భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Biostatistics BT2023

Lecture 9 + 10
Correlation and regression

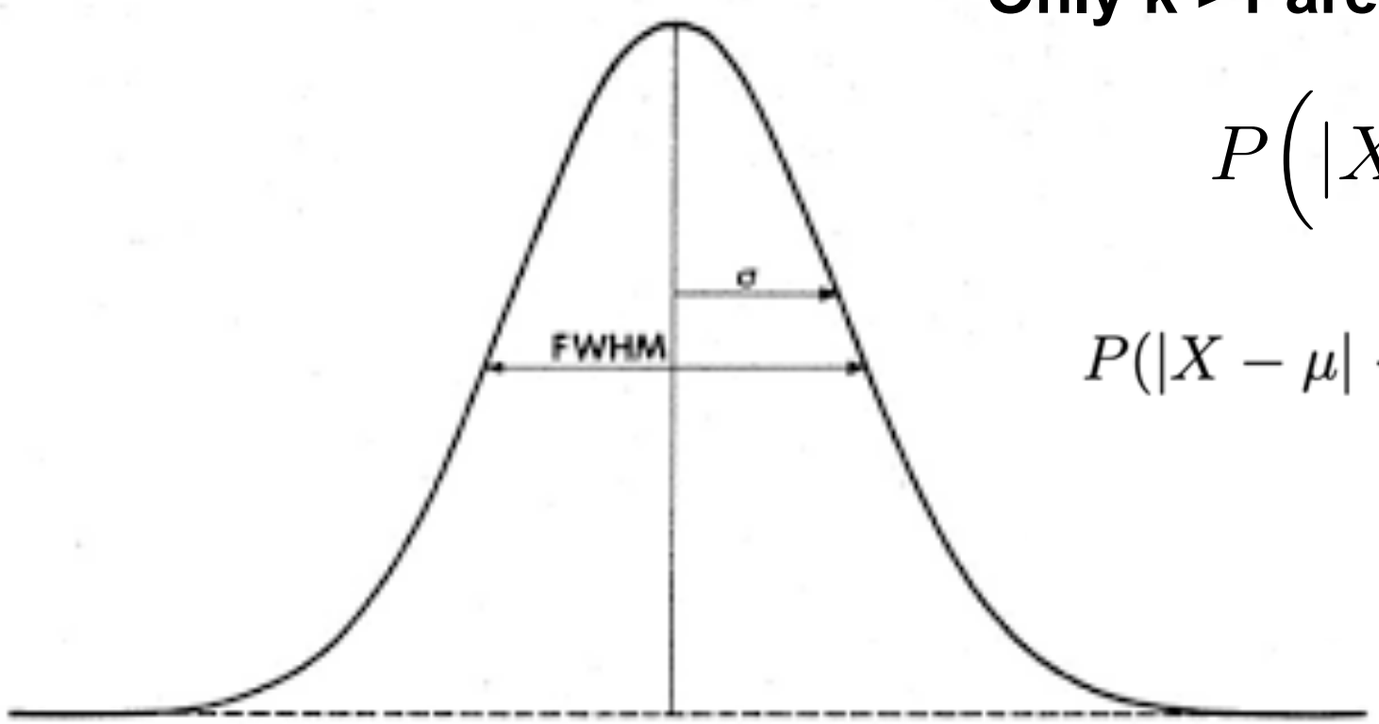
Himanshu Joshi
5 and 8 September 2023

Measure of dispersion

Chebyshev's inequality

The rule is often known as Chebyshev's theorem, tells about the range of standard deviations around the mean, in statistics.In a probability distribution, no more than a certain fraction of values can be more than a certain distance from the mean.

$$P(r)\Big(|X-\mu| \geq k \times \sigma\Big) \leq \frac{1}{k^2}$$



Only k >1 are interesting because for k<1 it trivial

$$P\Big(|X-\mu| < k \times \sigma\Big) \geq 1 - \frac{1}{k^2}$$

$$P(|X-\mu| < k\sigma) = P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

$$P(|X-\mu| \geq r) \leq \frac{Var(X)}{r^2}.$$

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Moments of central measure

Skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

β >3 Leptokurtic
β <3 Platykurtic
β =3 Mesokurtic

$First\ moment\ : \mu_1$ $\quad \frac{\sum_N (X - \bar{X})}{N} \qquad \frac{\sum_N f(X - \bar{X})}{N}$

Always 0

$Second\ moment\ : \mu_2$ $\quad \frac{\sum_N (X - \bar{X})^2}{N} \qquad \frac{\sum_N f(X - \bar{X})^2}{N}$

Measure of variance

$Third\ moment\ : \mu_3$ $\quad \frac{\sum_N (X - \bar{X})^3}{N} \qquad \frac{\sum_N f(X - \bar{X})^3}{N}$

Measure skewness

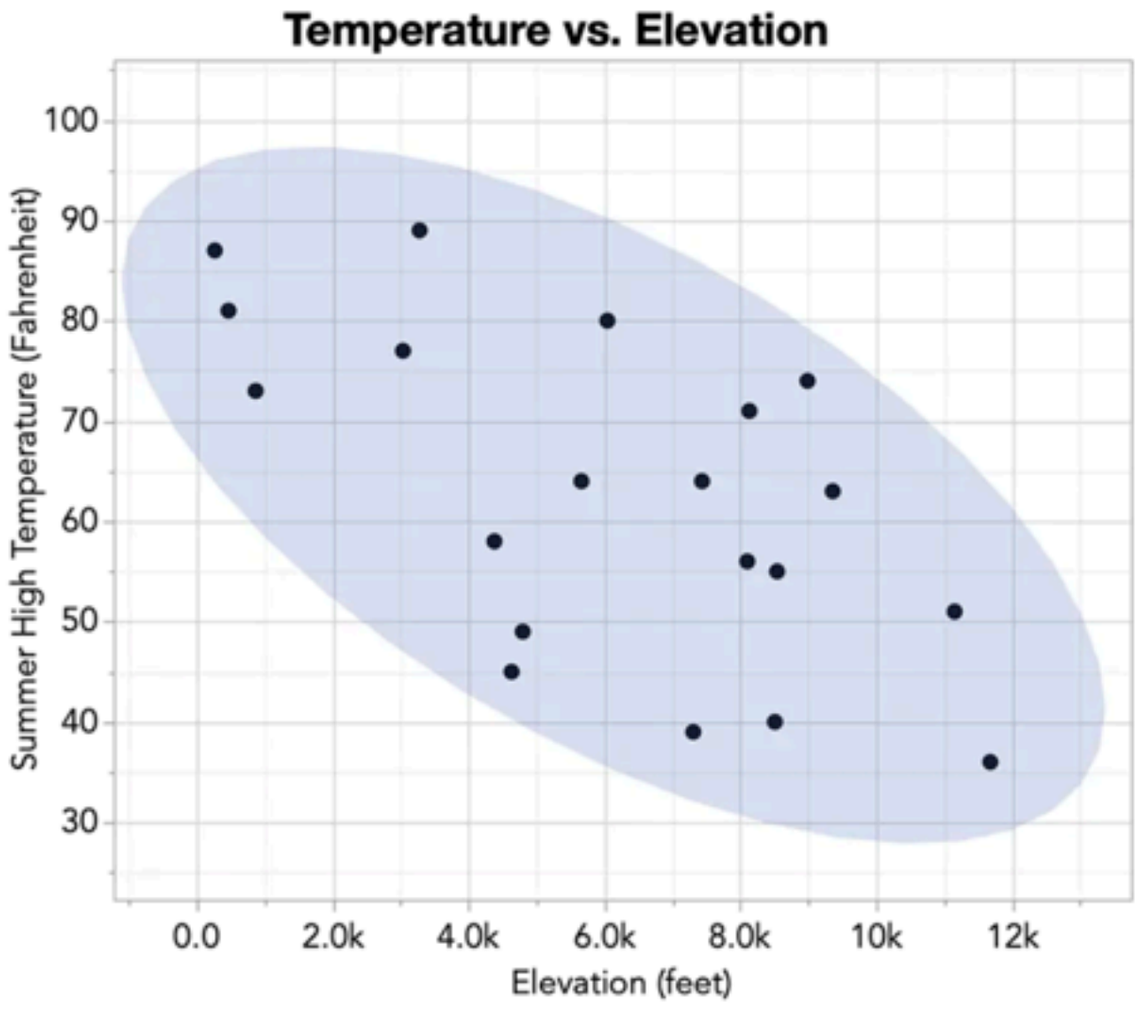
$Forth\ moment\ : \mu_4$ $\quad \frac{\sum_N (X - \bar{X})^4}{N} \qquad \frac{\sum_N f(X - \bar{X})^4}{N}$

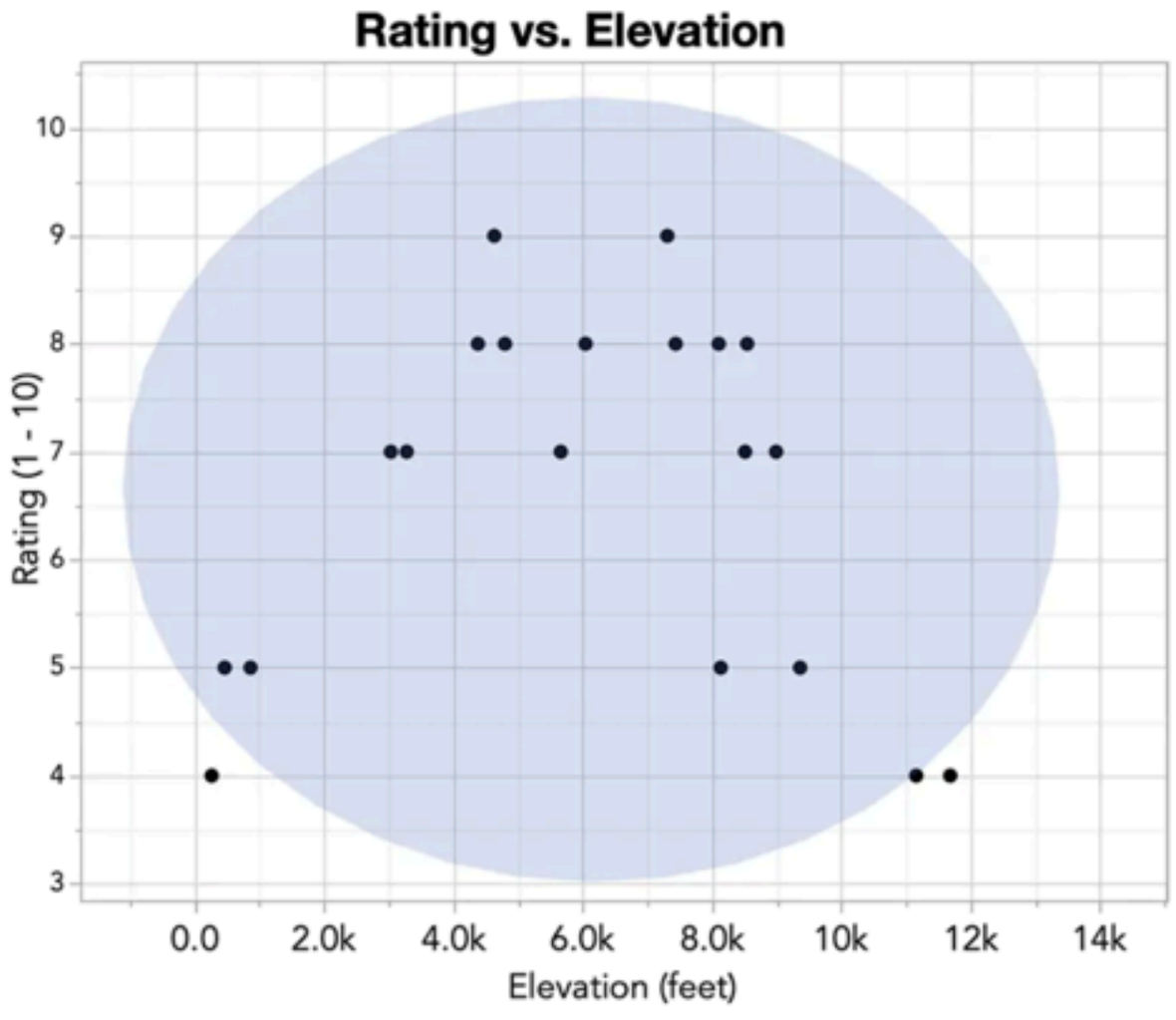
Measure Kurtosis

Causality and Correlation

Causality is an influence which leads to another event or production

We are often interested in if two incidents are related to each other, if so how to quantify them. Lets consider the following scatter plot





Covariance

Covariance in two quantities variable X and Y on a given set is given by

$$Cov.\ (X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n}(x_i)\frac{1}{n}\sum_{i=1}^{n}(y_i)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Correlation coefficient

$$r = \frac{\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}}{n}$$

$$\implies r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$where\ \sigma_{xy} = Cov(X, Y)$

Properties of the correlation coefficient

1. It will range from -1 to +1
2. Measures the closeness of the fit

$$dx = x - \bar{x}$$

$$dy = y - \bar{y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} dxdy - \left(\sum dx \sum dy\right)}{\sqrt{\sum dx^2 - \frac{\left(\sum dx\right)^2}{n}} \times \sqrt{\sum dy^2 - \frac{\left(\sum dy\right)^2}{n}}}$$

Blood sugar

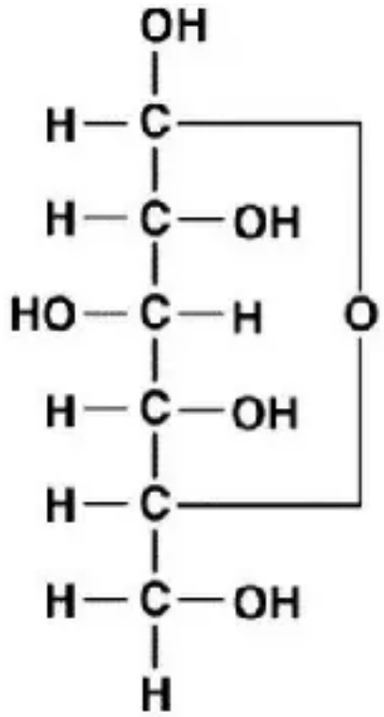
Units

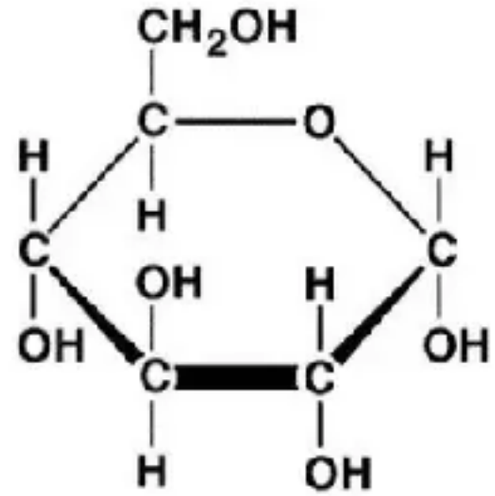
4.4 to 6.1 mm/liter

or 82 to 110 mg/deciliter

Four types of biomolecules

?





Glucose

Regression

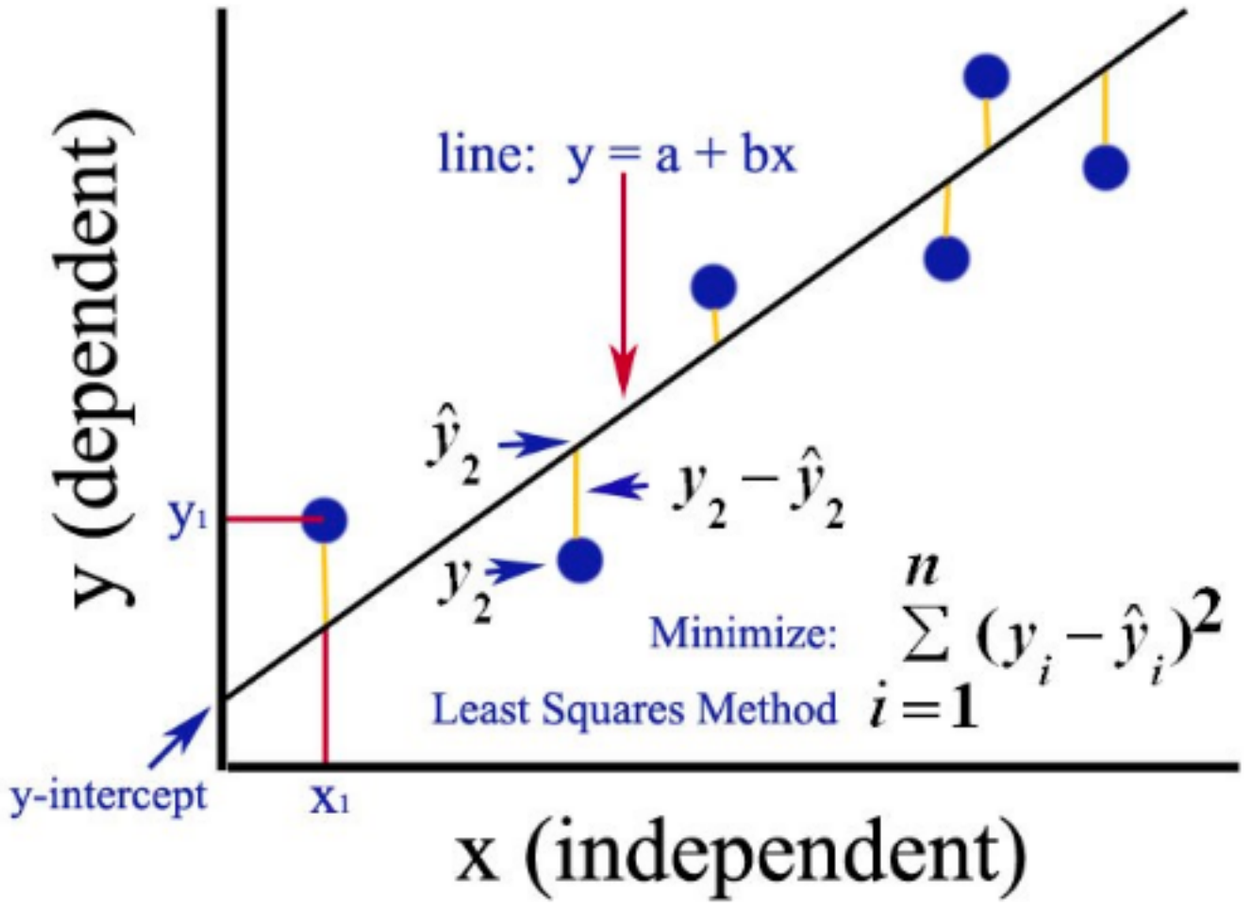
Regression shows the relationship b/w the average values of two variable. Its is very helpful in finding correlations.

1. **Linear regression** can be solved exactly
2. **Non-linear regression** solved using approximation or iteration

Linear regression by least square method

It is a method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum

Linear regression by least square method



$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Covariance Matrix
Principal component analysis

भारतीय साङ्केतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Next Class

2:30 PM Friday, 12 September 2023