Statistics and Estimation for Computer Science

İstanbul Teknik Üniversitesi

Mustafa Kamasak, PhD

Version: 2022.2.22

Descriptive Statistics

Descriptive Statistics

- Data should be explored to understand how it is distributed
  - Central tendency
  - Spread
  - Symmetricity
  - Flatness
- Data should be proprocessed
  - Invalid data
  - Missing values
  - Outliers
  - Normalized/standardized
  - Transformed
- Data should be visualized
  - Line plot
  - Bar graph
  - Histogram
  - Boxplot
  - ...

Mean - Central Tendency

- Population mean ($\mu = E(X)$) is not a random variable
- Sample mean ($\overline{x}$) is used as a measure of central tendency of distribution

$$\overline{x} = \frac{1}{N} \sum_{i=0}^{N-1} x_i$$

  - N is the sample size (number of instances in the sample)
  - x_i is the i^{th} instance in the sample
- For example:

$$x = [1,\ 3,\ 11,\ 5,\ 6]$$

  Then

$$\overline{x} = \frac{1}{5}(1 + 3 + 11 + 5 + 6)$$

Range of Data - Spread of distribution

- **Measure of data dispersion**
- Range of data is the difference of maximum and minimum values in the data
- For example

$$x = [1,\ 3,\ 11,\ 5,\ 6]$$

Then

$$\text{Range } x = 11 - 1 = 10$$

Standard Deviation - Spread of distribution

- Population std. dev. ($\sigma = E(X-\mu)^2$) is not a random variable
- Sample std. dev. (s) is also a **measure of data dispersion**

$$s = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \overline{x})^2}$$

  - Why $(x_i - \overline{x})^2$ instead of $(x_i - \overline{x})$?

  A $\sum_{i=0}^{N-1}(x_i - \overline{x})$ is always 0.

  - Why $N-1$?

  A Will be explained later.

- For example:

$$x = [1,\ 3,\ 11,\ 5,\ 6]$$

Then

$$s = \sqrt{\frac{1}{4}(1-\overline{x})^2 + \cdots + (6-\overline{x})^2}$$

Outliers (Aykırılıklar)

There may be outlier values in the data

$$x = [1,\ 3,\ 11253,\ 5,\ 6]$$

- Outliers may or may not be spurious data caused by temporary errors or rarely seen correct data point. It can never be known.
- Their probability of appearance is very low
- They are different than (further from) normal data
- They substantially affect estimated parameters

$$x = [1,\ 3,\ 11,\ 5,\ 6] \rightarrow \bar{x} = 5.2$$

$$x = [1,\ 3,\ 11253,\ 5,\ 6] \rightarrow \bar{x} = 2253.6$$

- Outliers are detected and cleaned
- Parameter estimation methods exist that are robust to outliers $\rightarrow$ Use ranks instead of values

Median (Ortanca)

- Median is also a measure of central tendency of a distribution
- Mean is sensitive to outliers $\rightarrow$ use median
- Order data in ascending way and assign ranks

$$x = [1,\ 3,\ 5,\ 6,\ 11253]$$
$$ranks = [1,\ 2,\ 3,\ 4,\ 5]$$

- Mean of rank is

$$\bar{r} = \frac{1}{5}(1+2+3+4+5) = 3$$

$$\text{Median} = x[\bar{r}] = 5$$

- If $\bar{r}$ is not integer (happens when sample size is even), then the average of indices around $\bar{r}$ is used.
- For example, if $\bar{r} = 3.5$, then median$=0.5(x[3] + x[4])$

Median

$$x = [1,\ 3,\ 11,\ 5,\ 6] \rightarrow \text{Median } x = 5,\ \overline{x} = 5.2$$

$$x = [1,\ 3,\ 11253,\ 5,\ 6] \rightarrow \text{Median } x = 5,\ \overline{x} = 2253.6$$

Median

- Use of mean/median is also important when population distribution is skewed
  - Typically for symmetric (no-skew) distributions mean$\approx$median
  - Right-skewed dist $\rightarrow$ mean $>$ median
  - Left-skewed dist $\rightarrow$ mean $<$ median

- Income data is very right-skewed. Consider mean personal income in US. What happens if you take out billionaires?

Trimmed Mean

- Trim data at the lower and higher tails before computation of mean
- For 10% trimmed mean, 5% of the upper and 5% of the lower data points are removed.
- The rest of the data (90%) is used to compute mean.
- Not preferred if sample size is small

Percentile (Yüzdelik)

- Range is sensitive to outliers $\rightarrow$ use percentiles
- Percentile of a data is the percentile of data that is smaller or equal to the value
- For example 15^{th} percentile of the data corresponds to the value for which 15% of the values are smaller or equal to that value.

- 25^{th} percentile $\rightarrow$ 1^{st} quartile
- 50^{th} percentile $\rightarrow$ 2^{nd} quartile / median
- 75^{th} percentile $\rightarrow$ 3^{rd} quartile

$$x = [1,\ 3,\ 5,\ 6,\ 11253]$$
$$percentile = [20, 40, 60, 80, 100]\%$$

Interquatile Range (IQR)

- Standard deviation is sensitive to outliers $\rightarrow$ use IQR

$$x = [1,\ 3,\ 11,\ 5,\ 6] \rightarrow s = 3.37$$

$$x = [1,\ 3,\ 11253,\ 5,\ 6] \rightarrow s = 4499.70$$

- IQR is defined as the difference between 3^{rd} and 1^{st} quartile
- Robust estimator of standard deviation

- IQR is also used for outlier detection
- Values higher than $Q3 + 1.5 * IQR$ are outliers
- Values lower than $Q1 - 1.5 * IQR$ are outliers

Interquatile Range (IQR)

$$x = [1,\ 3,\ 5,\ 6,\ 11253]$$
$$ranks = [1,\ 2,\ 3,\ 4,\ 5]$$
$$percentile = [20, 40, 60, 80, 100]\%$$

- 1^{st} quartile $\rightarrow Q1 = \frac{1*3+3*1}{4} = 1.5$ (linear interpolation)
- median $\rightarrow Q2 = 5$
- 3^{rd} quartile $\rightarrow Q3 = \frac{5*1+6*3}{4} = 5.75$

- IQR = 5.75-1.5 = 4.25
- Lower limit = 1.5-1.5*4.25 = -4,875
- Upper limit = 5.75+1.5*4.25 = 12.125
- $11253 \notin [-4,875,\ 12.125] \rightarrow$ outlier!

IQR

$$x = [1,\ 3,\ 11,\ 5,\ 6] \rightarrow IQR = 4.25,\ \ s = 3.37$$

$$x = [1,\ 3,\ 11253,\ 5,\ 6] \rightarrow IQR = 4.25,\ \ s = 4499.70$$

Outlier Detection - Z-score

- Z-score can be thresholded to detect outliers
- Z-score

$$z_i = \frac{x_i - \overline{x}}{s}$$

  where $\overline{x}$ is the data mean, and s_x is the standard deviation of data.

$$\overline{x} = \frac{1}{N} \sum_i x_i$$

$$s = \sqrt{\frac{1}{N-1} \sum_i (x_i - \overline{x})^2}$$

- Typically $|Z| \geq 2.5$ can be assumed outlier
- May not be good for asymmetric (skewed) distributed data

Outlier Detection - Hypothesis Testing

- There are hypothesis testing based methods as well
- Grubb's test

$$G = \frac{\max_i |x_i - \overline{x}|}{s}$$

- G statistic is thresholded by

$$\frac{N-1}{\sqrt{N}} \sqrt{\frac{t_\alpha^2}{N - 2 + t_\alpha^2}}$$

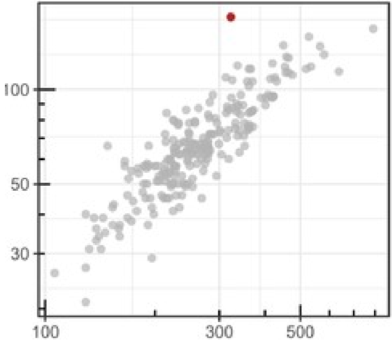
- Dixon's Q-test

$$Q = \frac{\min_i |x_j - x_i|}{\text{range}}$$

  where x_j is the data point tested for being outlier
- Q statistics thresholded by values obtained from table
- Both tests require normal distributed data

Outliers in Multivariate Data

- Data is typically multivariate/multidimensional
- For each instance, a vector is obtained
- For example, a person's age, height, weight is a 3-tuple data which are highly correlated
- For multivariate data, iqr & z-score may not be enough $\rightarrow$ Model data and find abnormalities
- With more than 2 variates, it is easy to visualize/detect by looking



Outliers in Multivariate Data - Dbscan

- Dbscan (Density Based Spatial Clustering of Applications with Noise)
- Groups together points that are closely packed together
- Inputs
  - Distance metric
  - Radius for neighborhood ϵ
  - `minPts` used to define core points

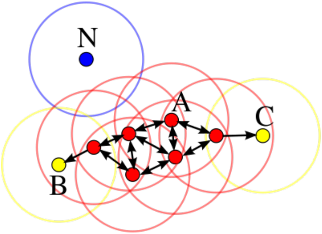


Image from `https://en.wikipedia.org/wiki/DBSCAN`

Outliers in Multivariate Data - Dbscan

- with `minPts` $= 4$
- Point A and the other red points are core points (area surrounding these points in ϵ radius contain at least 4 points including the point itself).
- They are all reachable from one another, they form a single cluster.
- Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well.
- Point N is a noise point that is neither a core point nor directly-reachable.

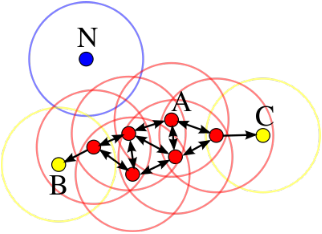


Image from https://en.wikipedia.org/wiki/DBSCAN

Outliers in Multivariate Data - Isolation Forest

- Outliers are by definition few and different
- A binary tree is formed by
  - Selection a random dimension
  - Selection a random value between [min, max] values of this dimension
- Using subsets of data, different binary trees can be formed $\rightarrow$ forest
- Isolation forest algorithm requires unlabeled data $\rightarrow$ unsupervised

Outliers in Multivariate Data - Isolation Forest

- With a new instance, length of path from each tree in the forest is computed and averaged
- Typically, outliers have shorter paths compared to normal data points

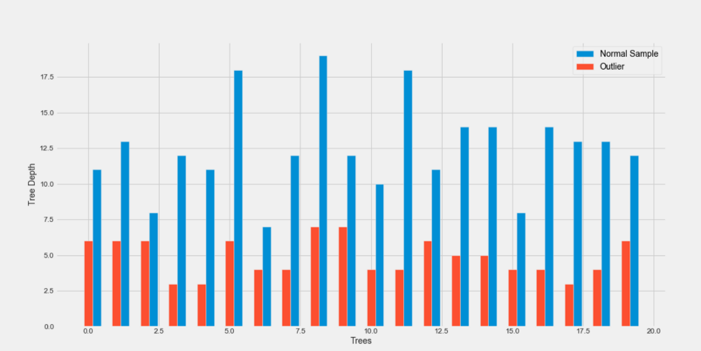


Figure from `https://towardsdatascience.com/isolation-forest-from-scratch-e7e5978e6f4c`

Handling Outliers

- Detected outliers can be deleted if its believed to be impossible
- Truncation: Set all values above a lower and upper limit to the limit

$$x_i \begin{cases} \ell & \text{if } x_i \leq \ell \\ x_i & \text{if } \ell \leq x_i \leq u \\ u & \text{if } x_i \geq u \end{cases}$$

- Winsoring: Set all outliers to a specified percentile of the data
- 90% Winsorizing means
  - data below 5% is set to 5%
  - data above 95% is set to 95%

```
from scipy.stats.mstats import winsorize

winsorize([92, 19, 101, 58, 1053, 91, 26, 78,
    10, 13, -40, 101, 86, 85, 15, 89, 89, 28,
    -5, 41], limits=[0.05, 0.05])
```

Moments of Data

- When the distribution of data will be investigated, higher moments are used.
- Definition of a moment of a function/distribution around a point c is defined as follows:

$$M^r = \sum_i (x_i - c)^r f(x_i)$$

- If the moment is taken around mean, then it is named central moment.
- If the moment is normalized with standard deviation

$$sM^r = \frac{\sum_i (x_i - c)^r f(x_i)}{\sigma^i}$$

Moments of Data

- First moment around 0 $\rightarrow$ mean
- Second central moment $\rightarrow$ variance
- Third standardized central moment $\rightarrow$ skewness (çarpıklık)
- Fourth standardized central moment $\rightarrow$ kurtosis (basıklık)

Skewness (Çarpıklık)

- Third standardized central moment - not a random variable for population distribution

$$\text{Skewness} = E\left(\frac{(X-\mu)^3}{\sigma^3}\right)$$

- **Measure of asymmetry**
- For symmetric distribution (such as Normal distr.) skewness=0

- For positive skew (right skewed) skewness $>$ 0)
- Positive skew $\rightarrow$ larger tail above mean

- For negative skew (left skewed) skewness $<$ 0
- Negative skew $\rightarrow$ larger tail below mean

Skewness

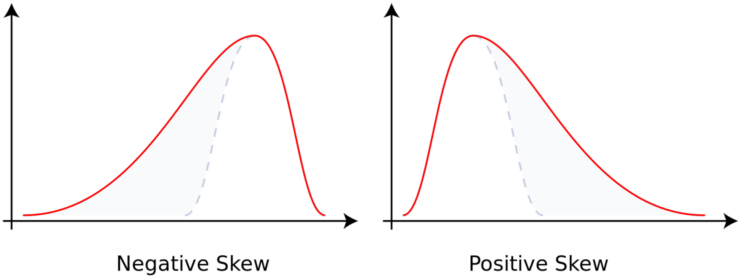


Figure taken from `https://en.wikipedia.org/wiki/Skewness`

Kurtosis (Basıklık)

- Fourth standardized central moment

$$\text{Kurtosis} = E(\frac{(X-\mu)^4}{\sigma^4})$$

- **Measure of peakedness/tailedness**

- For Normal distribution kurtosis $= 3$

Kurtosis

- More peak/less tail – kurtosis > 3
- Less peak/more tail – kurtosis < 3

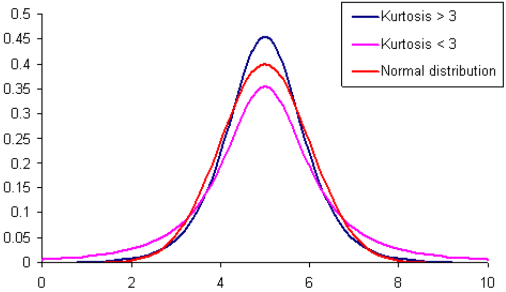


Figure taken from `https://modelassist.epixanalytics.com/display/EA/Kurtosis`

Kurtosis

- Sometimes, kurtosis is defined with respect to Normal distribution
- Excess kurtosis is defined as

$$\text{Excess Kurtosis} = E\left(\frac{(X-\mu)^4}{\sigma^4}\right) - 3$$

- For Laplace distr excess kurtosis is 3
- For Normal distr excess kurtosis is 0
- For uniform distr excess kurtosis is -1.2

Kurtosis

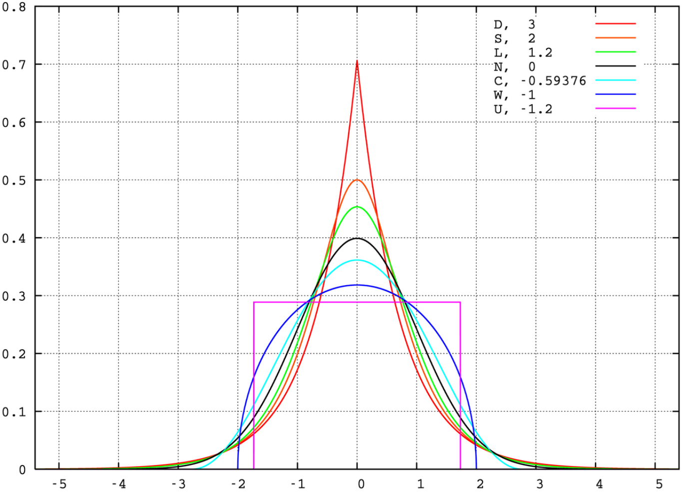


Figure taken from `https://en.wikipedia.org/wiki/Kurtosis`

Kurtosis

Classification of distributions in terms of their kurtosis:

- Mesokurtic/Mesokurtotic – Distr with zero excess kurtosis
- Leptokurtic/LeptoKurtotic – Distr with positive excess kurtosis
- Platykurtic/Platykurtotic – Distr with negative excess kurtosis

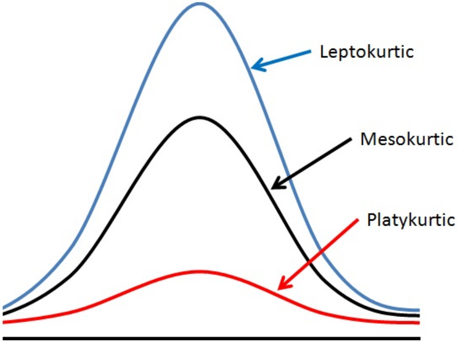


Figure taken from https://www.analystforum.com/forums/cfa-forums/cfa-level-ii-forum/91346370

Skewness/Kurtosis from Data

- Use $\overline{x}$ isntead of μ
- Use s instead of σ
- Use averaging instead of expectation

$$\text{Skewness} = \frac{\frac{1}{N}\sum_i^N (x_i - \overline{x})^3}{s^3}$$
$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_i^N (x_i - \overline{x})^4}{s^4}$$

Skewness/Kurtosis from Data

How to use skewness/kurtosis of data:

- With $\overline{x}$ and s, they give extra information about data distribution
- There are tests that use skewness/kurtosis to test if the data has normal distribution

Visual Methods for Distribution

- Need to check if data comes from a certain distribution
- Typically computed moment values gives hint about data distribution
- Inspection of data conformity with a probability distribution can be visualized
- There are also statistical methods $\rightarrow$ will be covered later

- Visual inspection of probability distribution of 2 sources to see if both comes from the same distribution or not

PP Plot

- Plot emprical cdf vs theoretical cdf
- Plot two emprical cdf against eachother.
- If the plot is on a straight line, data comes from that distribution family (with different mean/std. dev)
- If the plot is on a 45 degree straight line, data comes from the distribution (with the same mean/std. dev)

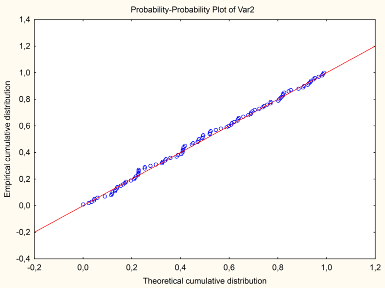


Figure taken from `https://en.wikipedia.org/wiki/P-P_plot`

QQ Plot

- Plot emprical quartiles vs theoretical quartiles
- Plot two emprical quartiles against eachother.
- If the plot is on a 45 degree straight line, data comes from that distribution family (with different mean/std. dev)

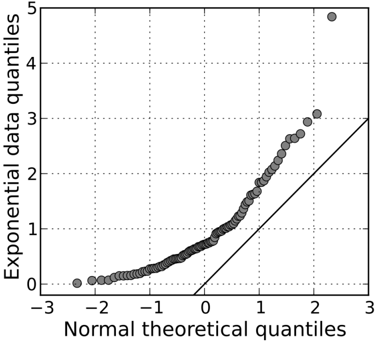
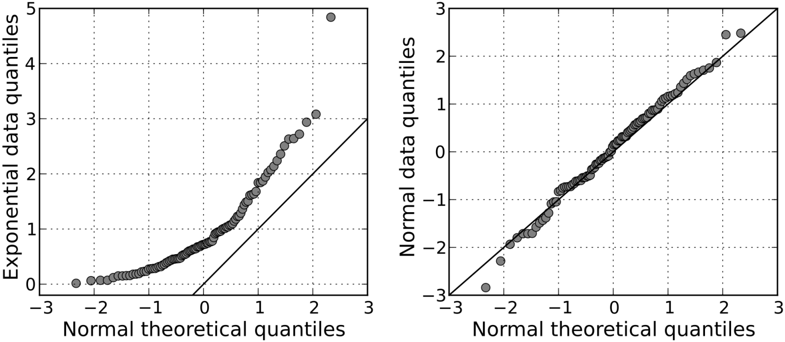


Figure taken from `https://en.wikipedia.org/wiki/Q-Q_plot`

QQ Plot



Figures taken from `https://en.wikipedia.org/wiki/Q-Q_plot`

Data Standardization

- Data come with various mean, variance, range etc.
- Sometimes data normalization/standardization is required to handle multivariate data
- There are various methods
  - Min-max normalization

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

  - Outlier cause normal data to squeeze in a small range
  - z-score standardization

$$\text{z-score}_i = \frac{x_i - \overline{x}}{s}$$

  - z-score: distance an observation from the mean, expressed in standard deviation units

Data Transformation

- Some algorithms may require certain data distributions (such as normal)
- Data should be transformed to have a certain distribution - for example: Data may have have exponential distribution $\rightarrow$ need to have normal distr.
- Recall from probability theory: Let X be a random variable with
  - probability distribution function (pdf) $f(x)$
  - cumulative distribution function (cdf) $F(x)$
- $U = F_X(x)$ random variable U will have uniform distribution
- Let Z be a random variable with cdf $F_Z()$. $F_Z^{-1}(u)$ will convert uniform distributed U in to Z
- There are direct transformations
  - Box-Cox transform (exponential$\rightarrow$normal dist)
  - Box-Muller transform (uniform$\rightarrow$normal dist)

Multivariate Data -Covariance

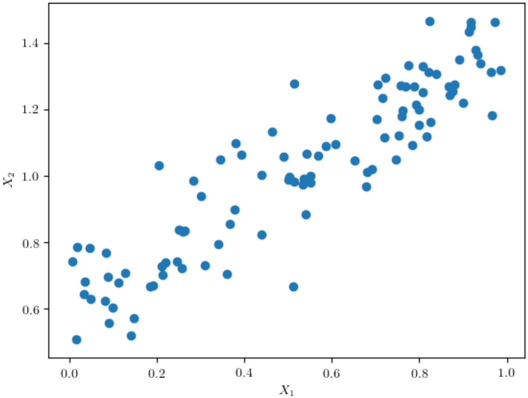
- When we deal with multivariate data, relation of variables against eachother is important
- Co $\rightarrow$ together
- vary $\rightarrow$ change
- Co-variance is a measure of **linear** change of variables

$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$

- Covariance $\approx 0 \rightarrow$ Uncorrelated (not independent), may have nonlinear relation
- High positive covariance $\rightarrow$ Variables change linearly at the same direction
- High negative covariance $\rightarrow$ Variables change linearly at different direction
- High?

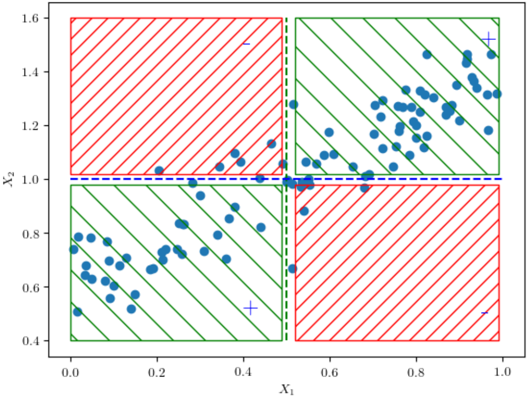
How does Covariance Works

Consider X_1 and X_2 who are related as follows



How does Covariance Works

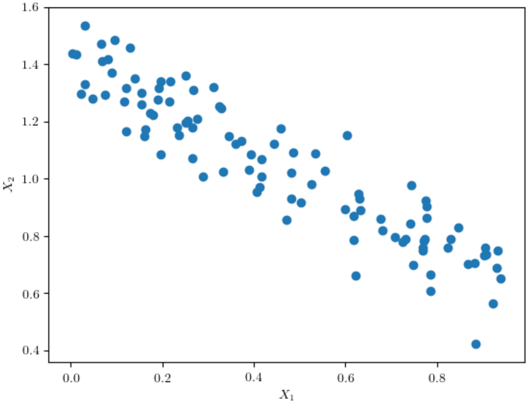
$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$



Covariance is **positive**.

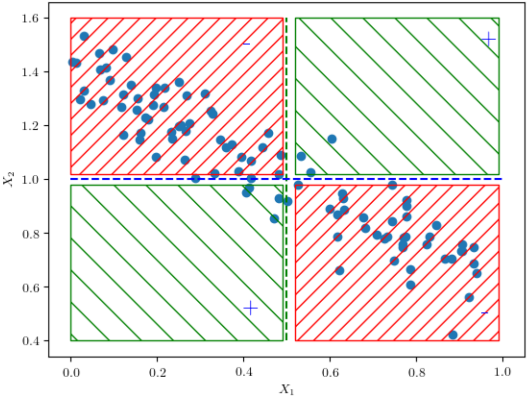
How does Covariance Works

Consider X_1 and X_2 who are related as follows



How does Covariance Works

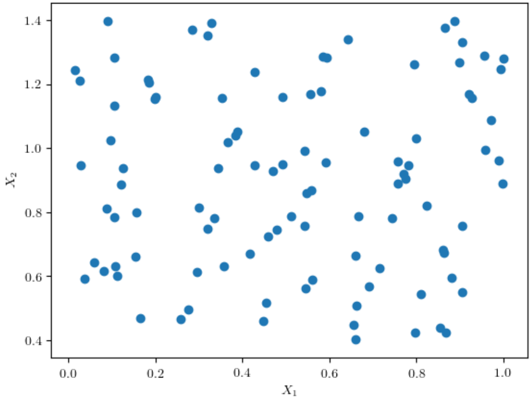
$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$



Covariance is **negative**.

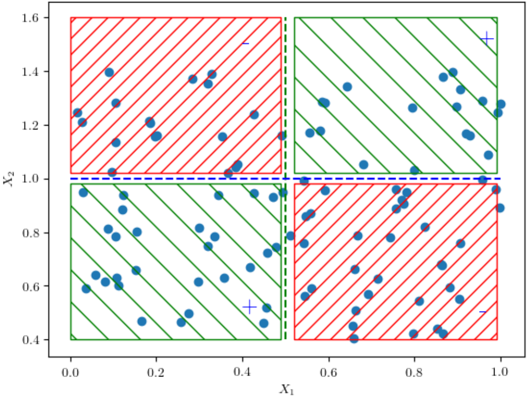
How does Covariance Works

Consider X_1 and X_2 who are related as follows



How does Covariance Works

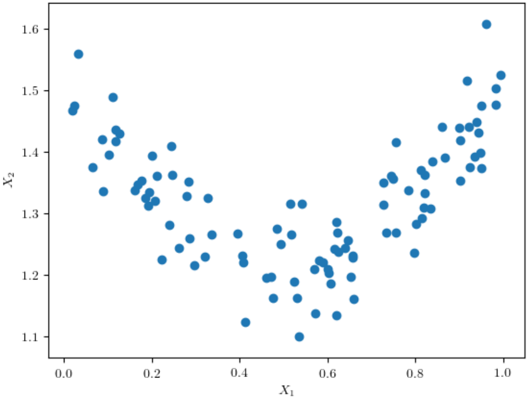
$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$



Covariance is **close to zero**. X_1 and X_2 seems unrelated.

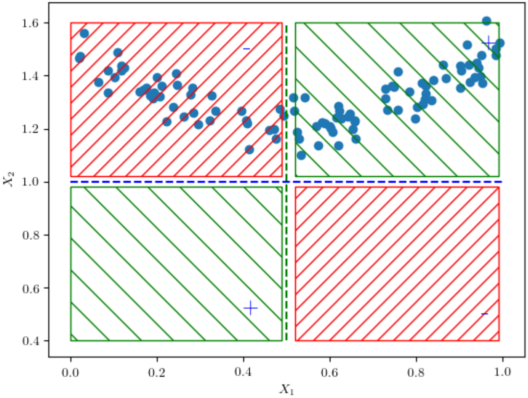
How does Covariance Works

Consider X_1 and X_2 who are related as follows



How does Covariance Works

$$Cov(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$



Covariance is **close to zero**. X_1 and X_2 are definitely related.

Covariance to Correlation Coefficient

- Covariance has no limits
- Covariance is related to units

Pearson Correlation Coefficient

- Pearson correlation coefficient
- Definition:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}$$

- Correlation coefficient ρ is a value in [-1,1] range.
- $-1 \leq \rho_{XY} \leq 1$ and $|\rho_{XY}| \leq 1$
- $|\rho_{XY}| = 1$ when X and Y are linearly related.
- $|\rho_{XY}| = 0$ when X and Y are uncorrelated.
- Uncorrelated does not mean independent (except Normal distr)

  Correlation coefficient between
  - height & weight ?
  - IQ & GPA ?
  - IQ & Income ?

Pearson Correlation Coefficient of Data

- Use averaging for expectation
- Use sample mean ($\overline{x}$) for population mean (μ)
- Use sample std mean (s) for population std dev (σ)
- Definition:

$$r_{xy} = \frac{\frac{1}{N-1}\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

- Correlation coefficient can only indicate linear relations
- Sensitive to outliers
- $|r| > 0.8$ means strong correlation
- $|r| < 0.3$ means weak correlation
- $r = 0$ means no correlation $\rightarrow$ uncorrelated
- Uncorrelated does not mean unrelated. It means no linear relation.

Pearson Correlation Coefficient of Data

- Sensitive to outliers $\rightarrow$ use Spearman correlation coefficient

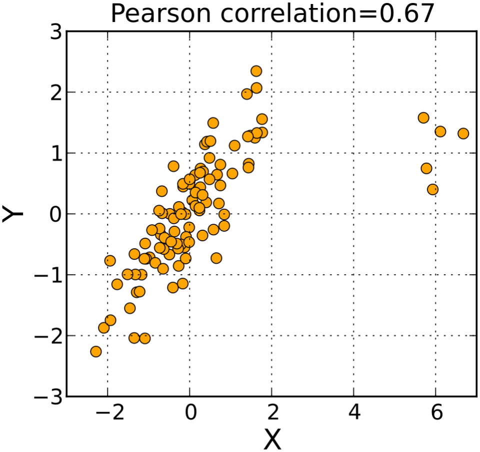


Figure taken from https://en.wikipedia.org/wiki/Spearman_rank_correlation_coefficient

Spearman Correlation Coefficient of Data

- Use ranks of data instead of data values
- Definition:

$$r_{xy} = 1 - \frac{6\sum_i^N (rx_i - ry_i)^2}{N(N^2 - 1)}$$

- rx_i is the rank of data x_i

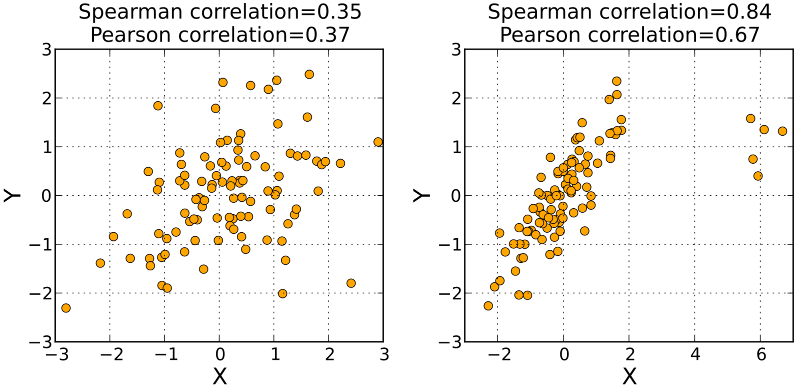


Figure taken from `https://en.wikipedia.org/wiki/Spearman_rank_correlation_coefficient`

Decorrelation - Whitening

- Correlation between variables are sometimes not desired
- By transformation, variables can be decorrelated and covariance matrix can be I
- This process is called whitening

Decorrelation

- Let X_1 and X_2 are correlated variables
- A transform is required

$$Y_1 = aX_1 + bX_2$$
$$Y_2 = cX_1 + dX_2$$

  such that Y_1 and Y_2 are uncorrelated.
- In matrix notation

$$\mathbf{Y} = G\mathbf{X}$$

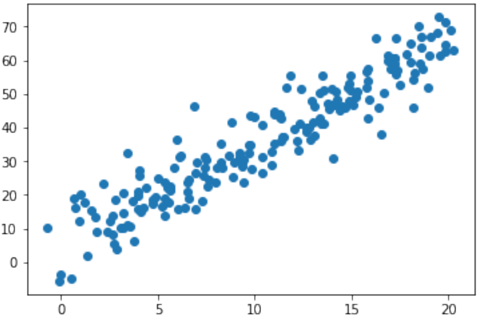
  where

$$G = \begin{bmatrix} \mathrm{a} & \mathrm{b} \\ \mathrm{c} & \mathrm{d} \end{bmatrix}$$

- Covariance matrix of X_1, X_2 is

$$\Sigma_X = \begin{bmatrix} \sigma_{X1}^2 & \rho\sigma_{X1}\sigma_{X2} \\ \rho\sigma_{X1}\sigma_{X2} & \sigma_{X2} \end{bmatrix}$$

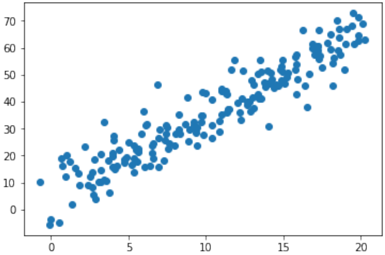
Decorrelation

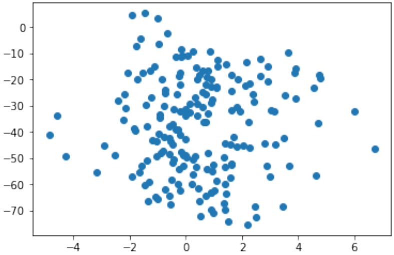


Decorrelation

- If $\mathbf{Y}$ is uncorrelated $\rightarrow \Sigma_Y$ is a diagonal matrix
- Find eigenvalues ($\Lambda = diag\{\lambda_X\}$) and eigenvectors ($V$) of Σ_X such that $\Sigma_X = V\Lambda V^T$
- Let $G = V^T$ ($Y = V^T X$), then Σ_Y will be diagonal and $Cov(Y_1, Y_2) = 0$
- However diagonal elements σ_{Y1} and σ_{Y2} will not be 1

$$\Sigma_Y = \begin{bmatrix} \sigma^2_{Y1} & 0 \\ 0 & \sigma_{Y2} \end{bmatrix}$$





Whitening

- Let $G = \Lambda^{-0.5} V^T$ ($Y = \Lambda^{-0.5} V^T X$), then Σ_Y will be diagonal and diagonal items will be 1.
- $Cov(Y_1, Y_2) = 0$

$$\Sigma_Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$