Mathematics for Artificial Intelligence

Yogesh Kulkarni

January 24, 2021

About Content

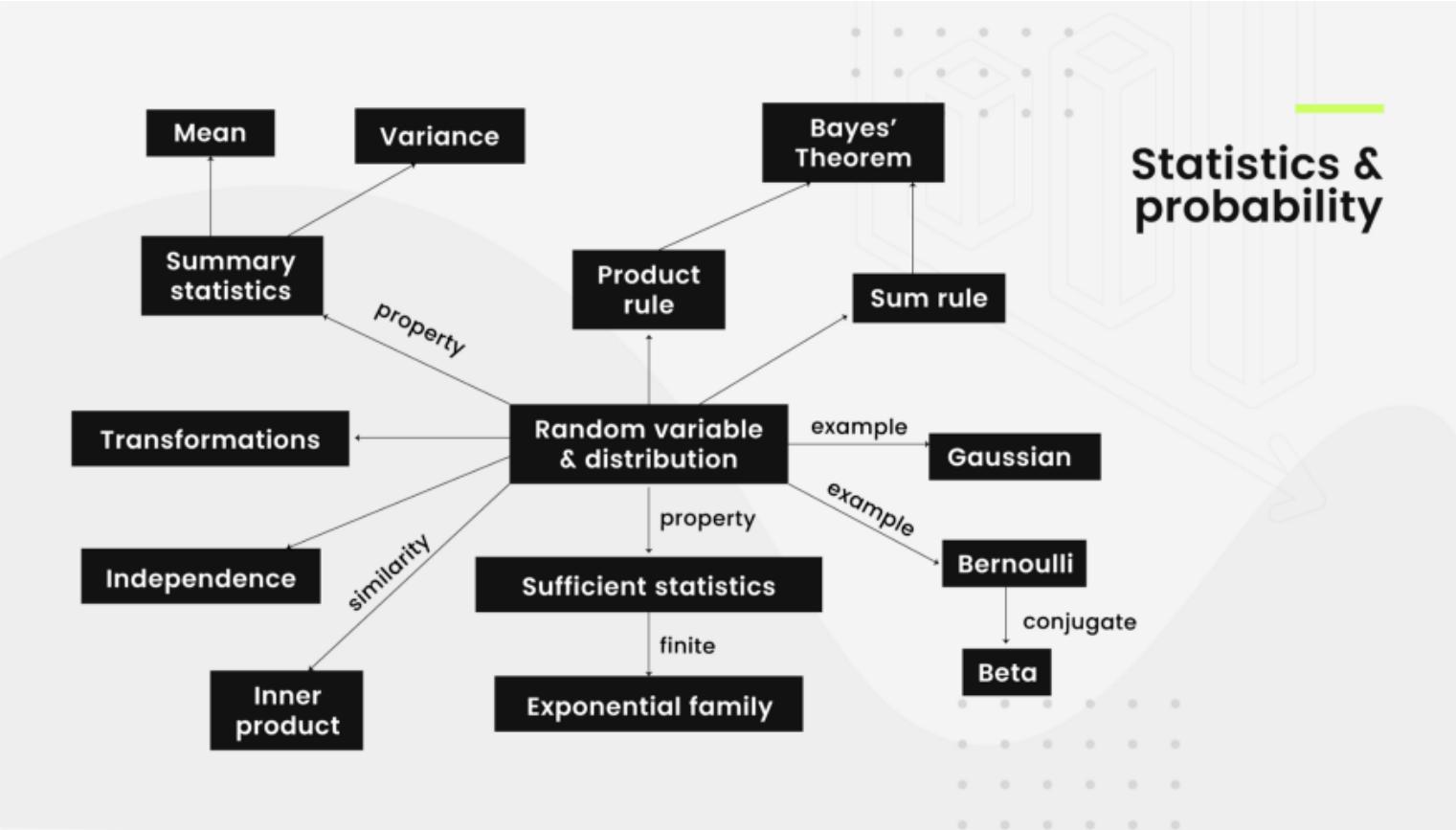
- Bare essential mathematics needed for Machine/Deep Learning
- Just to get you started.
- Each field within Machine/Deep Learning can go extremely deep.
- This is to give you enough foundation needed.

About Me

- I am not a mathematician.
- And this course won't turn YOU into one!!! (if you are not already)
- But, I will surely attempt to make you get interested in the learning, more.

Statistics

Landscape: Statistics



(Ref: The NOT definitive guide to learning math for machine learning - Favio Vazquez)

Descriptive Statistics

Descriptive Statistics

- Describes the data characteristics
- To make sense of the data
- To make rational decisions
- E.g. Demographics, clinical data.
- Measures of Central Tendencies
- Measures of Variability
- Measures of Shape

Why Descriptive Statistics?

- Population: the whole
- Sample: small subset of the population
- Gauging Population by examining traits of Sample.
- Example Question: Finding height of Americans?
- Not going to measure everyone height, but that in a **representative** sample.
- Example: Election sampling?

Why Descriptive Statistics?

- To check the accuracy and precision of the process
- To reduce variability and improve process capability
- To know the truth about the real world

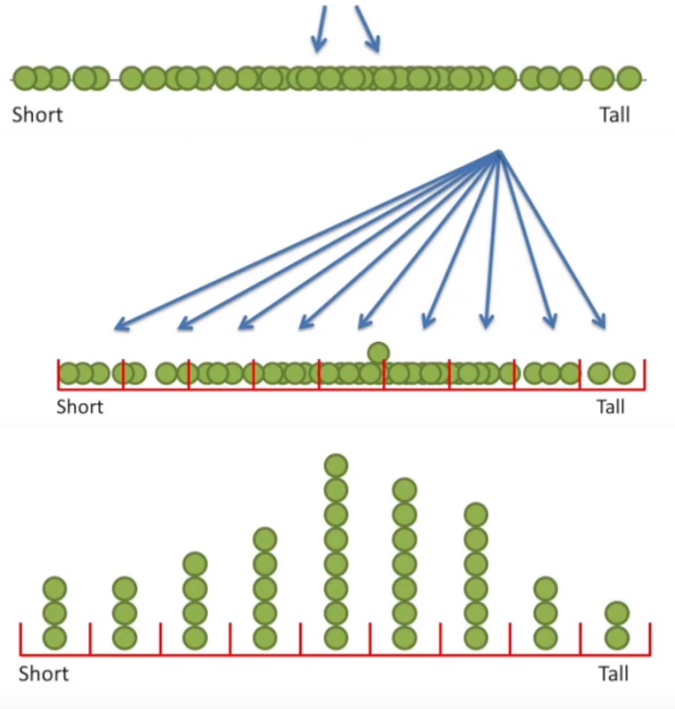
Basic Terms

Histogram

Example

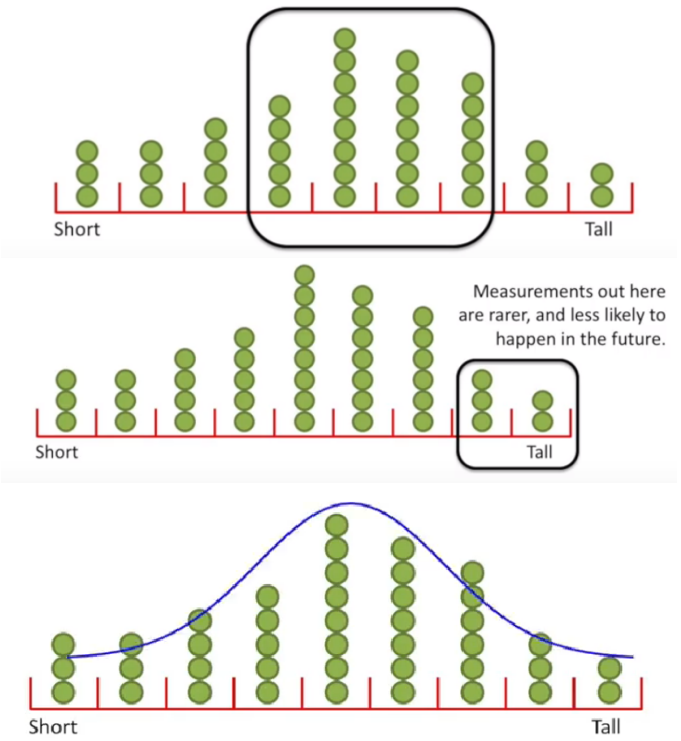
- Say, we are measuring height of people.
- Plotting them on X axis.
- The dots would look very crowded where there are many close or repetitive observations.
- Some dots get hidden.
- We can improve the visualization, by plotting frequency (number of occurrences) on Y axis.
- But in case of contiguous variable, like, exact measurements are rare. So we 'bin' them and measure occurrences.
- Thats Histogram.

(Ref: StatQuest: What is a Histogram? - Josh Starmer)



Histogram

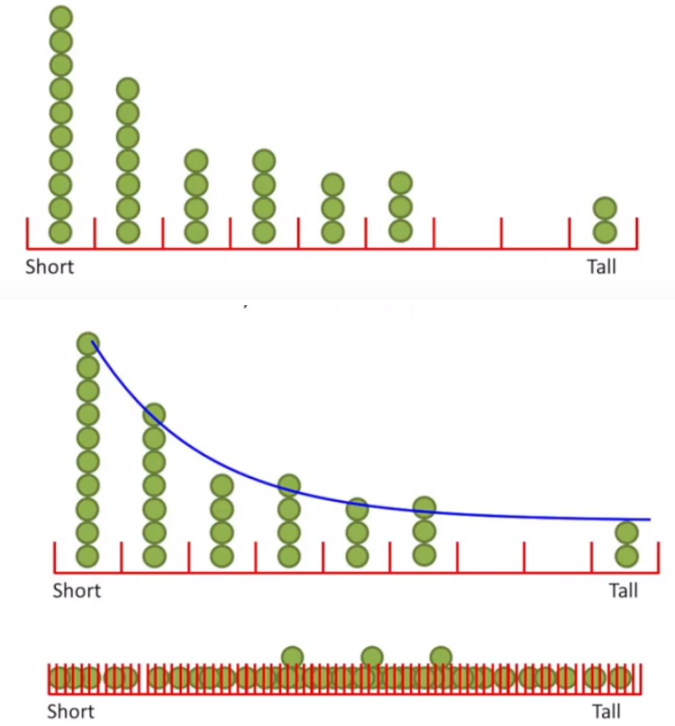
- Histogram can be used to predict probability of getting (future) measurements.
- Getting measurement (as shown in the box) in the middle region is more likely.
- Measurements at both the ends are rare.
- We can approximate this histogram of observations by a 'distribution'.
- Looks like 'Normal' distribution, or a bell-curve



(Ref: StatQuest: What is a Histogram? - Josh Starmer)

Histogram

- If the frequency of measurements seem decreasing, it may be an exponential distribution.
- Binning criterion is critical. They can not be too narrow or too wide.
- Try different bin widths/formulas to plot a histogram.



(Ref: StatQuest: What is a Histogram? - Josh Starmer)

Descriptive Statistics Example

Descriptive Statistics

- Describes features of data sets using numbers
- Individual row: Data
- Full table: Dataset
- Purpose: Answer questions.

Mrs. Graham's 5th Grade Class	Scores on Spelling Test
Bella	43
Betty	45
Bobby	32
Bonnie	45
Booker	38
Boston	45
Botania	50
Boyle	45
Bunder	31

Questions

Mrs. Graham's 5th Grade Class	Scores on Spelling Test
Bella	43
Betty	45
Bobby	32
Bonnie	45
Booker	38
Boston	45
Botania	50
Boyle	45
Bunder	31

- What is Bobby's score?
- Out of? (Total # entries)
- Highest/Lowest scores?

Questions

Mrs. Graham's 5th Grade Class	Scores on Spelling Test
Bella	43
Betty	45
Bobby	32
Bonnie	45
Booker	38
Boston	45
Botania	50
Boyle	45
Bunder	31

- Class average?
- Most Common/frequent Score?
- Any other questions?

Numerical Measures

- Highest to Lowest Score: RANGE
- Most Common Score: MODE
- Average Score: MEAN
- Any other measures?

Descriptive Statistics

- Examines ALL data (not sample)
- Cannot generalize to other datasets

Descriptive Statistics Example

Descriptive Tasks

id	Home Owner	Marital Status	Annual Income	Defaulted Barrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Objective: Derive patterns, summarize underlying relationships
- More exploratory of current state

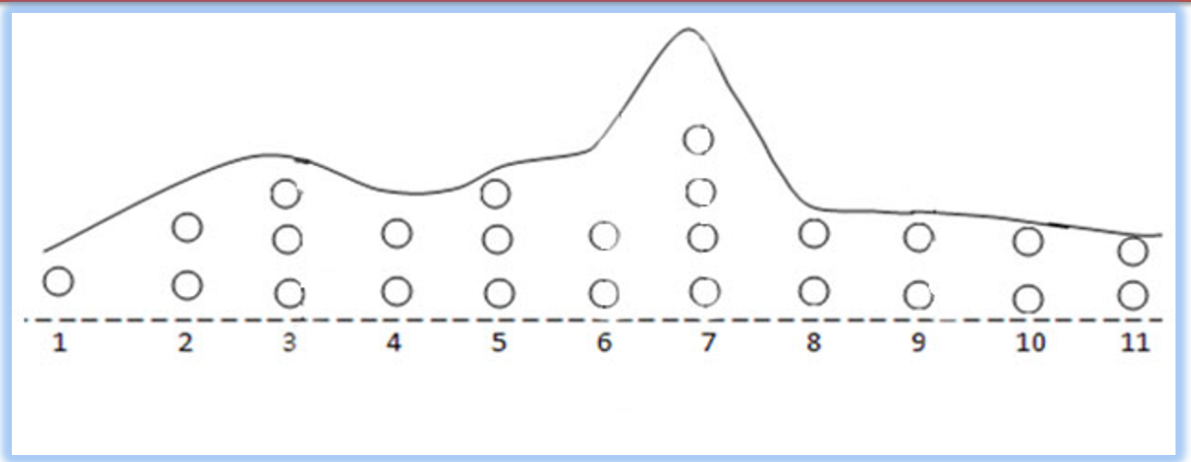
Data Example

1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

What sense it makes? Any pattern?

Visualize

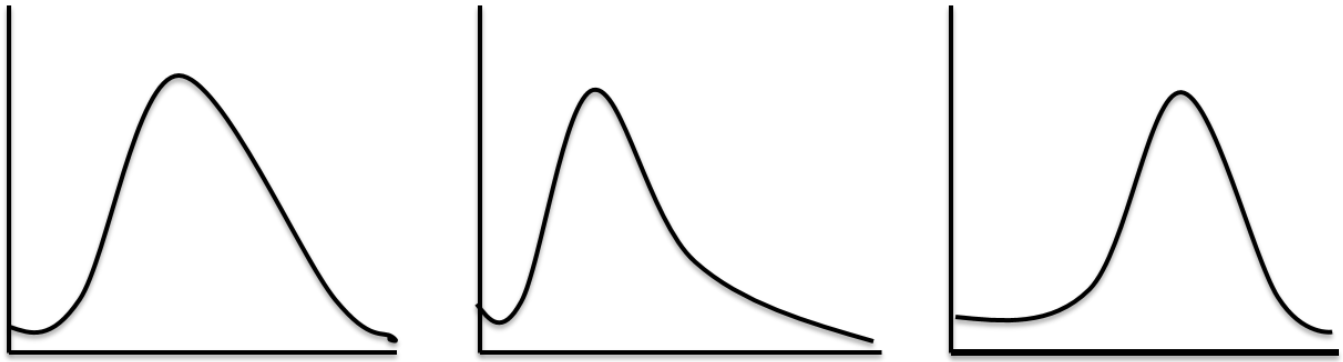
1	2	2	3	3	3	4	4	5
5	5	6	6	7	7	7	7	8
8	9	9	10	10	11	11		



Makes sense?

The Shape of The Distribution

Better to see



Symmetric? Skewed right/left?

Describing Data

Describing Data

- Univariate Analysis: Statistical moments (based on degree)
  - 1st degree: Central tendency: mean, median, mode
  - 2nd degree: Standard Deviation: how wide data is around mean
  - 3rd degree: Skewness: Asymmetric around mean
  - 4th degree: Kurtosis: shape of skewness.
- BiVariate Analysis: covariance, correlation
- MultiVariate Analysis

Univariate Analysis

- Measure of Central Tendency
- Measure of Spread
- Measure of Asymmetry
- Measure of Skewness

Measure of Central Tendency

Mean

- Measure the "location" of a set of values
- Mean is a very, very common measurement
- But is sensitive to outliers

$mean(x) = \bar{x} = 1/n \sum x_i$

Mean

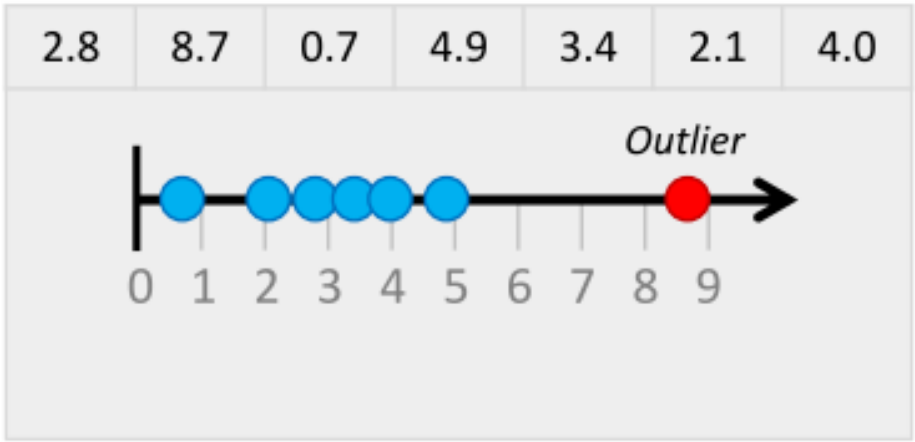
Data set from 25 subjects

1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Sum $= 153$
Mean $= 153/25 = 6.12$

Outliers

- Extreme data point
- May affect calculations
- Can occur in any given data set and in any distribution
- May indicate an experimental error or incorrect recording of data



Mean

Implement mean

```
def mean(datalist):
  :
  return m

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
result = mean(lst)
print("Mean : {}".format(result))
```

Result?

Mean

```
1 def mean(datalist):
    total = 0
3   m = 0
    for item in datalist:
5     total += item
    m = total / float(len(datalist))
7   return m
```

Mean : 13.157894736842104

Median

- Commonly used instead of mean if outliers are present
- Median is the middle value
- if odd number of values are present; average of the two middle values if even number of values
- Not easily affected by outliers (extreme values).
- Always exists and unique.

$$median(x) = x_{r=1} for \quad odd, 1/2(x_r + x_{r+1}) for \quad even$$

Median

Data set from 25 subjects

1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Median = 6

Medians are less reliable: medians of samples drawn from same population vary more widely than sample means.

Median

Implement median

```
def median(datalist):
  :
  return m

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
result = median(lst)
print("Median : {}".format(result))
```

Result?

Median

```
def median(datalist):
    n = len(datalist)
    numsort = sorted(datalist)
    mid = n // 2
    m = 1
    if n % 2 == 0:
        lo = mid - 1
        hi = mid
        m = (numsort[lo] + numsort[hi])/2
    else:
        m = numsort[mid]
    return m
```

Median : 9

Mode

- The value that has the highest frequency.
- Requires no calculation, only counting
- Often used with categorical values.
- The mode (especially with discrete / continuous data) may reveal value that symbolizes a missing value.
- Not a stable measure : it depends only a few values
- May not exist
- May not be unique

Mode

Data set from 25 subjects

1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Median = 7

Mode

Implement mode

```
def mode(datalist):
  :
  return m

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
result = mode(lst)
print("Mode : {}".format(result))
```

Result?

Mode

```
def frequency_distribution(datalist):
  freqs = dict()
  for item in datalist:
    if item not in frees.keys():
      freqs[item] = 1
    else:
      freqs[item] += 1
  return freqs

def mode(datalist):
    d = frequency_distribution(datalist)
    print(d)
    most_often = 0
    m = 0
    for item in d.keys():
        if d[item] > most_often:
            most_often = d[item]
            m = item
    return m
```

Mode : 3

Mode

Another implementation. Counter returns dictionary of frquencies and values.

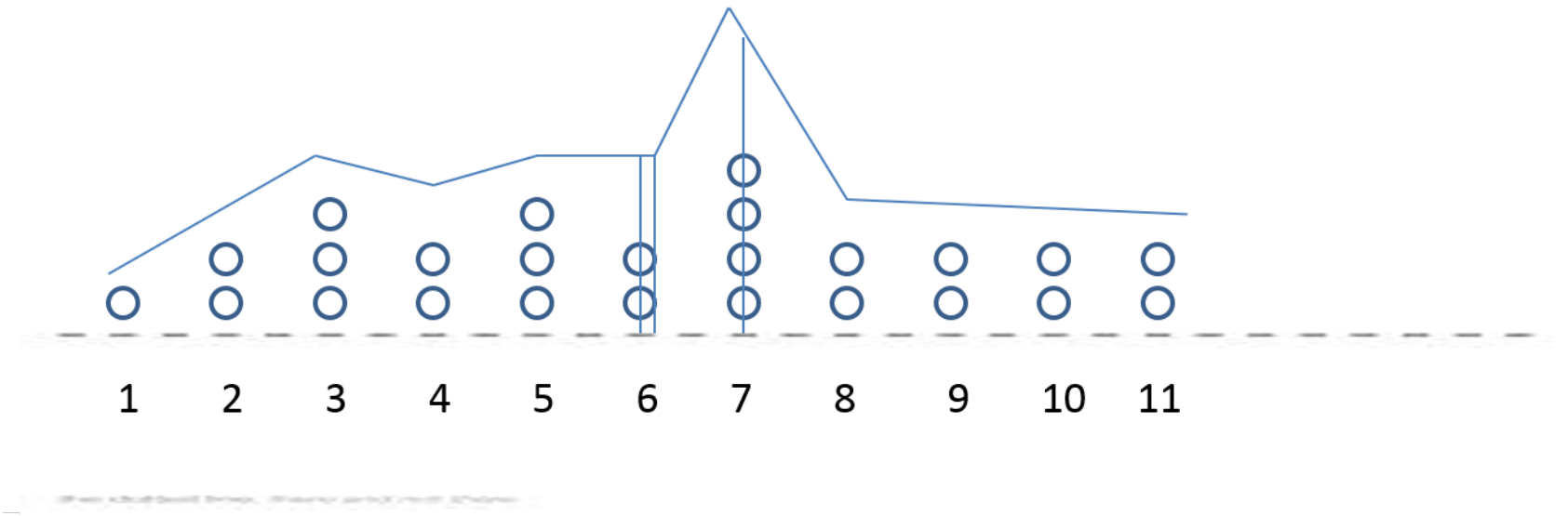
```
from collections import Counter

def mode2(x):
    counts = Counter(x)
    max_count = max(counts.values())
    return [x_i for x_i, count in counts.items() if count == max_count]  #
     multiple modes are possible
```

Mode : [3]

Central Tendency

- Mean: Summarizes all the information in the data set
- Median: Splits the data sets into two halves: there are an equal number of values above and below it.
- Mode: The most common value in the data set.

Locating Central Tendency



Mean = 6.12
Median = 6
Mode = 7

Descriptive Statistics Exercise

Exercise

Our Data

```
lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
```

Store as list of integers and calculate Mean, Median and Mode

Descriptive Statistics Exercise

Exercise

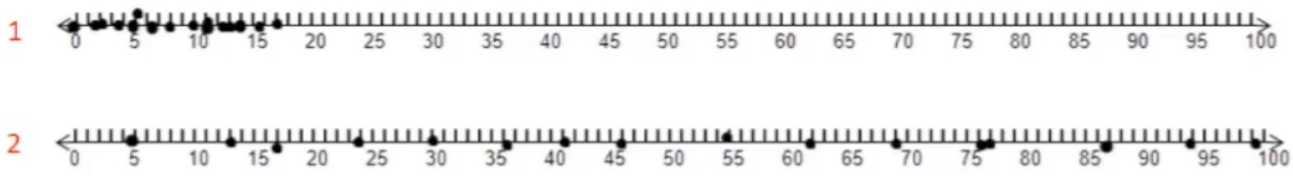
```
crater_diameter = [46, 51, 49, 82, 74, 63, 49, 70, 48, 47, 79, 48, 52, 55, 49,
     51, 58, 82, 72, 45]
print mean(crater_diameter)
print median(crater_diameter)
print mode(crater_diameter)
```

Code: *Result? 58.5,51.5,49*

Measure of Spread

Measure of Spread

In which example (below), the data is spread?



How do you quantify the spread?

Measure of Spread

- Range: The largest value minus the smallest value. Suffers from Outliers.
- Semi-Interquartile range: One half of the difference between the 75th percentile and the 25th percentile. Not affected by Outliers.
- Standard Deviation: The square root of the average of the squared deviations from the mean

Range

- Variation between the smallest and the largest values
- Can be misleading if most values are concentrated, but a few values are extreme

$$range(x) = max(x) - min(x)$$

Range

Data set from 25 subjects

1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Range = 11 -1 = 10

- A 'quick and easy' indication of variability
- No indication of dispersion within
- Unstable, as depends ONLY on Outliers/Extremes

Code: *Calculate and verify the answer*

Range

Implement my_range. It cannot be called as "range" is already there in Python, so a different name

```
def my_range(datalist):
  :
  return min, max, diff

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
min,max,diff = my_range(lst)
print("Range: Min {}, Max {}, Diff {}".format(min,max,diff))
```

Range

```
1  def my_range(abclist):
       smallest = abclist[0]
3      largest = abclist[0]
       range_of_values = 0
5      for item in abclist[1:]:
           if item < smallest:
7              smallest = item
           elif item > largest:
9              largest = item
       range_of_values = largest - smallest
11     return smallest, largest, range_of_values
```

Range: Min 2, Max 44, Diff 42

Range

min max functions are available on list

```
1 def my_range2(x):
    return max(x) - min(x)
3
  diff = my_range2(lst)
5 print("Range: {}".format(diff))
```

Range: 42

Percentiles

- For ordered data, percentile is useful.
- Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile xp is a value of x such that p% of the observed values are less than xp.
- Example: the 75th percentile is the value such that 75% of all values are less than it.

Quantile

Quantile can be of any number between 0 to 1. Quartiles are about quarters so they are quantiles of 0.25, 0.5,0.75
Quantiles are cut points in set of data. They can represent the bottom ten percent of the data or the top 75% or any % from 0 to 100.

Quantile

Implement quantile

```
def quantile(datalist):
  :
  return q

def interquartile_range(x):
  :
  return iqr

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
result1 = quantile(lst,0.10)
result2 = quantile(lst,0.25)
result3 = quantile(lst,0.75)
result4 = quantile(lst,0.90)
result5 = interquartile_range(lst)
print("Q10 {}, Q25 {}, Q50 {} Q90 {} IQR
     {}".format(result1,result2,result3,result4,result5))
```

Quantile

```
def quantile(datalist,num):
    index = int(num * len(datalist)) # slicing parameter
    return sorted(datalist)[index]
    # For values :
    # if num > .5:
    #     return sorted(datalist)[index:]
    # else:
    #     return sorted(datalist)[:index]

def interquartile_range(x):
    return  quantile(x, 0.75) - quantile(x, 0.25)
```

Q10 [2], Q25 [2, 3, 3, 3], Q50 [21, 22, 23, 42, 44] Q90 [42, 44] Q10 3, Q25 3, Q50 21 Q90 42 IQR 18

Semi-Interquartile Range

- Quartiles are Quantiles at 25% and 75%.
- Inter Quartile Range (IQR) is between 25% and 75%.
- More resistant to extreme values than the range
- Does not utilize all the values in the data or set for its computation
- If small, the values are concentrated near the median

Semi-Interquartile Range

- 75th percentile: the value in the date set which is exceeded by 75% of the total number of items in the set
- 25 x (0.75) = 18.75
- 18.75 : rank of the 75th percentile
- 18th and 19th items, both 8
- 75th percentile = 8

Data Set from 25 subjects				
1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Semi-Interquartile Range

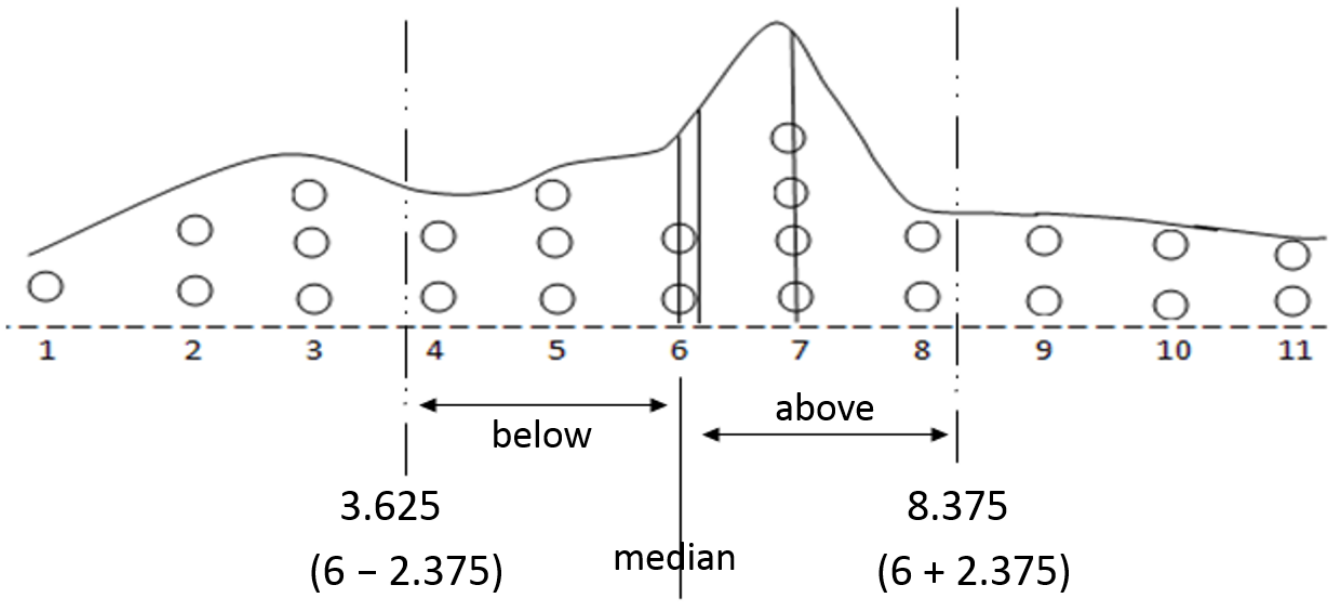
- 25th percentile: the value in the date set which is exceeded by 75% of the total number of items in the set
- 25 x (0.25) = 6.25
- 6.25 : rank of the 25th percentile
- 6th item = 3 and 7th item = 4
- 25th percentile = 3 + (0.25)(4-3)
- 25th percentile = 3.25

Data Set from 25 subjects				
1	2	2	3	3
3	4	4	5	5
5	6	6	7	7
7	7	8	8	9
9	10	10	11	11

Code: *Calculate and verify the answer*

Semi-Interquartile Range

- 75th percentile = 8
- 25th percentile = 3.25
- SIQR =1/2 (8 - 3.25)
- SIQR = 2.375
- Semi-interquartile range = 2.375



Standard Deviation

- How far each value if from the mean
- Uses all the values in the data for its computation
- If small, the values are concentrated near the mean.
- If LARGE, the values are scattered widely about the mean
- z score: how many std deviations from the mean.

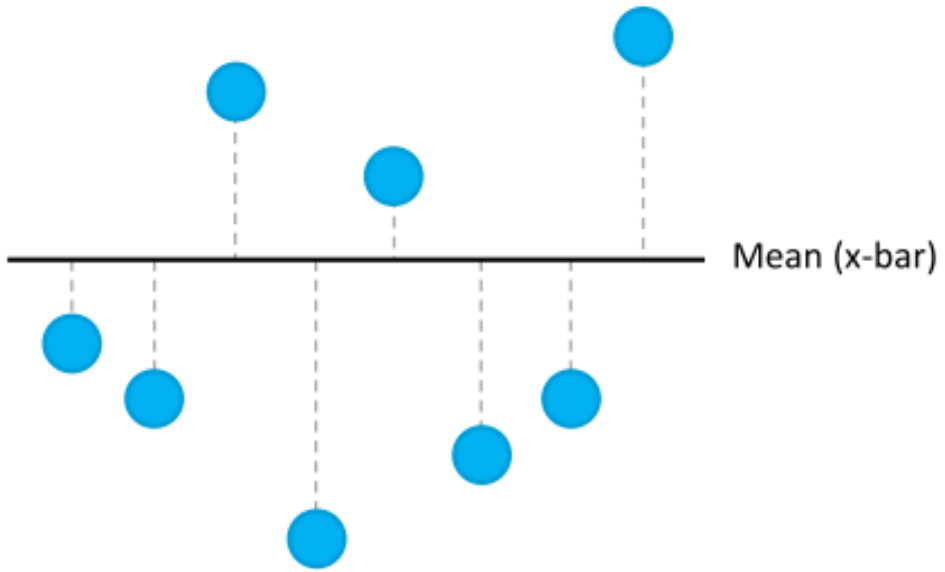
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

s = standard deviation

$\bar{x}$ = mean

x = values of the data set

n = size of the data set

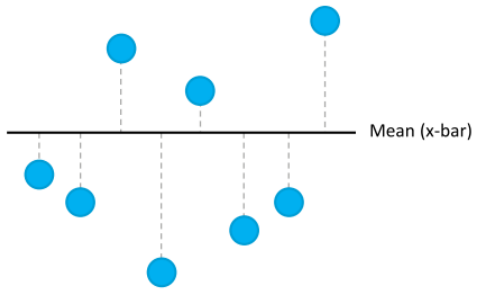


Variance, Standard Deviation

Implement variance and standard deviation.

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

s = standard deviation
$\bar{x}$ = mean
x = values of the data set
n = size of the data set




```
def variance(datalist):
  :
  return v

def std_dev(datalist):
  :
  return s

lst = [9,3,7,2,7,10,23,44,12,42,19,11,22,5,3,4,3,21,3]
result1 = variance(lst)
result2 = std_dev(lst)
print("Variance {}, Std Dev {}".format(result1,result2))
```

Standard Deviation

To reparametrize Covariance value between -1 and 1, need to divide by std devs. Empirical Rule for symmetric bell-shaped distributions

- About 68% of the values will lie within 1 standard deviation of the mean
- About 95% of the values will lie Within 2 standard deviation
- About 99.7% of the values will lie within 3 standard deviation of the mean

$variance(x) = s_x^2 = 1/(n-1)\sum(x_i - \bar{x})^2$
$sd(x) = s_x = \sqrt{1/(n-1)\sum(x_i - \bar{x})^2}$

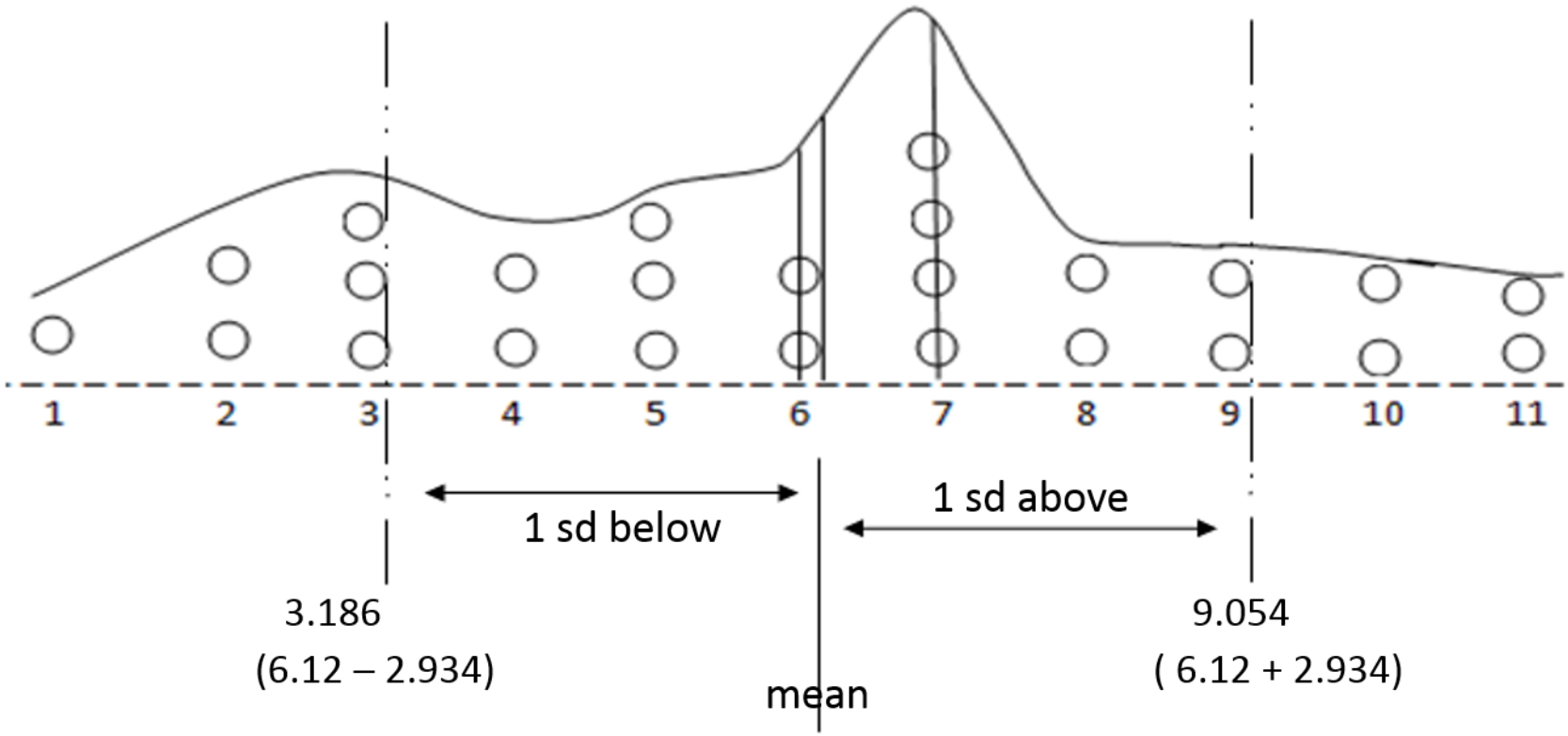
Variance, Standard Deviation

```
def de_mean(x):
  """translate x by subtracting its mean"""
  x_bar = mean(x)
  return  [x_i - x_bar for x_i in x]

def sum_of_squares(diffs):
  sum_of_squares = 0
  for df in diffs:
          sum_of_squares += (df) ** 2
     return sum_of_squares

def variance(x):
  """assumes x has at least two elements"""
  n = len(x)
  deviations = de_mean(x)
  return  sum_of_squares(deviations) / (n - 1)

def std_dev(anotherlist):
  std_dev = variance(anotherlist) ** 0.5
  return std_dev
```

Variance 158.36257309941527, Std Dev 12.584219208970229

Standard Deviation

Standard Deviation = 2.934



Descriptive Statistics Exercise

Exercise

```
crater_diameter = [46, 51, 49, 82, 74, 63, 49, 70, 48, 47, 79, 48, 52, 55, 49,
     51, 58, 82, 72, 45]

print range_min_max(crater_diameter)
print avg_dev(crater_diameter)
print variance(crater_diameter)
print std_dev(crater_diameter)
```

Code: *Result?(45,82,37),11.25,161.45,12.7062*

Exercise

Find the mean, median, range and standard deviation for the following set of data:

`2.8, 8.7, 0.7, 4.9, 3.4, 2.1 & 4.0`

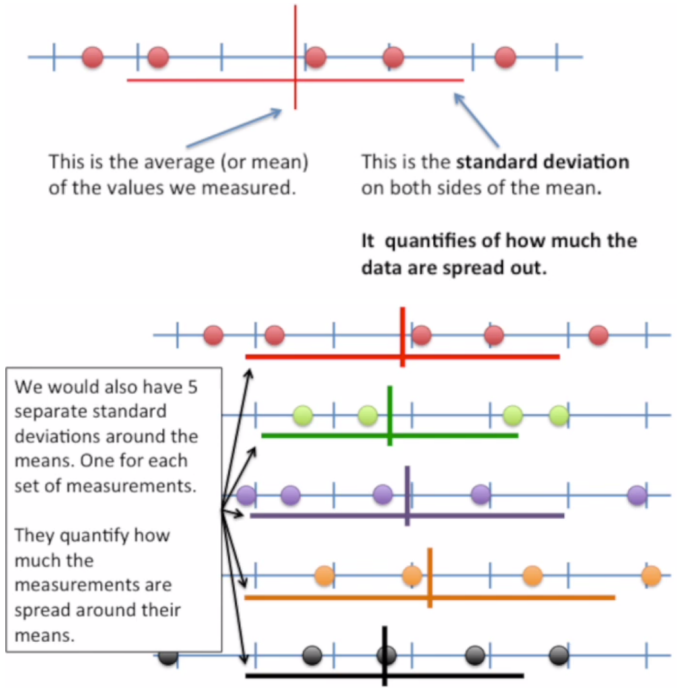
Exercise

Find the mean, median, range and standard deviation for the following set of data:

21	19	20	24	23	21	26	23
25	24	19	19	21	19	25	19
23	23	15	22	23	20	14	20
15	19	20	21	17	15	16	19
13	17	19	17	22	20	18	16
17	18	21	21	17	20	21	21
21	17	17	19	21	22	25	20
19	20	24	28	26	26	25	24

Difference between Standard Deviation and Standard Error

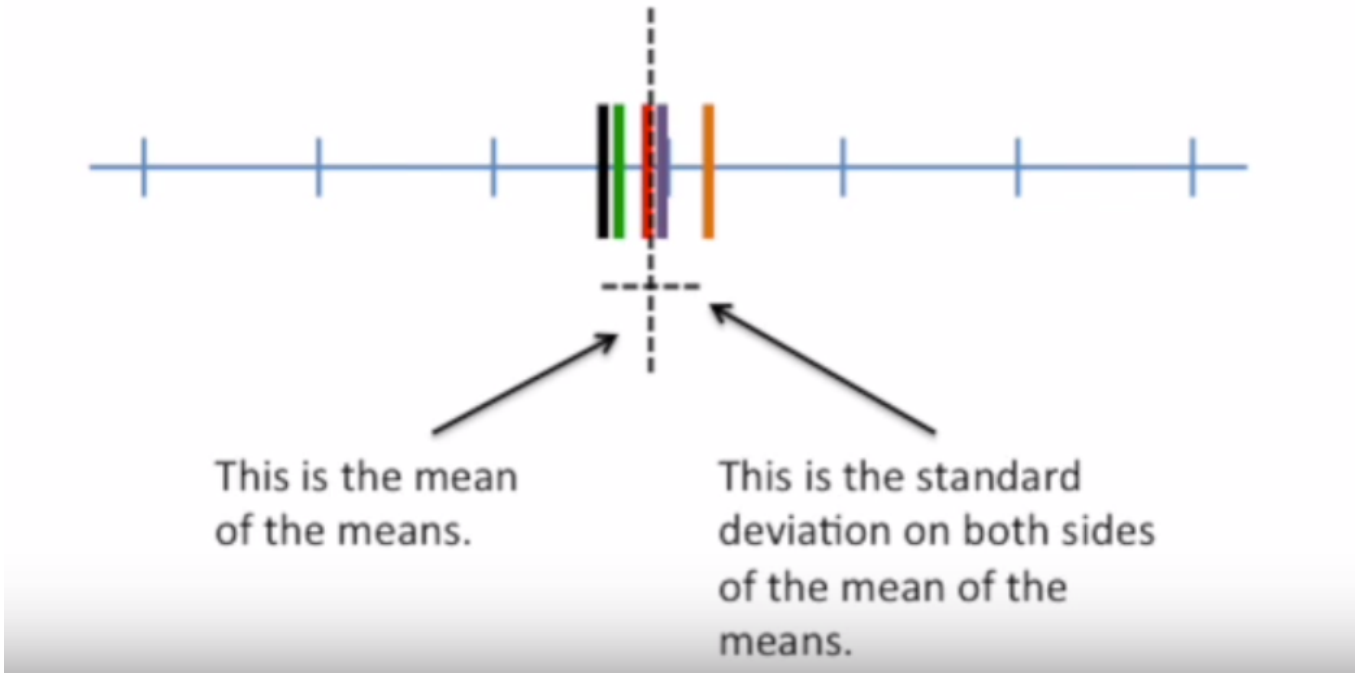
- For a set of normally distributed observations you have mean and standard deviation.
- If you do this for different samples, you get their own respective means and standard deviations.



(Ref: StatQuest: Difference between Standard Deviation and Standard Error - Josh Starmer)

Difference between Standard Deviation and Standard Error

- Plotting those sample means, and sample standard deviations, can form another (meta?) distribution
- Standard deviation of this meta distribution is called Standard Error



(Ref: StatQuest: Difference between Standard Deviation and Standard Error - Josh Starmer)

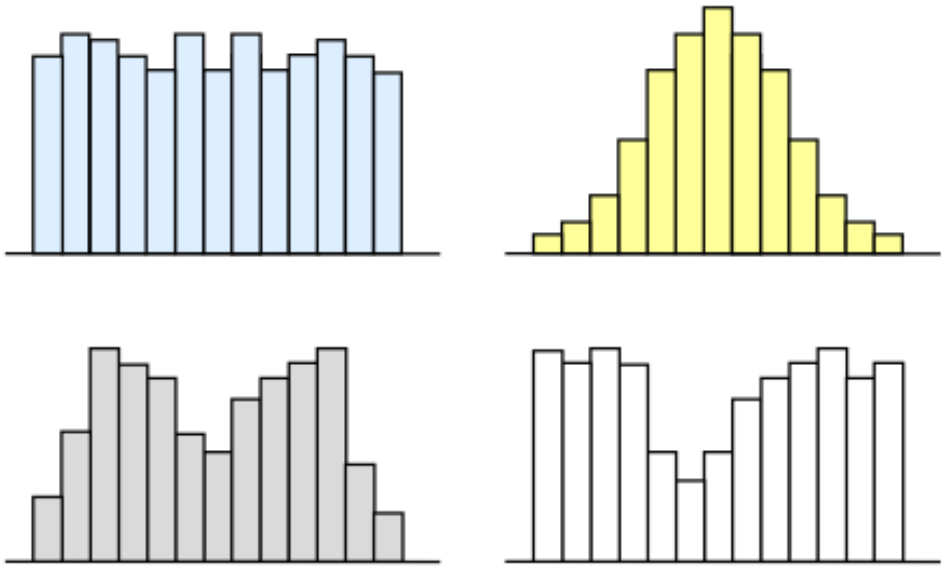
Measure of Asymmetry

Measures of Shape

- To have a general idea of its shape, or distribution
- Helps identifying which descriptive statistic to use
- Symmetrical or nonsymmetrical
- Skewness.
- Kurtosis.

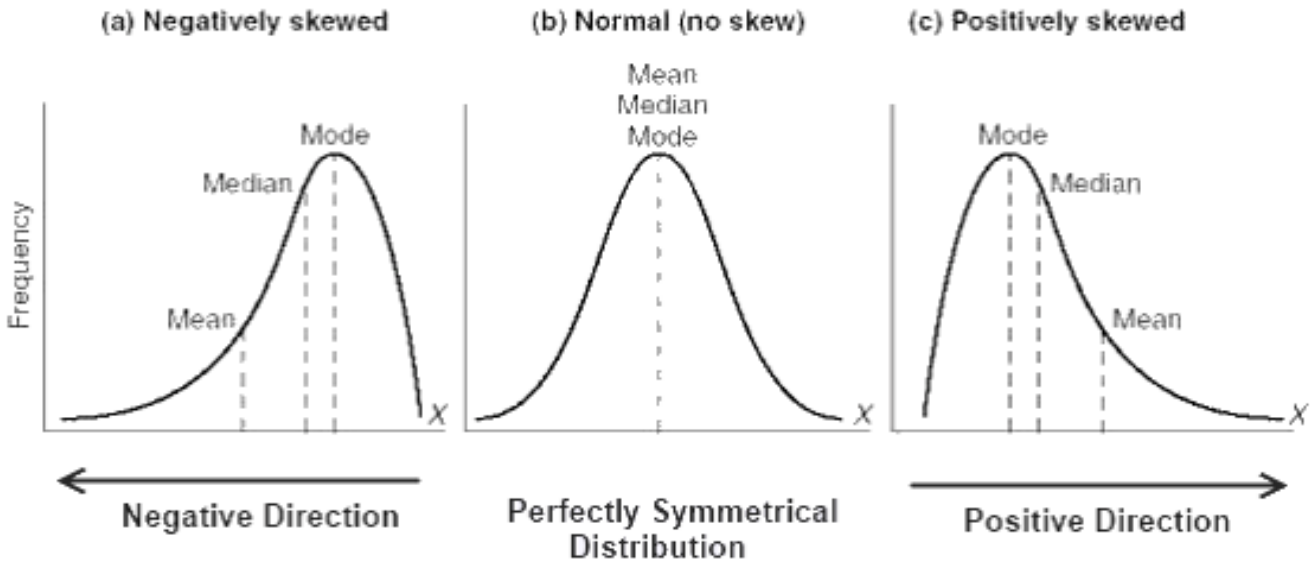
Symmetric

- Uniform.
- Normal.
- Camel-back.
- Bow-tie shaped.



Skewness

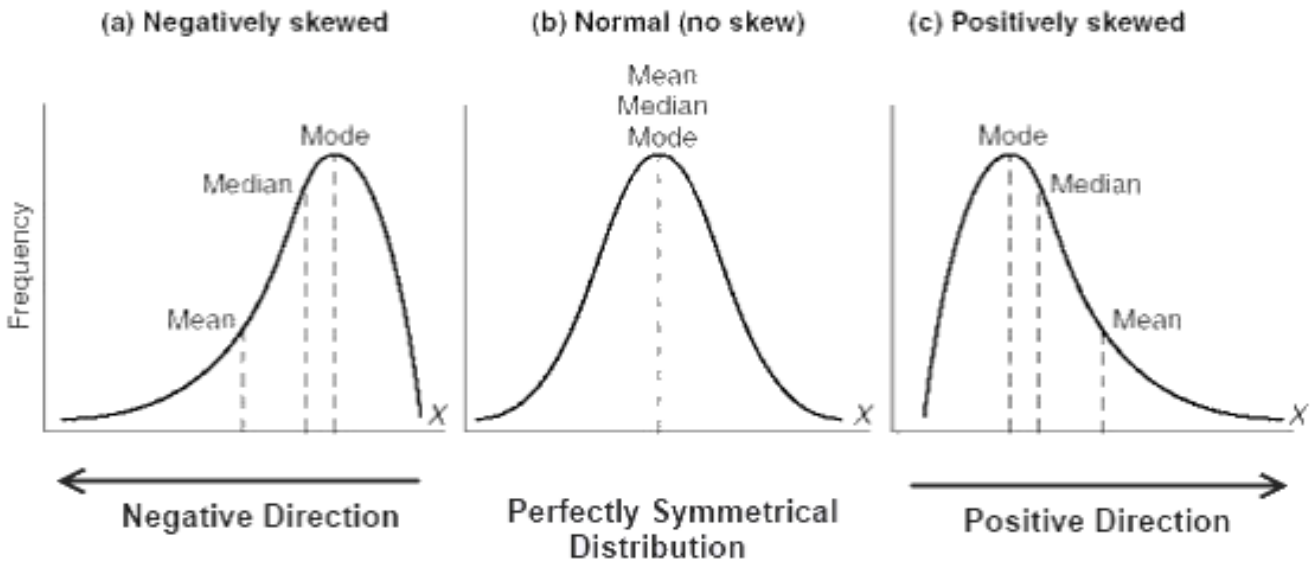
Measures the degree to which the values are symmetrically distributed about the center



If the distribution of values is skewed, then the median is a better indicator of the middle, compare to the mean.

Skewness

For perfectly symmetrical distribution, like Normal Distribution (middle figure):

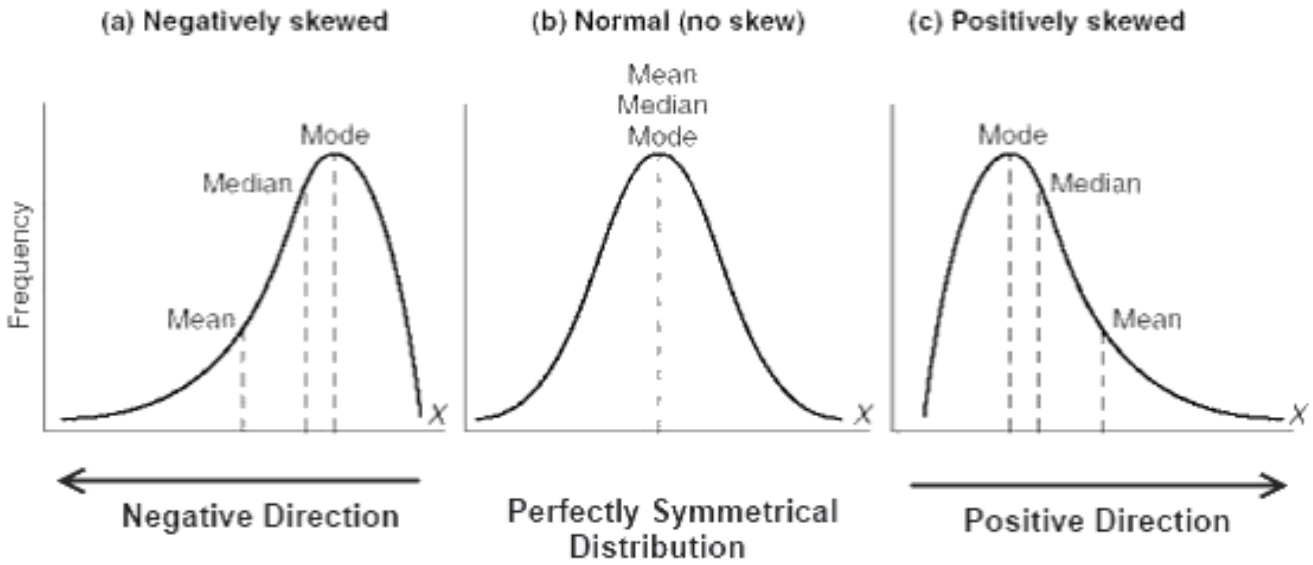


- Whats the mean?: the middle axis point
- Whats the mode?: Highest frequency, top most point
- Whats the median?: Half split of the curve is at the middle.

All Points/Axes are same.

Skewness

For skewed distribution (left and right figures):



- Whats the mean?: towards tail, as most of the heavy (+ve or -ve) points are there
- Whats the mode?: Highest frequency, top most point
- Whats the median?: somewhere between these two

All Points/Axes are different. Sides of Mean and Mode can decide right/left skewness.

Pearson's Skewness Coefficient

Karl Pearson coefficient of Skewness $sk_p = \frac{3(\mu - median)}{\sigma}$

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the difference.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed.
- A large positive value means the distribution is positively skewed.

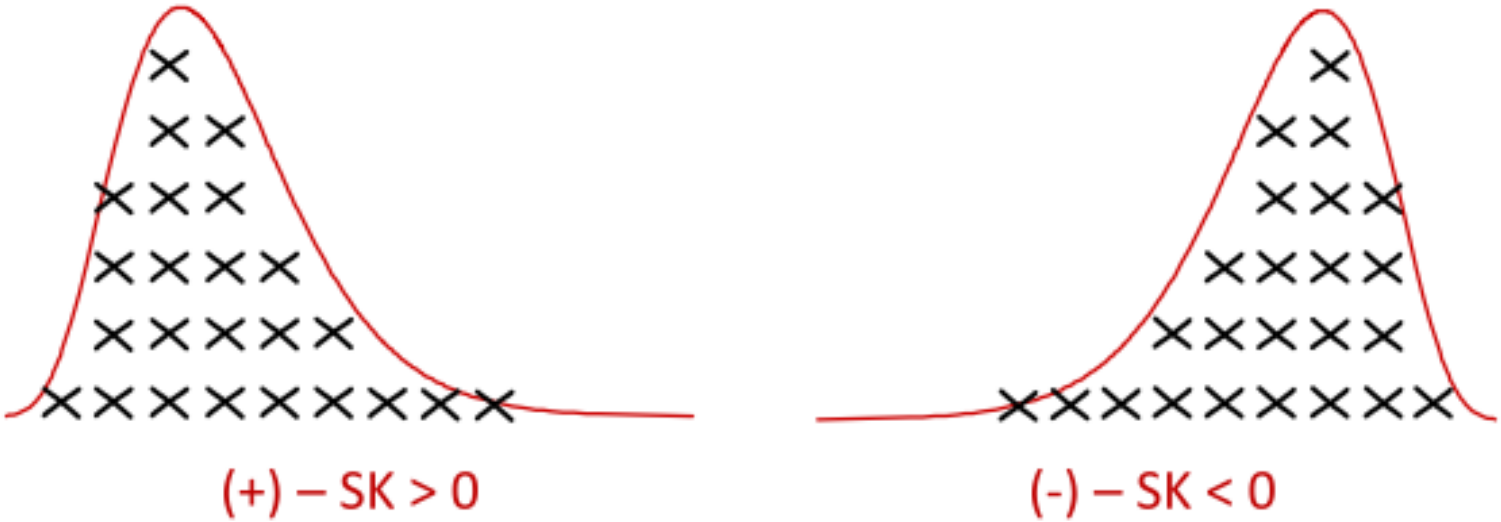
3rd Moment Skewness Coefficient

$$sk_t = \frac{\sum(x_i - \mu)^3}{\sigma^3}$$

- If the power would have been 1 (instead of 3) then $\sum(x_i - \mu)$ would have been 0. +ve and -ve will cancel each other.
- Odd moments are increased when there is a long tail to the right and decreased when there is a long tail to the left.

Skewness

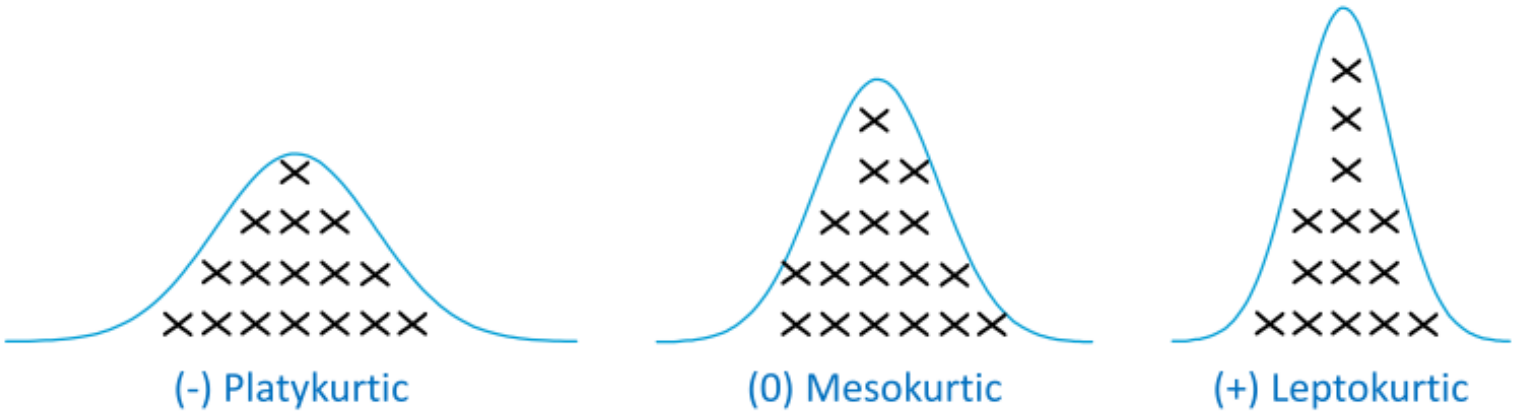
- Zero indicates perfect symmetry
- Negative value implies left-skewed data
- Positive value implies right-skewed data.



Measure of Skewness

Kurtosis

- Measures the degree of flatness (or peakness)
- Clustered around middle? More peak, more kurtosis value
- If values spread evenly, flatted, less kurtosis value



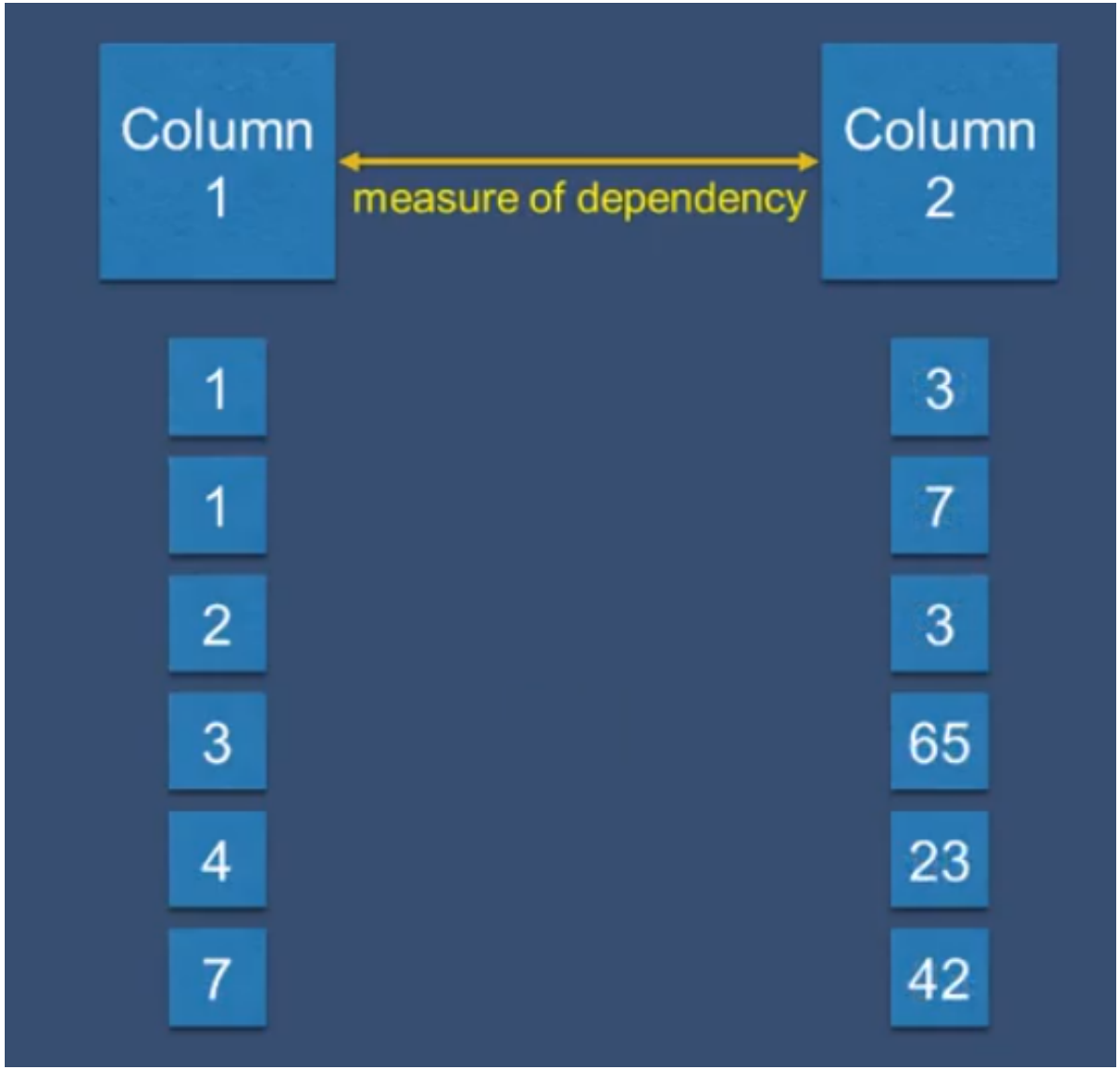
Kurtosis Skewness Coefficient

$$sk_k = \frac{\sum (x_i - \mu)^4}{\sigma^4}$$

- Since the exponent in the above is 4, the term in the summation will always be positive
- Moments of even order are increased when either tail is long.
- Kurtosis is a measure of outlier content. High if longer the tails so more the outliers.
- The third and fourth moments are the smallest examples of these so are used for skewness and kurtosis measures.

Bi-variate Analysis

Bi-variate Analysis



Correlations and Covariance

Covariance and Correlation

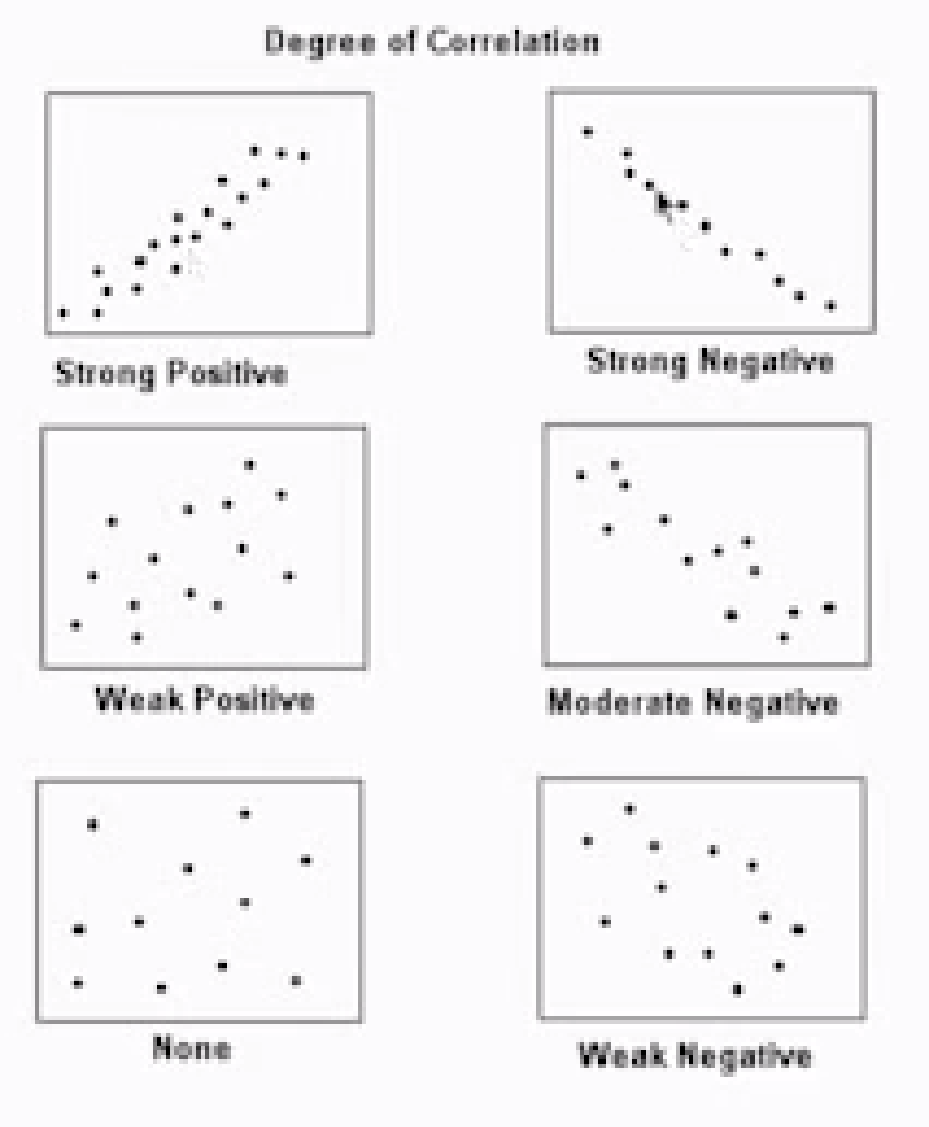
Both show association between two variables

- Positive: If one goes up, the other does too and vice versa.
- Example: Height and weight
- Not always, but tendency
- Another example: Temperature and Ice-creame sales
- Negative: Temperature and sale of woolen clothes

Correlation

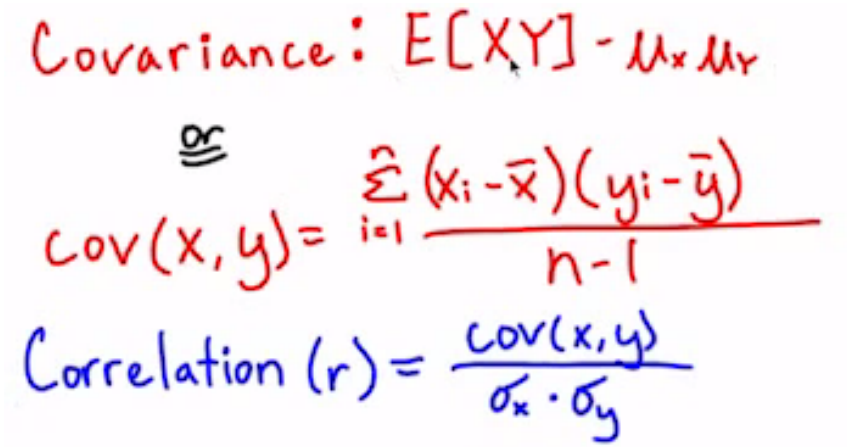
- Correlation is a value standardized between -1 to 1
- Relation between two variables is linear,
- Directly proportional in case of Positive Corr
- Inversely proportional in case of Negative Corr
- The value of corr is the factor of proportionality
- No correlation, ie no dependence so value $= 0$

Covariance and Correlation



Covariance

Implement covariance, the paired analogue of variance. The variance measures how a single variable deviates from its mean, covariance measures how two variables vary in tandem from their means.




```
x = [2, 3, 0, 1, 3]
y = [ 2, 1, 0, 1, 2]
result1 = covariance(x,y)
result2 = correlation(x,y)print("CoVariance {}, Correlation
     {}".format(result1,result2))
```

Covariance

Covariance is like a dot product and tell how two quantities (centered, meaning subtracted by Mean) are together/similar.

```
def elemwise_multi(v, w):
    """v_1 * w_1 + ... + v_n * w_n"""
    return sum(v_i * w_i for v_i, w_i in zip(v, w))

def covariance(x, y):
  n = len(x)
  return  elemwise_multi(de_mean(x), de_mean(y)) / (n - 1)
```

CoVariance 0.8

Correlation

Covariance is like a dot product normalized by standard deviation.

```
def correlation(x, y):
  stdev_x = std_dev(x)
    stdev_y = std_dev(y)
  if stdev_x > 0 and stdev_y > 0:
    return  covariance(x, y) / stdev_x / stdev_y
  else:
    return  0 # if  no variation, correlation is zero
```

Correlation 0.7333587976225691

R^2

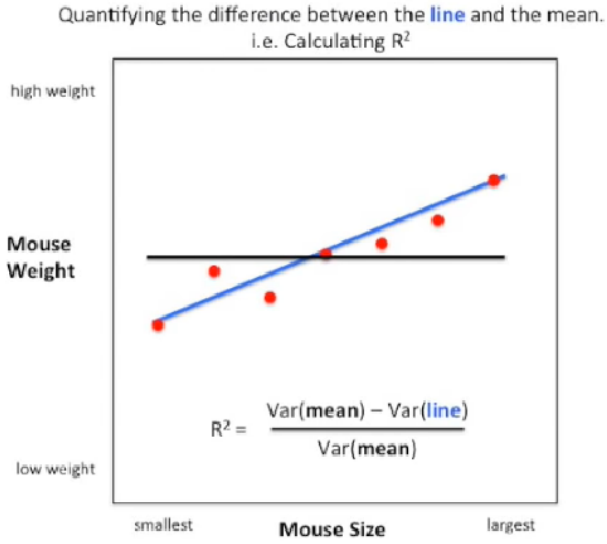
- Correlation, the regular 'R' has values from -1 to 1 and is good enough to tell you that the two quantitative variables are strongly related.
- Why do you need R^2 then?
- Plain R is not easier to interpret.
- Example: $R = 0.7$ is twice as good as $R = 0.5$
- But its more clear when $R^2 = 0.7$ is 1.4 times as good as $R^2 = .5$

(Ref: StatQuest: R-squared explained - Josh Starmer)

R^2

- R^2 is used to decide the quality of the linear fitting.
- $Var(mean)$ represents the variation of just the mean line, ie black line.
- $Var(line)$ represents the variation calculated using he fitted line, ie blue line.
- Taking just relative ratio to make R^2 in range 0 t 1 and as a percentage.
- If the value is 0.81, it means there is 81% less variation around fitted line than the benchmark black line.
- So, if one variable is input (size) and one is output (weight), then we say that 81% of weight variation is explained by size.

(Ref: StatQuest: R-squared explained - Josh Starmer)

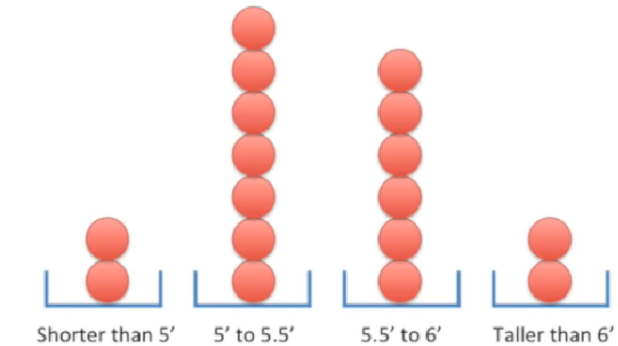


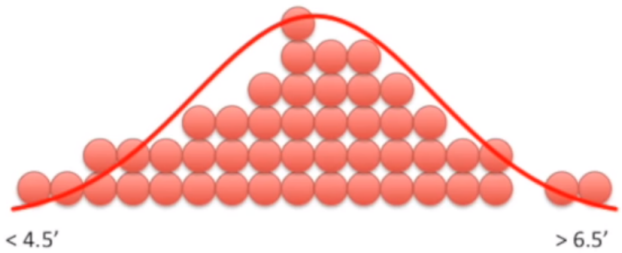
Distributions

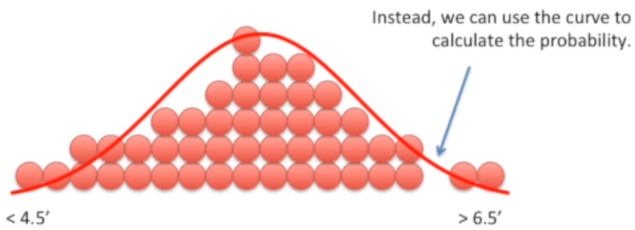
What is a Statistical Distribution?

Example

- Say, we are measuring height of people and we plotted histogram.
- Measurements in the middle are more than at the ends (extreme values).
- If we decrease the bin size to very small width, and with lots of observations, we will get a continuous curve approximating the tops.
- The curve gives a general sense, so we still can calculate frequencies of the missing observations.







Thats 'normal' distribution.

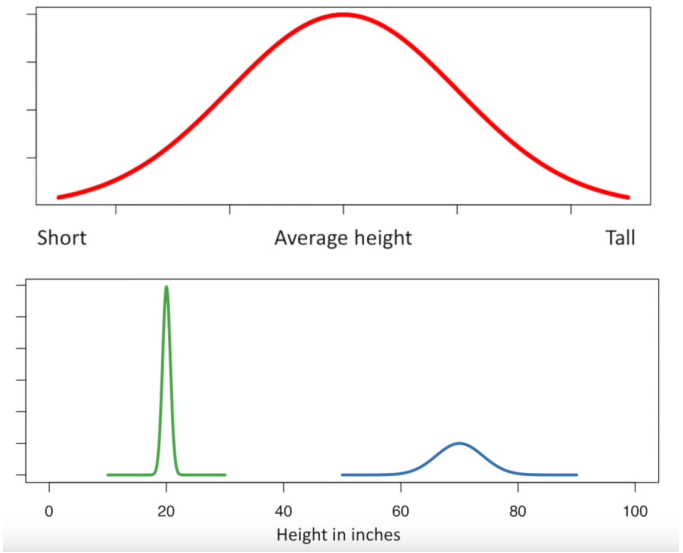
(Ref: StatQuest: What is a statistical distribution? - Josh Starmer)

'Normal' or 'Gaussian' Distribution

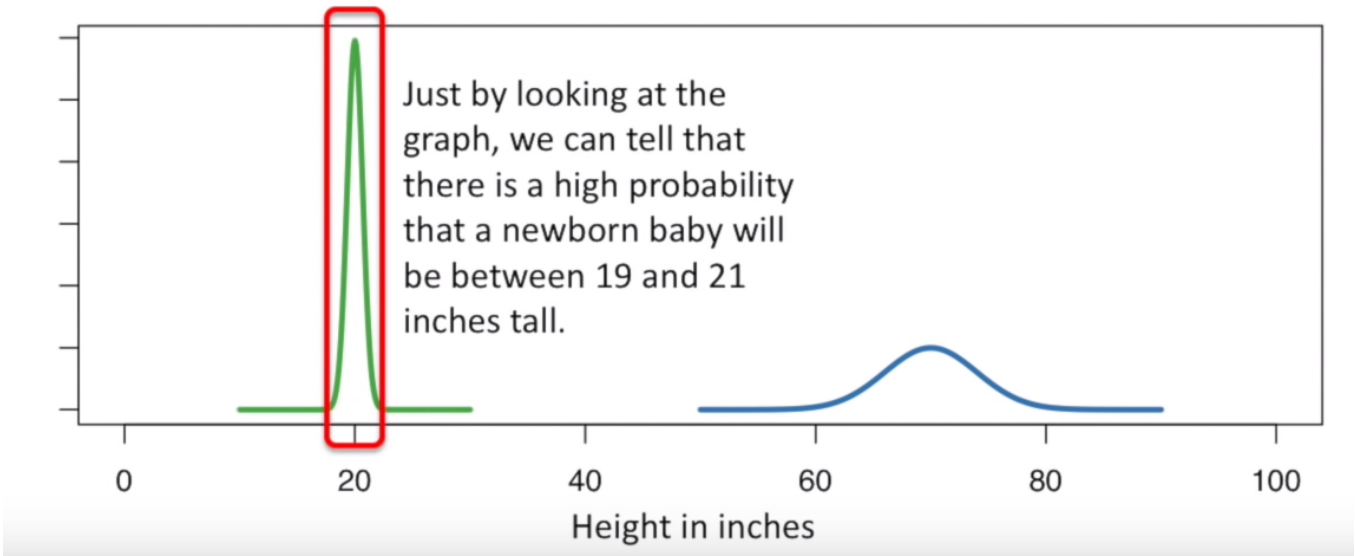
Also called 'Bell Curve'.

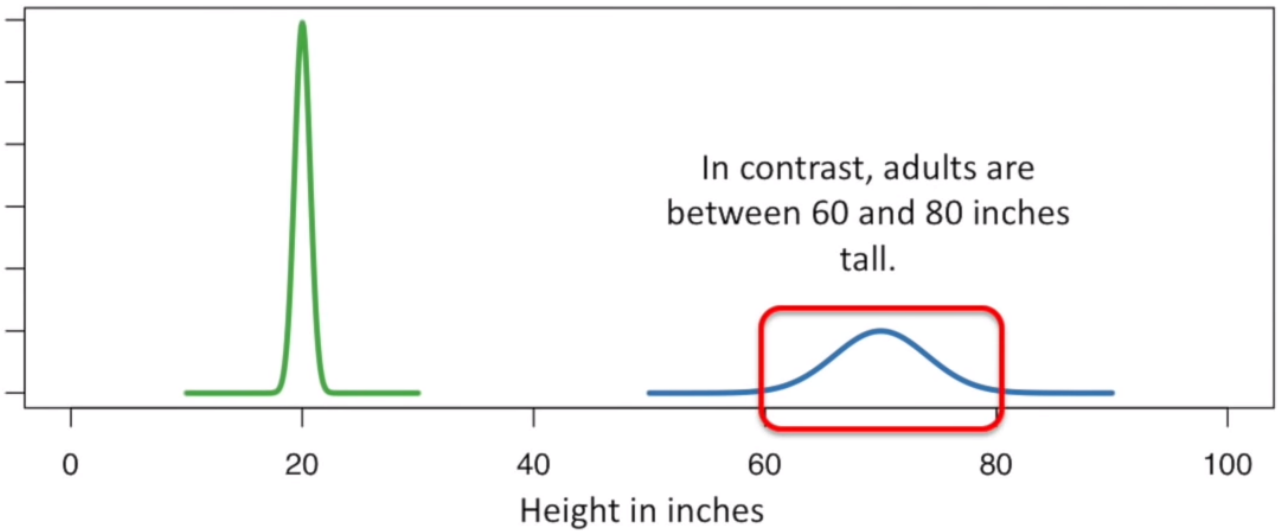
- More people are in the middle region.
- Thats also a relative probability saying that its more likely to find someone of average height.
- Its less likely (low probability) to find someone very tall or very short.
- Another example: two normal distributions of the height of male humans when born and as adults.
- For babies, almost all of them have similar heights (20 inches), but as adults variations are more.

(Ref: StatQuest: What is a statistical distribution? - Josh Starmer)

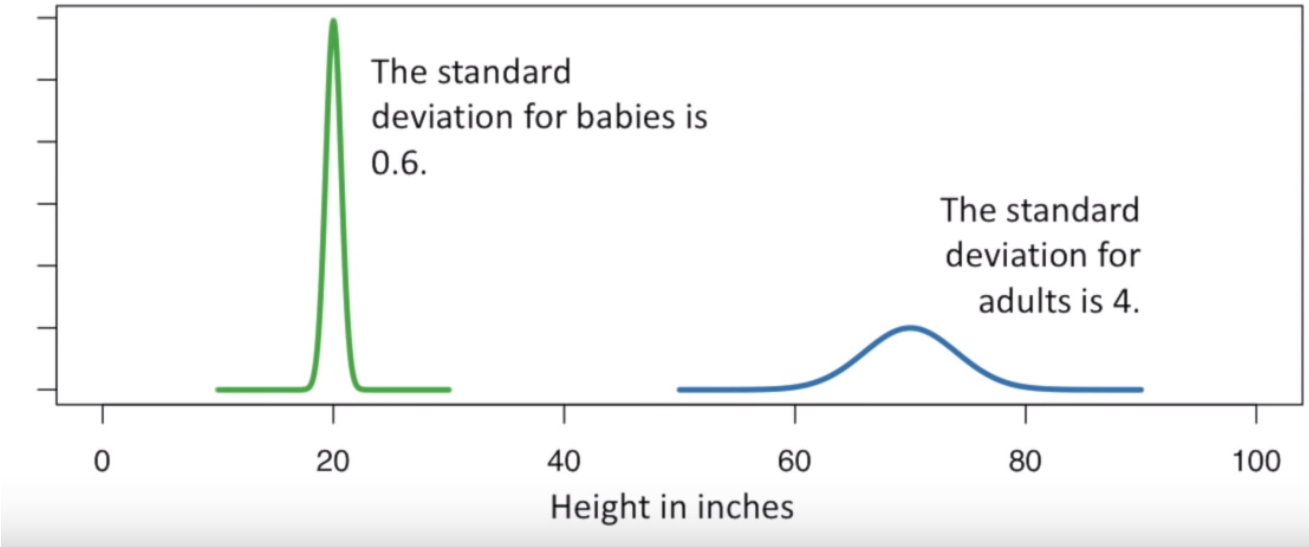


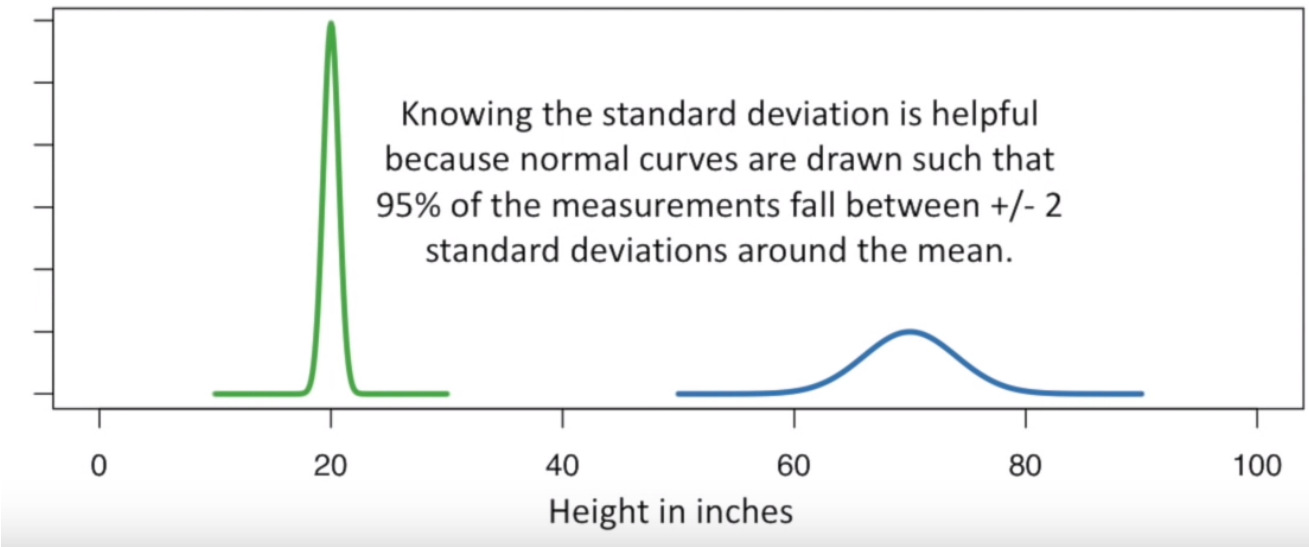
'Normal' or 'Gaussian' Distribution



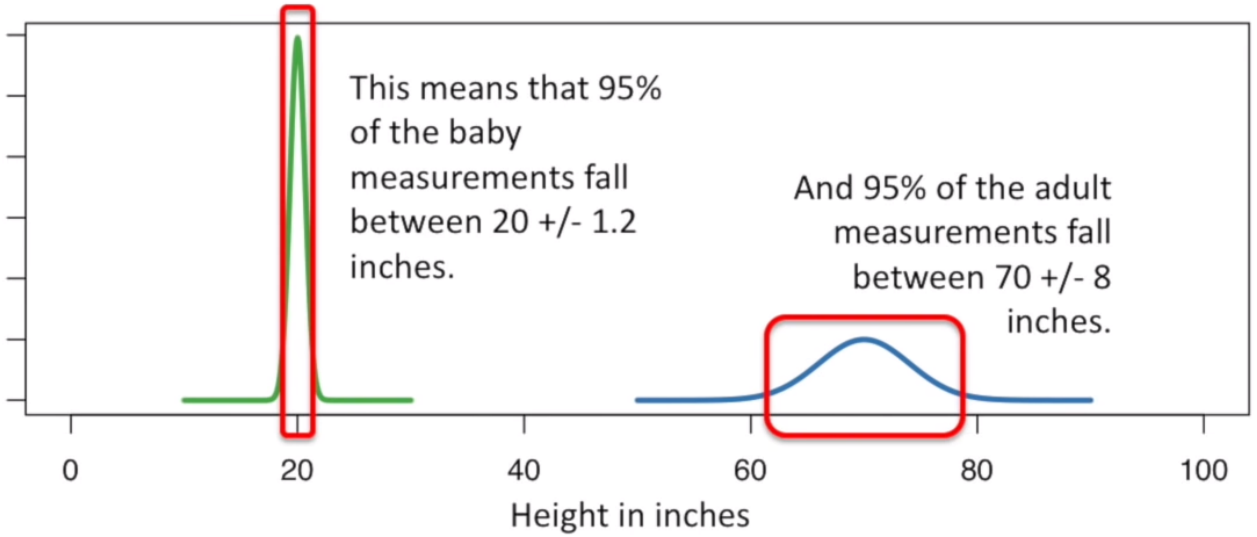


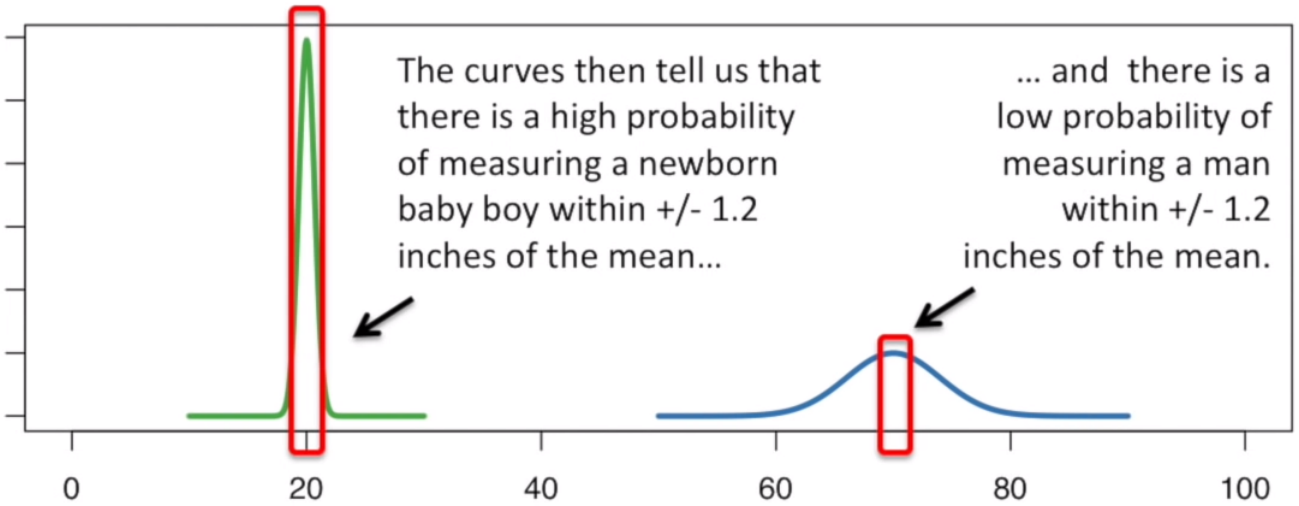
Width of the curve is represented by Standard deviation.





‘Normal’ or ‘Gaussian’ Distribution





'Normal' or 'Gaussian' Distribution

- Normal Distribution is observed widely in nature, eg heights, weights, incomes, etc
- The reason behind this is: The Central Limit Theorem.
- Briefly, if you plot averages of samples, they form normal distribution.

(Ref: StatQuest: What is a statistical distribution? - Josh Starmer)

Mathematically, The normal distribution

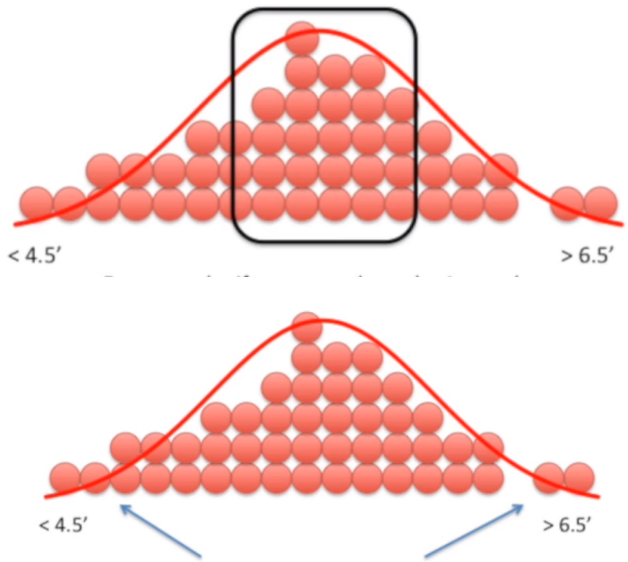
The normal (or Gaussian) distribution is a continuous, symmetric distribution. The **standard normal distribution**, denoted $N(0,1)$ is a normal distribution with mean 0 and variance 1. The normal distribution $N(\mu, \sigma^2)$ has mean μ and variance σ^2.

From the properties of expected values and standard deviations given earlier:

- If Z is standard normal, then $\mu + \sigma Z$ is $N(\mu, \sigma^2)$.
- If Z is $N(\mu, \sigma^2)$, then $(Z - \mu)/\sigma$ is standard normal.

Sampling a Distribution

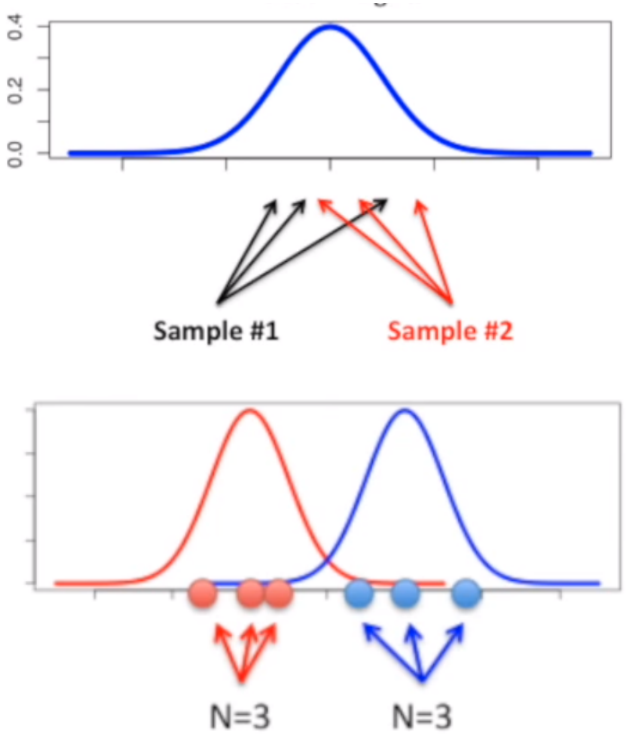
- If we take one sample from a Normal Distribution, most likely it will be a value near middle.
- Once in a while we may get values from ends as well.
- Why do you take samples: to explore the statistics.
- Rather than measuring all, to explore using only a few values.



(Ref: StatQuest: What is a statistical distribution? - Josh Starmer)

Sampling a Distribution

- Tests can be used to see if the samples represent the original population data.
- We can measure similarity between two samples as well, by t-test.
- Since the original distribution form both is same, should give large p value (meaning both are mostly same)
- With just 3 samples for different distributions if p value is small, then increase the sample size.



Calculating normal probabilities meaning sampling normal distribution.

(Ref: StatQuest: What is a statistical distribution? - Josh Starmer)

Calculating normal probabilities

Suppose we have a random variable T which has mean μ and variance σ^2. If we are willing to assume T is normal, how can we calculate $P(T > c)$ for some constant c?
Normal probability tables are available on the web and in most statistics software packages. Often only a table for the standard normal distribution is available, but this is sufficient, since

$$\begin{aligned} P(T > c) &= P((T - \mu)/\sigma > (c - \mu)/\sigma) \\ &= 1 - P(Z \leq (c - \mu)/\sigma). \end{aligned}$$

Thus we can look up the value of $(c - \mu)/\sigma$ in a standard normal probability, table or use a software package, e.g. in R we would use

```
1 1 - pnorm((c-mu)/s)
```

Normal distribution rules of thumb

- The normal distribution is symmetric – around half of a normal sample lies below the mean and half lies above the mean.
- Around 68% of a normal sample lies within one standard deviation of the mean. For example, if we have 1000 points from a normal distribution with mean 10 and variance 4, around 680 of the points will lie between 8 and 12.
- Around 95% of a normal sample lies within two standard deviations of the mean. Continuing with the previous example, around 950 of the points will lie between 6 and 14.
- Around 99% of a normal sample lies within three standard deviations of the mean. Continuing with the previous example, around 990 of the points will lie between 4 and 16.

Normal Distribution

The classic bell curve-shaped distribution and is completely determined by two parameters: its mean (mu) and its standard deviation (sigma).

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Implement.

Use `math.sqrt` and `math.pi` as well as `matplotlib.pyplot` for plotting

```
def normal_pdf(x, mu=0, sigma=1):
  :
  return v
```

Normal Distribution

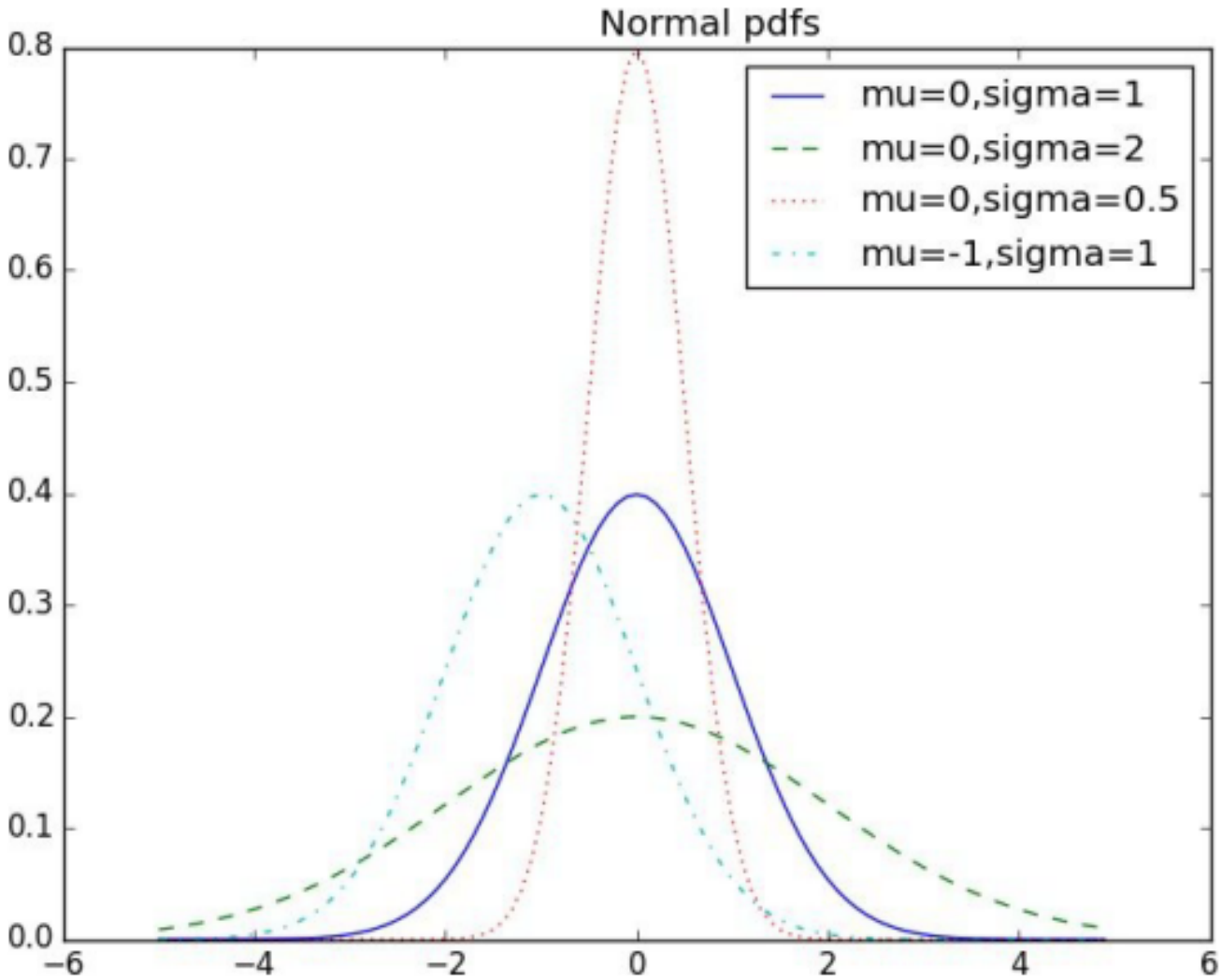
Data and plotting routine

```
xs=[x /10.0 for x in range(-50, 50)]

plt.plot(xs,[normal_pdf(x,sigma=1) for x in xs],'-')
plt.plot(xs,[normal_pdf(x,sigma=2) for x in xs],'--')
plt.plot(xs,[normal_pdf(x,sigma=0.5) for x in xs],':')
plt.plot(xs,[normal_pdf(x,mu=-1) for x in xs],'-')
plt.legend()
plt.title("Normal pdfs")
plt.show()
```

Normal Distribution

```
import math
import matplotlib.pyplot as plt
def normal_pdf(x, mu=0, sigma=1):
  sqrt_two_pi = math.sqrt(2*math.pi)
  return  (math.exp(-(x-mu)**2/2/sigma**2) / (sqrt_two_pi* sigma))
```

Normal Distribution



Log transforms

Some quantities that vary over several orders of magnitude are best analyzed on the log scale.
For example, if we observe these values:

$$14, 28, 3, 60, 39, 13, 1, 9, 3, 55$$

We can take $\log_2$ to get their approximate values as powers of 2:

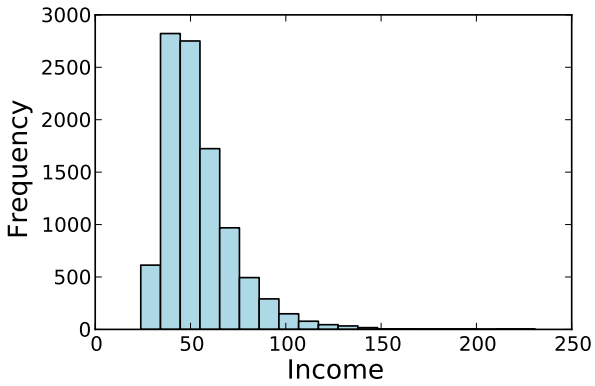
$$3.8, 4.8, 1.6, 5.9, 5.3, 3.7, 0, 3.2, 1.6, 5.8.$$

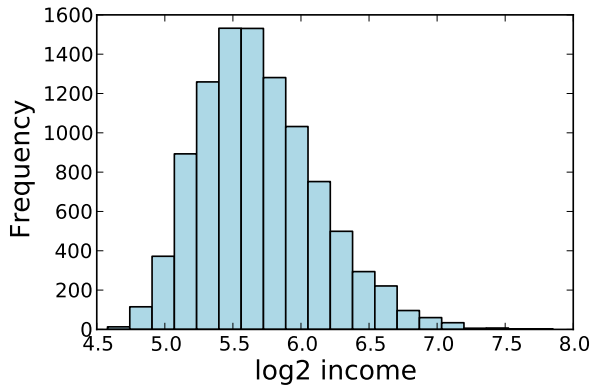
It usually doesn't matter what base is used, since we can convert from one base to another by scaling:

$$\log_b(x) = \log_a(x)/\log_a(b)$$

Symmetrizing effect of log transforms

The log transform symmetrizes right-skewed distributions:





It's common to transform data to make it more symmetric, and usually that's the right thing to do (but don't overdo it...).

Properties of log transforms

Remember the key properties of logarithms:

$$\log(ab) = \log(a) + \log(b) \qquad \log(a^b) = b\log(a).$$

As a consequence, if we take data $X_1, \ldots, X_n$ and scale it to get $Z_i = cX_i$, then

$$\log(Z_1), \ldots, \log(Z_n) = \log(c) + \log(X_1), \ldots, \log(c) + \log(X_n)$$

Thus changing the units of the original data becomes a shift by $\log(c)$ units for the log-transformed data.

Mean values and log transforms

If we observe data $X_1, \ldots, X_n$ and take a log transform to get $Y_i = \log X_i$, then the mean value of the logged data is:

$$\begin{aligned}
\bar{Y} &= n^{-1} \sum_i Y_i \\
&= n^{-1} \sum_i \log X_i \\
&= n^{-1} \log(X_1 \cdot X_2 \cdots X_n) \\
&= \log\left((X_1 \cdot X_2 \cdots X_n)^{1/n}\right).
\end{aligned}$$

$(X_1 \cdot X_2 \cdots X_n)^{1/n}$ is called the **geometric mean** of the X_i, so we see that the usual (arithmetic) mean of the log transformed data is the log of the geometric mean of the untransformed data.

Log transforms

We generally take the log of positive data that is substantially right skewed. If the data are roughly symmetrically distributed, there is no need to take a log transform, and you cannot take a log transform if any of the data values are less than or equal to zero.

Examples: We generally would log-transform income but not age.

Standardizing the sample mean

Suppose we have an id sample of n observations from a population with mean μ and variance σ^2.
We know that $\bar{X}$ has mean μ and variance σ^2/n (so the standard deviation is $\sigma/\sqrt{n}$).
The standardized sample mean is

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}.$$

Example calculation with the normal distribution

Suppose we have a sample $X_1, \ldots, X_{20}$ from a normal population with mean zero. We are told that the probability that $|\bar{X}|$ is greater than 1 is 0.2. What is the standard deviation of the X_i?
Using the fact that the normal distribution is symmetric,

$$\begin{aligned} 0.2 &= P(|\bar{X}| > 1) \\ &= P(\bar{X} > 1) + P(\bar{X} < -1) \\ &= 2P(\bar{X} > 1) \\ &= 2P(\sqrt{20}\bar{X}/\sigma > \sqrt{20}/\sigma) \\ &= 2P(Z > \sqrt{20}/\sigma). \end{aligned}$$

So $P(Z > \sqrt{20}/\sigma) = 0.1$. Now if we look at a table of the normal distribution, we see that the probability of a standard normal value being bigger than 1.28 is 0.1, so $\sqrt{20}/\sigma = 1.28$, and $\sigma = \sqrt{20}/1.28 \approx 3.49$.

Exercises

1. Suppose we observe a sample of 100 values from a normal population with mean 100 and standard deviation 10. Around how many of the values will be greater than 110?
2. Suppose we observe a sample of 150 values from a normal population with mean 80 and standard deviation 12. Write an R expression that will give the approximate number of values between 75 and 85.
3. Suppose we observe a sample of 200 values from a normal distribution with mean zero. Around 20 of the values that we observe are greater than 50. Approximately what is the standard deviation of the population we are sampling from?

Normal Distribution in Practice

Normal Distribution in Practice

- Applied to single variable continuous data e.g. heights of plants, weights of lambs, lengths of time
- Used to calculate the probability of occurrences less than, more than, between given values e.g.
  - The probability that the plants will be less than 70mm
  - The probability that the lambs will be heavier than 70kg
  - The probability that the time taken will be between 10 and 12 minutes
- Standard Normal tables give probabilities

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

How to use Normal Distribution table?

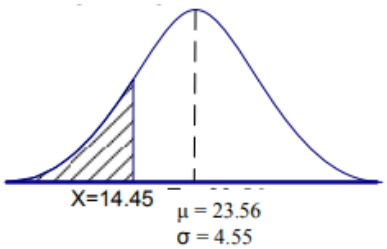
- First need to calculate how many standard deviations above (or below) the mean a particular value is, i.e., calculate the value of the "standard score" or "Z-score".
- Use the following formula to convert a raw data value, X, to a standard score: $Z = \frac{(X-\mu)}{\sigma}$
- eg. Suppose a particular population has $\mu = 4$ and $\sigma = 2$. Find the probability of a randomly selected value being greater than 6
- The Z score corresponding to $X = 6$ is $Z = \frac{(6-4)}{2} = 1$
- $Z = 1$ means that the value $X = 6$ is 1 standard deviation away from the mean
- Use standard normal tables to find $P(Z > 1) = 0.6587$

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Example

Wool fibre breaking strengths are normally distributed with mean $\mu = 23.56$ Newtons and standard deviation, $\sigma = 4.55$. What proportion of fibres would have a breaking strength of 14.45 or less?

- Draw a diagram, label and shade area required



- Convert raw score X to standard score Z: $Z = \frac{(14.45-23.56)}{4.55} = -2.0$
- That is, the raw score of 14.45 is equivalent to a standard score of -2.0.
- It is negative because it is on the left hand side of the curve.
- Use tables to find probability and adjust this result to required probability:

$$\begin{aligned} p(X < 14.45) = p(Z, -2.0) &= 0.5 - p(0 < Z < 2) \\ &= 0.5 - 0.4772 \\ &= 0.0228 \end{aligned}$$

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Inverse Process

To find a value for X, corresponding to a given probability

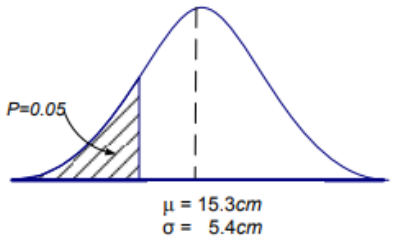
- Draw a diagram and label
- Shade area given as per question
- Use probability tables to find Z -score
- Convert standard score Z to raw score X using inverse formula

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Example

Carrots entering a processing factory have an average length of 15.3 cm, and standard deviation of 5.4 cm. If the lengths are approximately normally distributed, what is the maximum length of the lowest 5% of the load? (i.e., what value cuts off the lowest 5 %?)

- Draw a diagram, label and shade area required



- Use standard Normal tables to find the Z -score corresponding to this area of probability.
- Convert the standard score Z to a raw score X using the inverse formula: $X = Z \times \sigma + \mu$
- For $p(Z < z) = 0.05$ the Normal tables give the corresponding z-score as -1.645. (Negative because it is left of the mean.)
- Hence the raw score is

$$\begin{aligned} X &= Z \times \sigma + \mu \\ &= -1.645 \times 5.44 + 15.3 \end{aligned}$$

Binomial Distribution

Binomial Distribution

Example: Which drink people prefer: Tea or Coffee?

- If 4 out of 7 people prefer Tea, do we say that general population prefers Tea?
- Or it could be that people in general do not have any preference over each other and this observation is just due to random chance and a small sample size.
- If we had surveyed another set of 7 people, we have have had 4 people preferring Coffee.
- In such Yes/No outcomes, we use Binomial distribution as the model.
- We will see how data fits the model.
- If the model is a poor fit, we will reject the idea that both tea and coffee are loved equally (ie there is no preference within them).

(Ref: StatQuest: The Binomial Distribution and Test, Clearly Explained!!! - Josh Starmer)

Binomial Distribution

Example: Which drink people prefer: Tea or Coffee?

- If people really did not prefer Tea over Coffee (and vice versa), then we will assume that there is 50% chance they will pick Tea and 50% chance that they will pick Coffee.
- Out of 3 people, lets calculate probability of first 2 people choosing Tea and the last one choosing Coffee.
- 1st Tea: 0.5
- 2nd Tea: 0.5
- 3rd Coffee: 0.5
- Total: $0.5 \times 0.5 \times 0.5 = 0.125$
- This is true for 3 cases: TTC, TCT, CTT
- So, Probability that any 2 out of 3 people preferring Ta is sum of these $= 0.125 + 0.125 + 0.125 = 0.375$
- This is readily given by $p(x|n,p) = (\frac{n!}{x!(n-x)!})p^x(1-p)^{n-x}$

(Ref: StatQuest: The Binomial Distribution and Test, Clearly Explained!!! - Josh Starmer)

Binomial Distribution

$p(x|n,p) = (\frac{n!}{x!(n-x)!})p^x(1-p)^{n-x}$

- x: number of people preferring Tea (2)
- n : total number of people we asked (3)
- p: probability that someone will pick Tea (0.5)
- The first part of the formula is just combinations formula, how many ways 2 things can be arranged out of 3. Its $= (\frac{3!}{2!(3-2)!}) = 3$.
- p^x shows Tea probability x times, ie $p^x = 0.5^2 = 0.25$
- $(1-p)^{n-x}$ shows the remaining, ie coffee's probability $(1-0.5)^{3-2} = 0.5$
- So, the probability of x(the number of people who prefer Tea), given n (the total number of people asked) and p (probability of picking Tea) is $= (\frac{3!}{2!(3-2)!})(0.5)^x(1-0.5)^{3-2}$

(Ref: StatQuest: The Binomial Distribution and Test, Clearly Explained!!! - Josh Starmer)

Binomial Distribution

Going back to the original example: out of 7 , 4 prefer Tea. Can we say population prefers Tea?

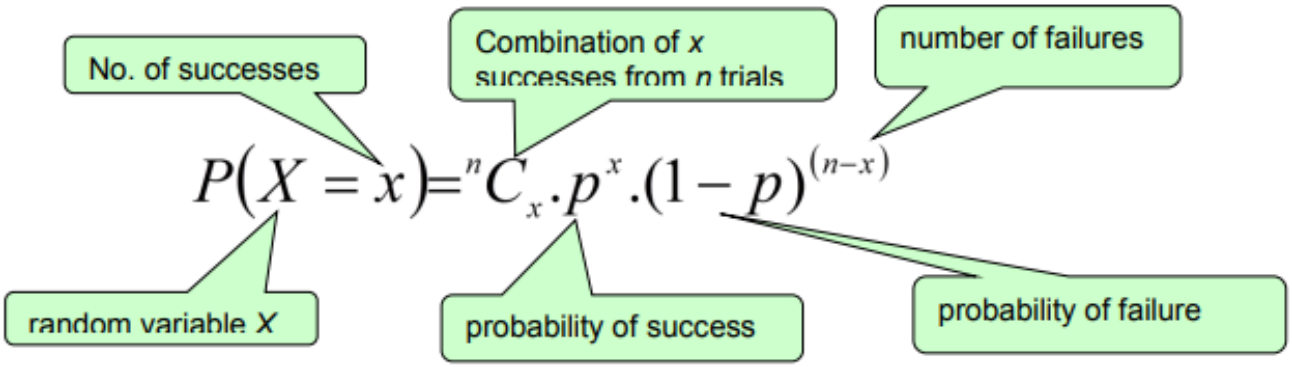
- x: 4
- n : 7
- p: 0.5
- So, the probability someone randomly would pick Tea is $= (\frac{7!}{4!(7-4)!})(0.5)^4(1-0.5)^{7-4} = 35 \times 0.5^4(1-0.5)^3 = 0.273$
- Whats the p value of 4 of 7 people preferring Tea? It is the probability of 4 of 7 people preferring Tea over Coffee, plus the probabilities of all other possibilities that are equally likely or rarer.
- Means, we need to calculate, TTTTCCC,TTTTTCC,TTTTTTC,TTTTTTT. First is observed, reaming three are rare possibilities.
- We will also need probabilities CCCCTTT, CCCCCTT, CCCCCCT, CCCCCCC. With this we are calculating two sided (tailed) p value.
- Tea heavy arrangements give $= 0.273 + 0.164 + 0.055 + 0.008$ similarly for Coffee heavy arrangements. Total probability $0.5 + 0.5 = 1$
- Thus this is a good fit, meaning both Tea and Coffee are equally preferred.
-

Binomial Distribution in Practice

- Applied to single variable discrete data where results are the numbers of "successful outcomes" in a given scenario.
  - Number of times the lights are red in 20 sets of traffic lights
  - Number of students with green eyes in a class of 40
  - Number of plants with diseased leaves from a sample of 50 plants
- Used to calculate the probability of occurrences exactly, less than, more than, between given values
  - The probability that the number of red lights will be exactly 5
  - The probability that the number of green eyed students will be less than 7
  - The probability that the number of diseased plants will be more than 10

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Binomial Distribution in Practice



Read as "the probability of getting x successes is equal to the number of ways of choosing "x successes from n trials" times "the probability of success to the power of the number of successes required" times "the probability of failure to the power of the number of resulting failures."

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Example

An automatic camera records the number of cars running a red light at an intersection (that is, the cars were going through when the red light was against the car). Analysis of the data shows that on average 15% of light changes record a car running a red light. Assume that the data has a binomial distribution. What is the probability that in 20 light changes there will be exactly three (3) cars running a red light?

- Write out the key statistics from the information given
  $p = 0.15, n = 20, X = 3$
- Apply the formula, substituting these values:
  $P(X = 3) = \binom{20}{3} 0.15^3 0.85^{17} = 0.243$
- That is, the probability that in 20 light changes there will be three (3) cars running a red light is 0.24 (24%)

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Poisson Distribution

Poisson Distribution in Practice

This is often known as the distribution of rare events. Firstly, a Poisson process is where DISCRETE events occur in a CONTINUOUS, but finite interval of time or space. The following conditions must apply:

- For a small interval the probability of the event occurring is proportional to the size of the interval.
- The probability of more than one occurrence in the small interval is negligible (i.e. they are rare events). Events must not occur simultaneously
- Each occurrence must be independent of others and must be at random.
- The events are often defects, accidents or unusual natural happenings, such as earthquakes, where in theory there is no upper limit on the number of events. The interval is on some continuous measurement such as time, length or area

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Poisson Distribution in Practice

- The parameter for the Poisson distribution is λ(lambda).
- It is the average or mean number of occurrences over a given interval.
- The probability function is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$
- Example: The average number of accidents at a level-crossing every year is 5. Calculate the probability that there are exactly 3 accidents there this year.
- Here, $\lambda = 5$, and $x = 3$.
- $p(X = 3) = \frac{e^{-5}5^3}{3!} = 0.1404$
- That is, there is a 14% chance that there will be exactly 3 accidents there this year

(Ref: Normal, Binomial, Poisson Distributions - Lincoln University)

Inferential Statistics

What is Inferential Statistics?

Inferring from whats not fully available.

- Discover properties of larger group by studying smaller group, with a quantified confidence that results generalize well with the larger group.
- There is a chance (here comes Probability) that sample's behavior can be similar to Population's.
- Used to compare variables/groups

Inferential Statistics Objectives

- To estimate, predict a population 'parameter' using sample value (called 'statistic')
- Eg Predicting election results based of Exit Polls.
- Test Hypotheses.
- E.g. Whether the new method realy works or not.

Estimate/Predict Population Parameter

- Population: Big/Whole dataset
- Sample: Small/test dataset
- Point Estimate: using a single value of sample (called 'statistic') to predict corresponding value in the population (called 'parameter')
- E.g. Mean of sample can very well be used to estimate Population mean.

Test Hypothesis

- Types: Null Hypothesis (H_0) and Alternate Hypothesis (H_1)
- Null Hypothesis: two groups under study are same
- Alternate Hypothesis: two groups are different
- Goal: to prove Null Hypothesis wrong

Inferential Statistics Types

Based on type of distribution:

- Parametric: Variable is assumed normally distributed
- NonParametric: Distribution of scores is severely skewed

Inferential Statistics Types

Based on type of underlying variable:

- For nominal variables: compare distributions, central tendencies, measure spreads
- For ordinal variables: non-parametric tests, rank order differences
- Continuous: Ratio or Interval variables: regression analysis

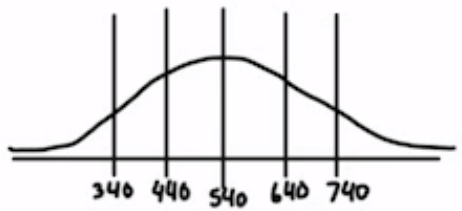
Some Statistical Terms

Z-score

- Meaning: number of std deviations a value is from the mean
- Example:
  - $\mu = 540$
  - $\sigma = 100$
  - $z =$?

GMAT Scores

Average: 540
Standard deviation: 100



Value	Z-score
640	1
340	-2
540	0
690	1.5
420	-1.2

$$\frac{420 - 540}{100} \Rightarrow \frac{-120}{100} = -1.2$$

Z-score

- 1 std deviation would be 540 ± 100 ie 640 or 440
- 2 std deviations would be 540 ± 200 ie 740 or 340
- For 690? Simple intrapolation $\frac{690-540}{100} = 1.5$
- $z = \frac{x-\mu}{\sigma}$
- z of a particular value tells us percentile

Z-score

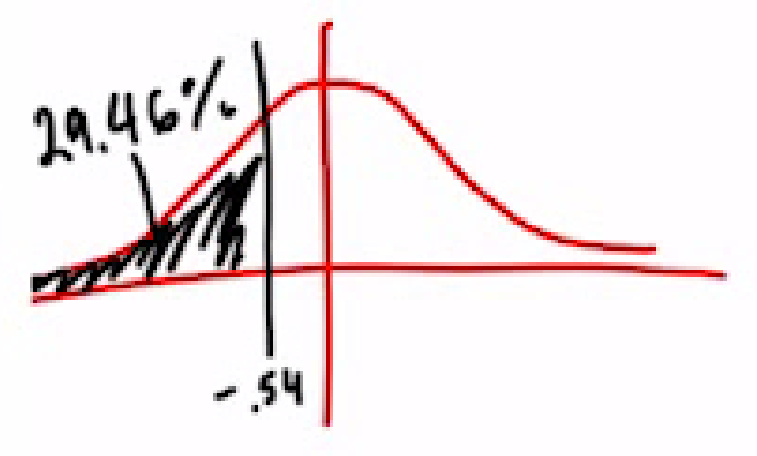
z-table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877

- To look for z score of -0.54 go to row -0.5 and then find column for 0.04. Its 0.2946
- This value tells us proportion of values to the left of that z
- So, for $z = -0.54$ there are 29.46% values to the left.

Z-score

z of a particular value tells us percentile



- You can go backwards as well
- If you want score which is at 95% percentile, then
- Find z value from table (almost 1.645)
- Find x

Inferences for Normal Distribution

GMAT example:

- Finding percentile for score 705
- Find z score first: $z = \frac{x-\mu}{\sigma} = \frac{705-540}{100} = 1.65$
- Go to z table, find the value, make it percentile: 95.05%

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803

Quiz

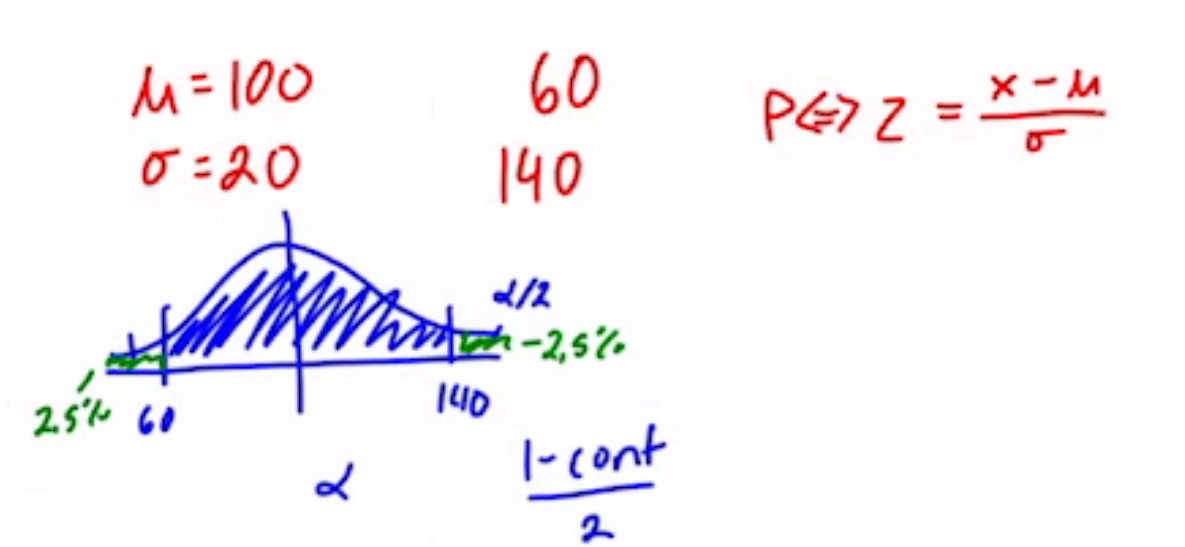
In a telephone call center, the typical agent averages 86 phone calls per day with a standard deviation of 11 calls. Management would prefer that agents make between 80 and 100 phone calls each day. What percentage of agents are meeting management's expectations?

Solution

- z for 100 calls: 100 - 86/11 = 1.27
- z for 80 calls: 80 - 86/11 = -.545
- value from table for z=1.27: 0.8980
- value from table for z=-.545: 0.2929
- Difff between them are the people doing correct amounts of call: about 60%

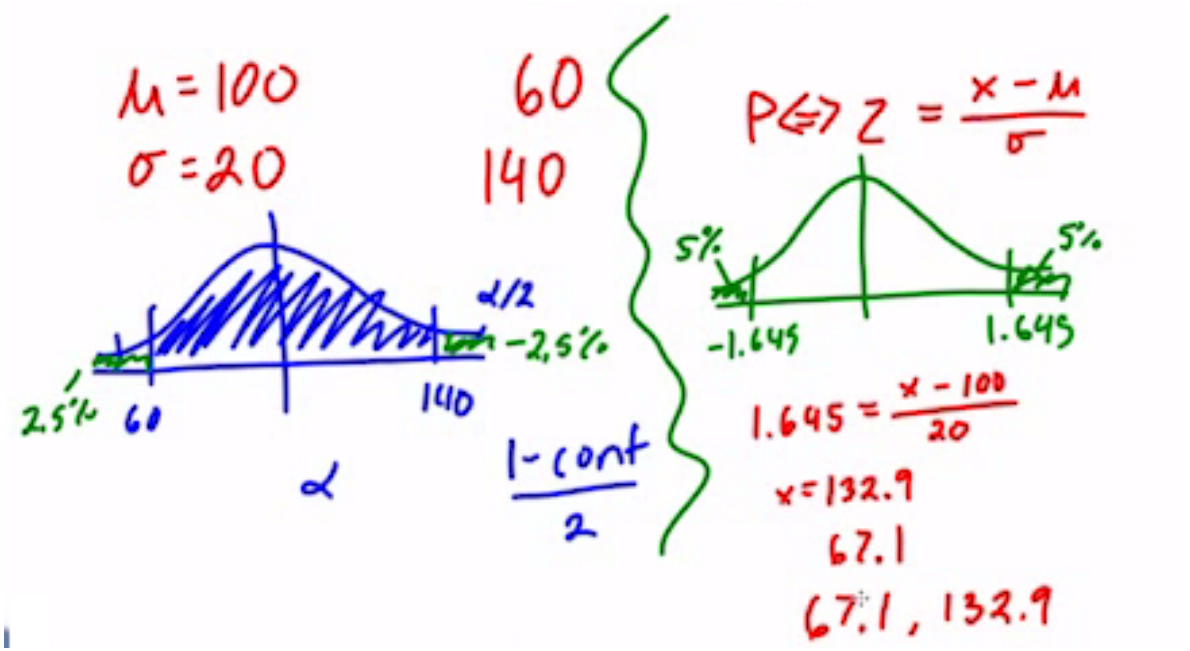
Confidence Interval

- In a restaurant, average 100 customers come, std dev is 20.
- 95% always falls between 2 std dev, so between 60 and 140.
- So, theer is 95% confidence that the number of customers will be between 60 to 140.
- Remaning non-confident portion is α so $\alpha/2$ on both sides.
- What if we want to be only 90% confident?



Confidence Interval

- What if we want to be only 90% confident?
- so $\alpha/2 = 0.05$
- Find z for that: Its about - 1.645
- Higher band is 1.645
- Caluclate x for both upper and lower boundaries: 132.9 and 67.1
- For 90% confidence, guest would be in range 67 to 133 guests.



Thanks ... yogeshkulkarni@yahoo.com