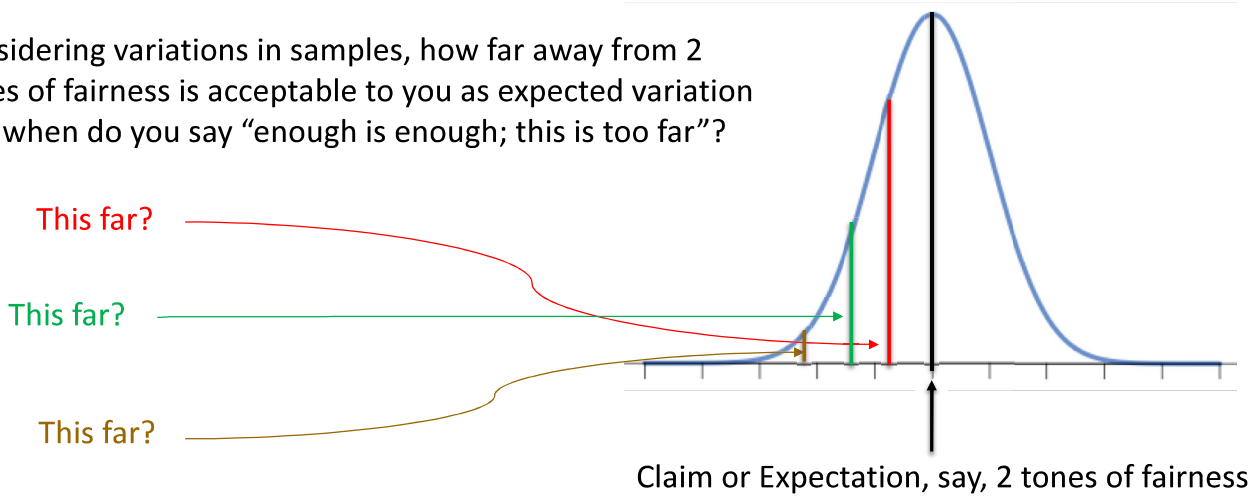


Hypothesis tests give a way of using samples to test whether or not statistical claims are likely to be true or not.



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.

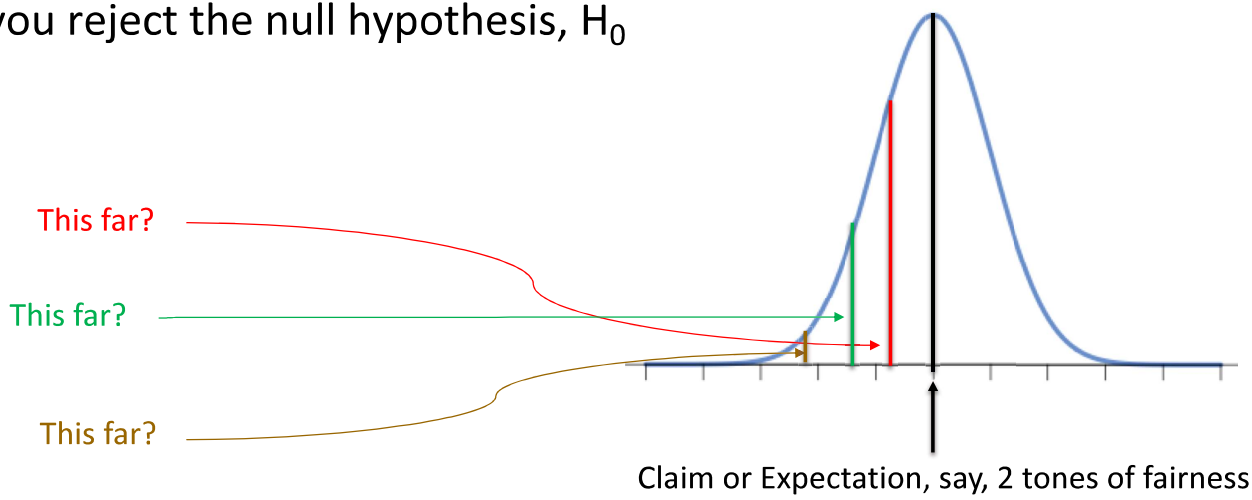
In this case, H_0: Tone = 2

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

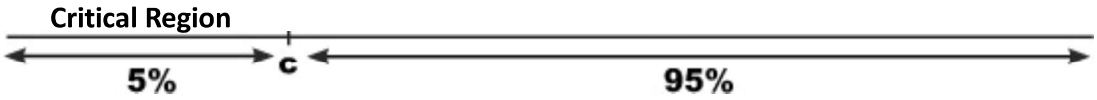
Step 2: Determine the critical region

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: Determine the critical region

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



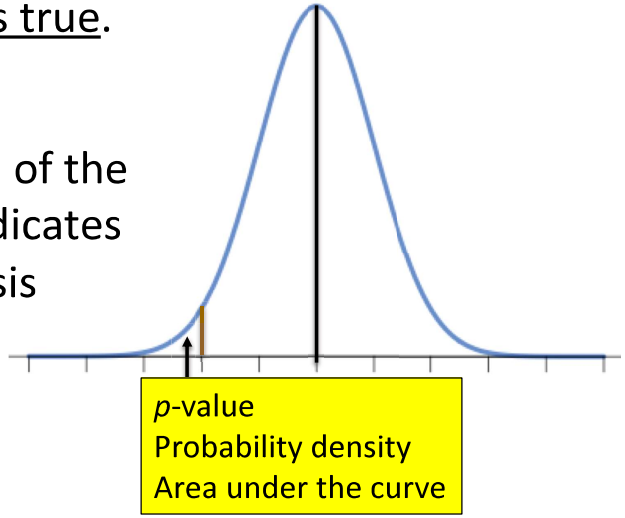
If the sample falls in the critical region, the null hypothesis that 2 tones of fairness increase, is rejected.

5% or 0.05 is called the Significance Level.

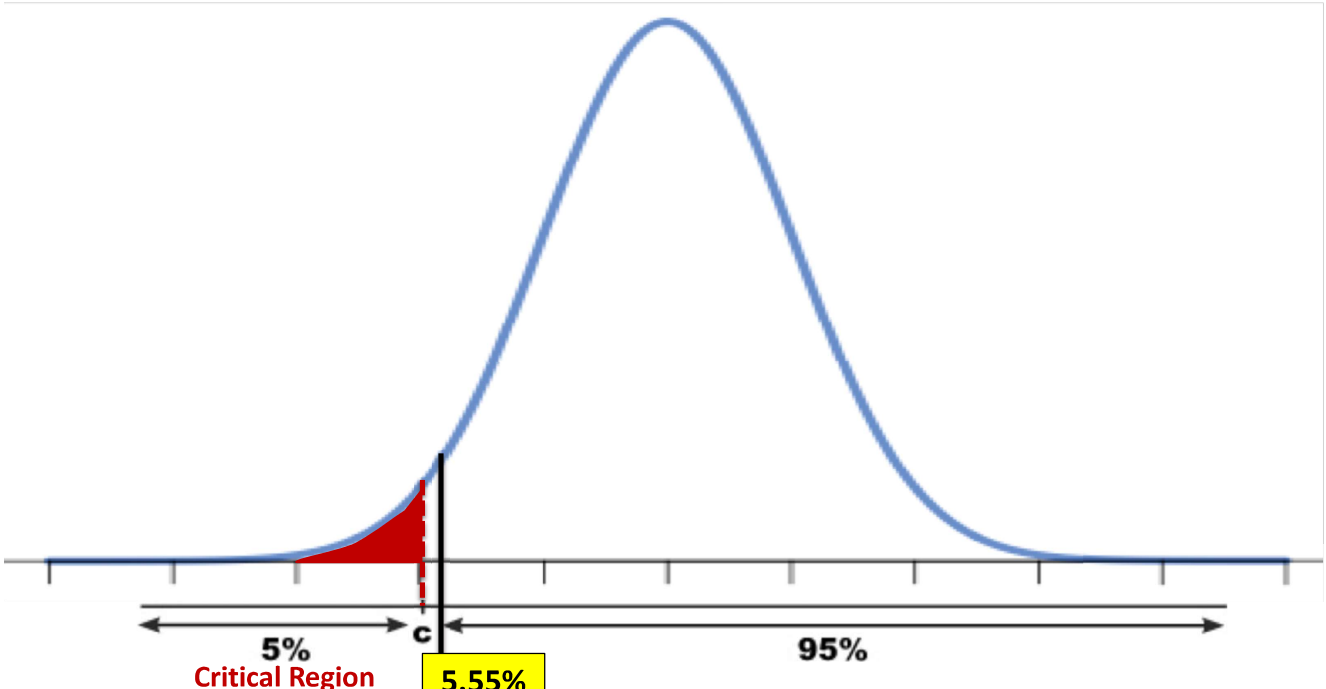
Step 3: Find the *p*-value for the sample

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

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



Step 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.

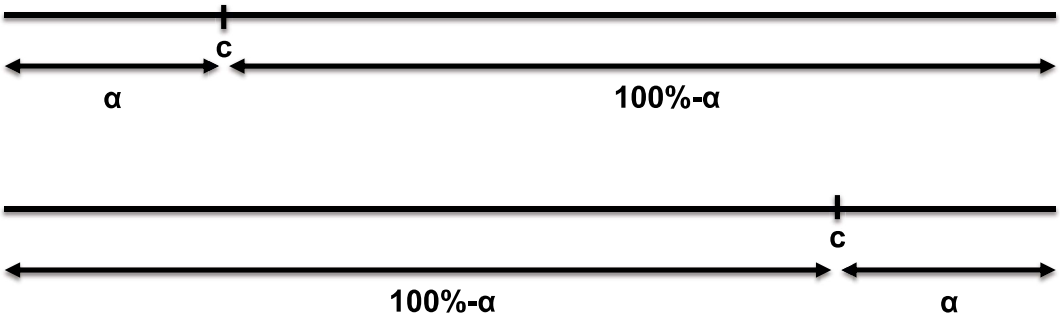
Do you need weaker evidence or stronger to reject the null hypothesis if you were testing at the 1% significance level instead of the 5% significance level?

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.

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.

It only states whether the evidence is enough to reject the null hypothesis or not ***at the chosen significance level***.

Common Test Statistics for Inferential Techniques

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

- z
- ***t***
- χ^2 (Chi-squared)
- **F**

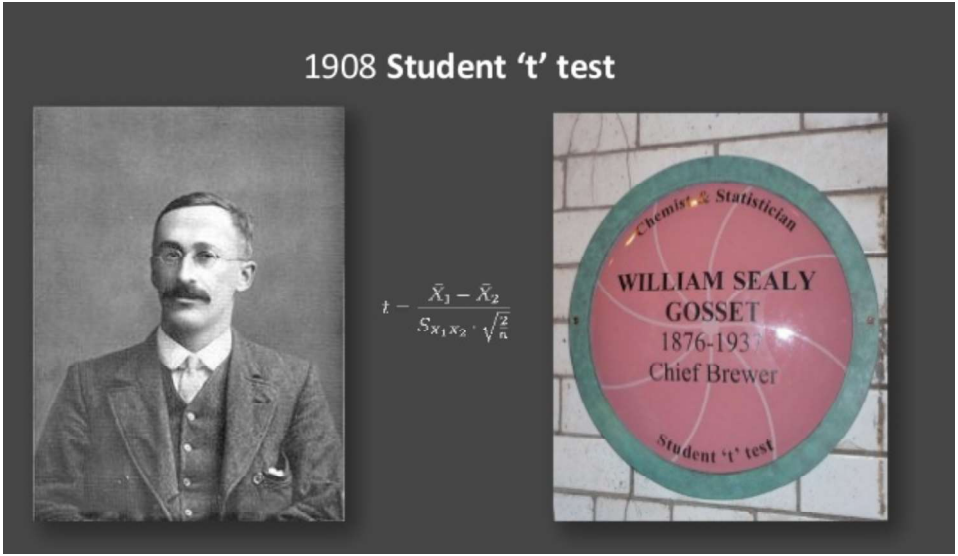
Closely related to Sampling Distribution of **Means**

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

Application of t and F Distributions in Regression

- The t-distribution is used to check the significance of each independent variable. z is used in Logistic Regression.
- The F-distribution is used to check the overall significance of the Linear Regression model.

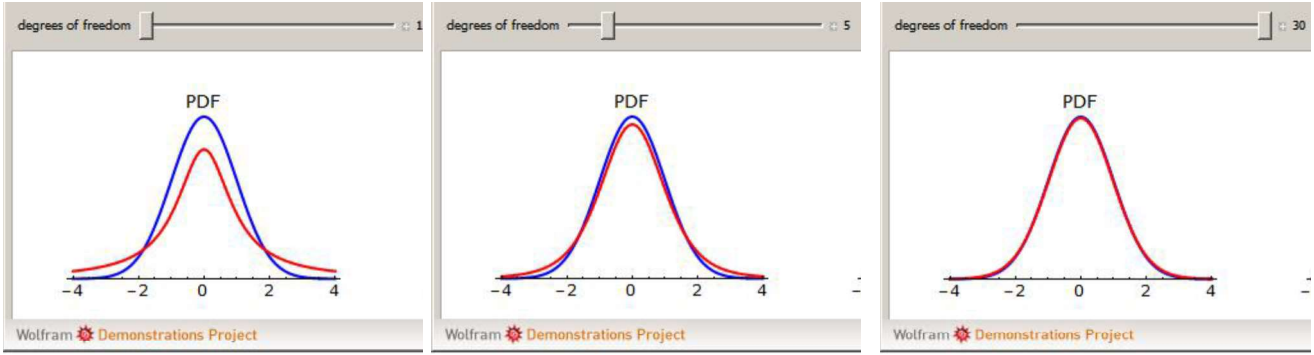
t-Distribution



Ref: http://image.slidesharecdn.com/2013-ingenious-ireland-theingeniousirishiet-slideshow-130524065705-phpapp01/95/2013-

t-Distribution

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



Ref: "Comparing Normal and Student's t-Distributions" from the Wolfram Demonstrations Project

t-Distribution

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

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

When sample size is considered, degrees of freedom are *n-1*.

Recall the Infosys stock hypothetical data created to explain the concept of variance on Day 1 of this module.

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

Properties of *t*-Distribution

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

t-Distribution - Example

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

A researcher wants to test at 5% significance if the true mean of the batch of tablets is indeed 100 mg. Assume the potency is normally distributed.

Data are as follows (in mg):

99.2	100.1	100.0	100.0	99.5
99.4	99.3	100.3	99.9	99.2

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

$H_0: \mu_1 = 100;\ H_1: \mu_1 \neq 100$

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

Two-tailed.

What are the degrees of freedom?

$\nu = 10 - 1 = 9$

t-Distribution – Example – R

R code: t.test(dosage, conf.level = 0.95, mu = 100)

Note: conf.level = 1 - α

How do you get the sample t-value?

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{99.69 - 100}{\frac{0.41}{\sqrt{10}}} = -2.3784$$

```
        One Sample t-test

data:  dosage
t = -2.3784, df = 9, p-value = 0.04134
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 99.39515 99.98485
sample estimates:
mean of x
    99.69
```

HYPOTHESIS TESTING APPROACH

Can population mean be equal to 100mg?

At 5% significance, we reject the null hypothesis that the true mean of the population is 100 mg because the p-value < 0.05, indicating the sample is in the critical region.

F DISTRIBUTION

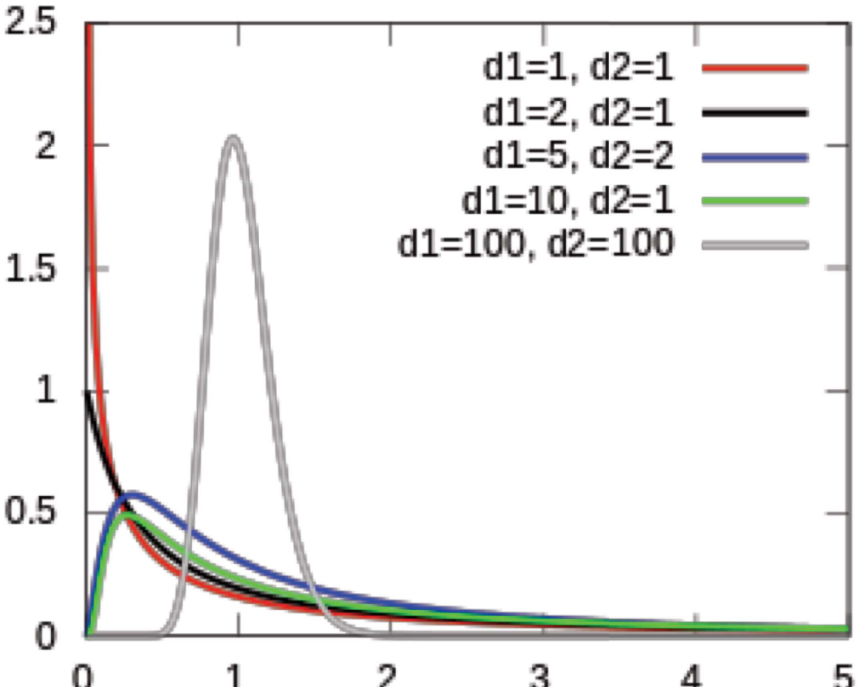
F distribution

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

F distribution

- Ratio of 2 variance estimates: $F = \frac{s_1^2}{s_2^2} = \frac{est.\sigma_1^2}{est.\sigma_2^2}$
- Ideally, this ratio should be about 1 if 2 samples come from the same population or from 2 populations with same variance, but sampling errors cause variation.

F distribution



Hypothesis test for 2 sample variances

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

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

How do you test this?

Hypothesis test for 2 sample variances

Data

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

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

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

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

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

Two-tailed.

What are numerator and denominator degrees of freedom?

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

Hypothesis test for 2 sample variances

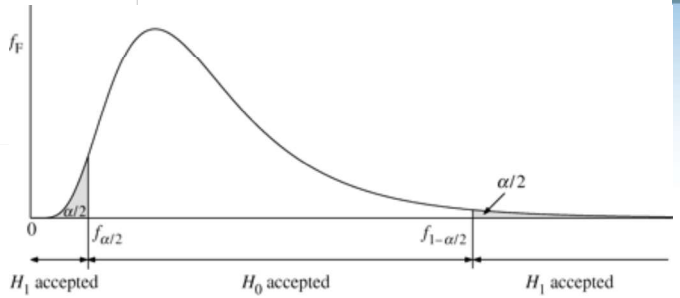
```
> machine1 <- c(22.3,21.8,22.3,21.6,21.8,21.9,22.4,22.5,22.2,21.6)
> machine2 <- c(22,22.1,21.8,21.9,22.2,22,21.7,21.9,22,22.1,21.9,22.1)
> Ftest <- var.test(machine1,machine2,conf.level = 0.95)
> Ftest

        F test to compare two variances

data:  machine1 and machine2
F = 5.625, num df = 9, denom df = 11, p-value = 0.009387
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.567761 22.005297
sample estimates:
ratio of variances
          5.624969
```

Will you reject the null hypothesis or not?

Reject ($p<0.05$). Population variances are not equal.



Hypothesis test for 2 sample variances

What are the business implications?

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

Some Good Resources on Statistics and Probability

- http://onlinestatbook.com
- http://stattrek.com
- http://www.khanacademy.org
- http://www.statsoft.com/Textbook
- http://vassarstats.net/textbook
- Applied Business Statistics by Ken Black
- Statistics For Business: Decision Making and Analysis by Robert Stine and Dean Foster