Unit 5 – Quantitative Methods: Correlation

Dr M. H. Tayarani

University of Hertfordshire

Objectives

After this lecture you will understand:

1. What correlation means.
2. How to compute correlation.
3. How to assess whether correlation is significant.

Definitions

- ***Correlation***: The extent to which a dependent variable (outcome) varies in proportion to an independent variable (treatment).
- ***Significance***: The probability that an observed effect is due to random variation of sampling.
- ***Effect***: How much the dependent variable increases or decreases with a corresponding increase in the independent variable.

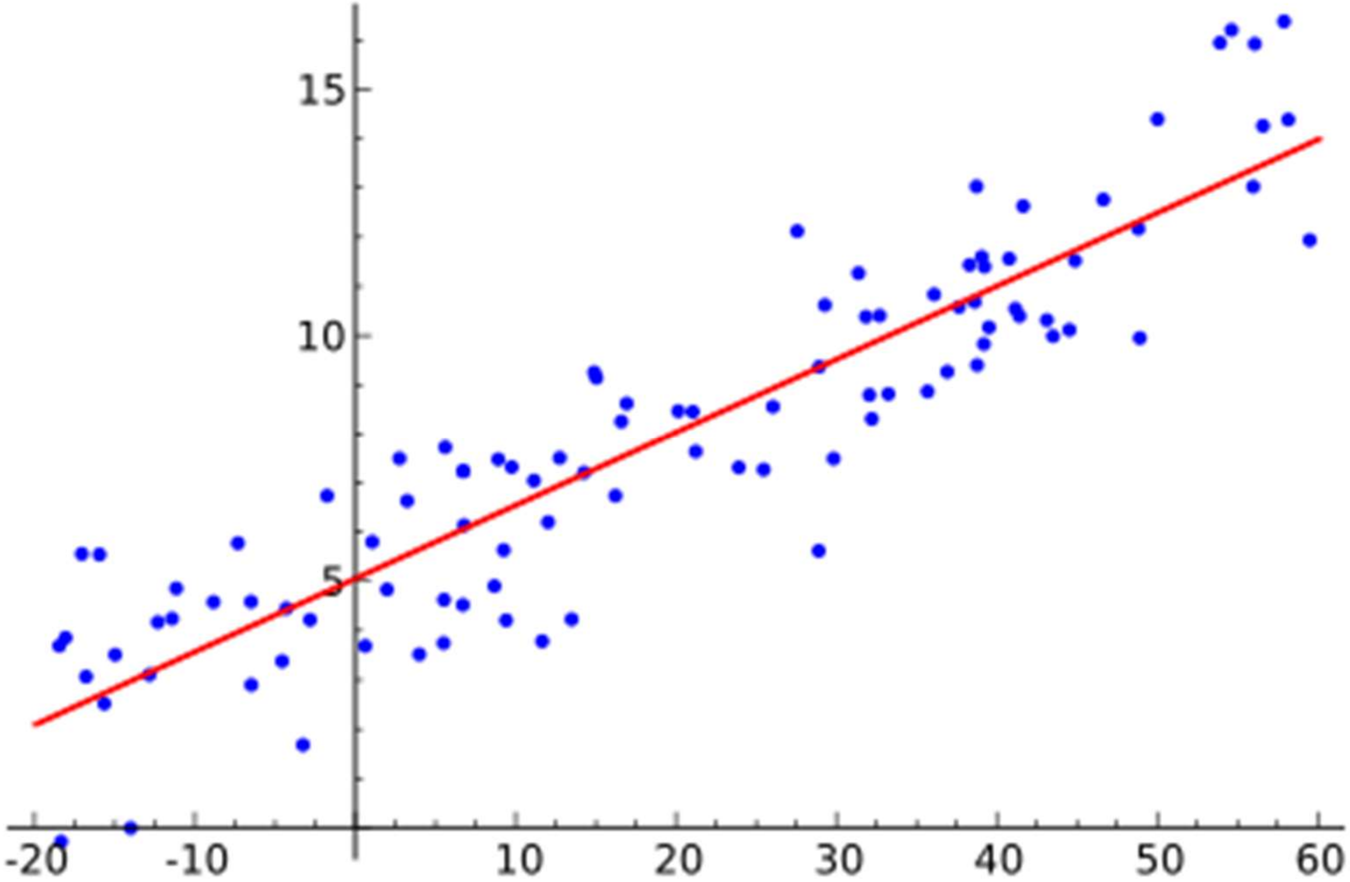
Definitions

- ***Interval data***: Numeric data where the difference between two values is meaningful.
- ***Ordinal data***: Numeric data than can be ordered (sorted), but the difference between values is not necessarily
  meaningful.
  How much more is "definitely" than "mostly?"
- ***Nominal (categorical) data***: Data with non-numeric values (colors, places, locations).

Ordinal Data vs Interval Data

1. Ordinal data are concerned about the order and ranking while interval data are concerned about the differences of value within two consecutive values.
2. Ordinal data place an emphasis on the position on a scale while interval data are on the value differences of two values in a scale.
3. There is no certainty of equality in ordinal data while there is a presence of equality in interval data.
4. The scale and value of differences in an ordinal sequence is not uniform while the two factors in interval data are uniform.
5. Interval data are considered more informative kinds of quantitative data compared to ordinal data.
6. Interval data are a form of parametric data while ordinal data are a form of non-parametric data.
7. Interval data can also be placed in an ordinal manner.

Linear Regression



Linear Regression

When there is one independent variable and one dependent variable, a linear model provides a formula for the response:
$y = \alpha + \beta x$
Linear regression attempts to derive such a formula by minimizing the error between the predicted and actual values in a sample.

Linear Regression

Given:

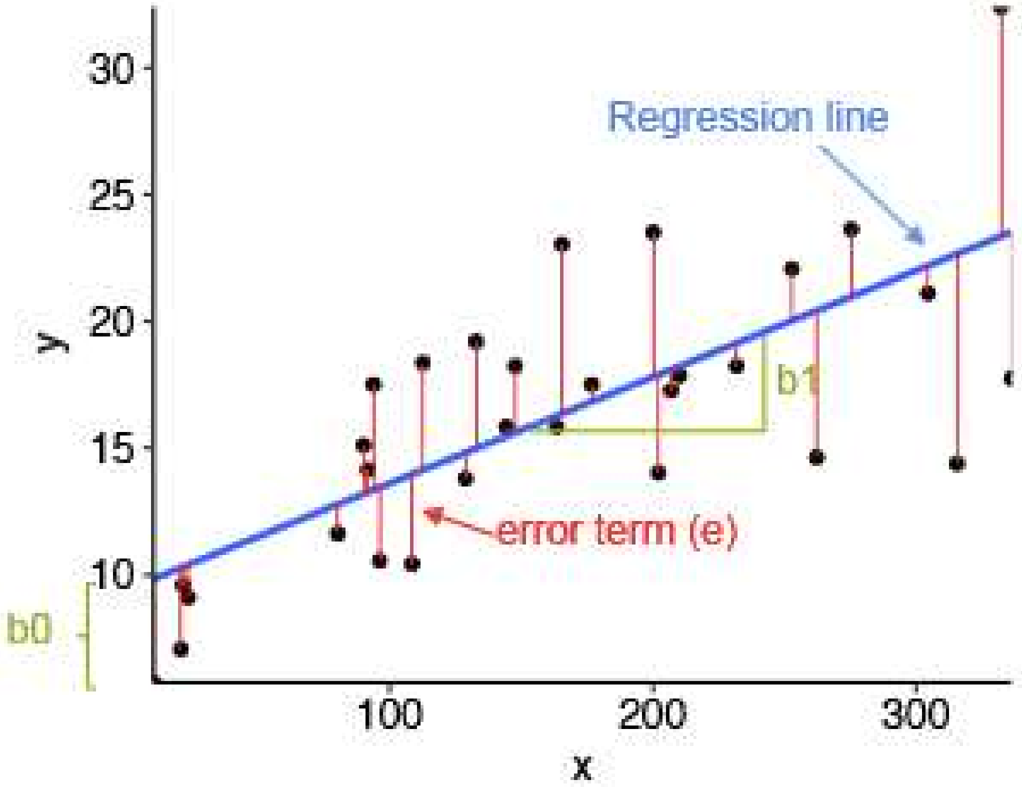
$y_i = \alpha + \beta x_i + \varepsilon_i$

where α,β are the values of the linear model for the underlying population, and ε_i is the difference between the observed value y_i in the sample, and the predicted value based on x_i, α and β,

then one approach is to minimize the sum of squares of ε_i, by finding values for α,β such that the following is minimized:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha_i - \beta x_i)^2$$

Linear Regression



Linear Regression

There are three measures that should be considered

1- Is the correlation significant: p-value determines the significance. A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance.

2- What is the Correlation coefficient: The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y.

3- Is the correlation substantial (Effect)? With a certain amount of change in the x axis, how much change is observed in the y axis.

What is Significance?

- Every sample is a more or less random subset of the underlying population.
- There is a finite (usually small, but always non-zero) probability that the sample points are not distributed in the same way as the underlying population.
- Statistical significance measures the probability that this is the case.
- *Significant ≠ substantial!*

Testing Significance

We pose two hypotheses:

- The null hypothesis (H_0) that the observed effect is due to random variations and is not significant.
- The alternative hypothesis (H_1) that the observed effect is not due to random variations.
- We reject the null hypothesis if the probability that the observed effect, as measured by a test statistic, is due to random variation is less than some threshold p (referred to as the $p - value$).
- In software engineering we use $p - value$ = .05.

P-value (Testing Significance)

- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to or in addition to preselected confidence levels for hypothesis testing.

Testing Significance

1. Statistics never prove anything; they just support hypotheses (or not).

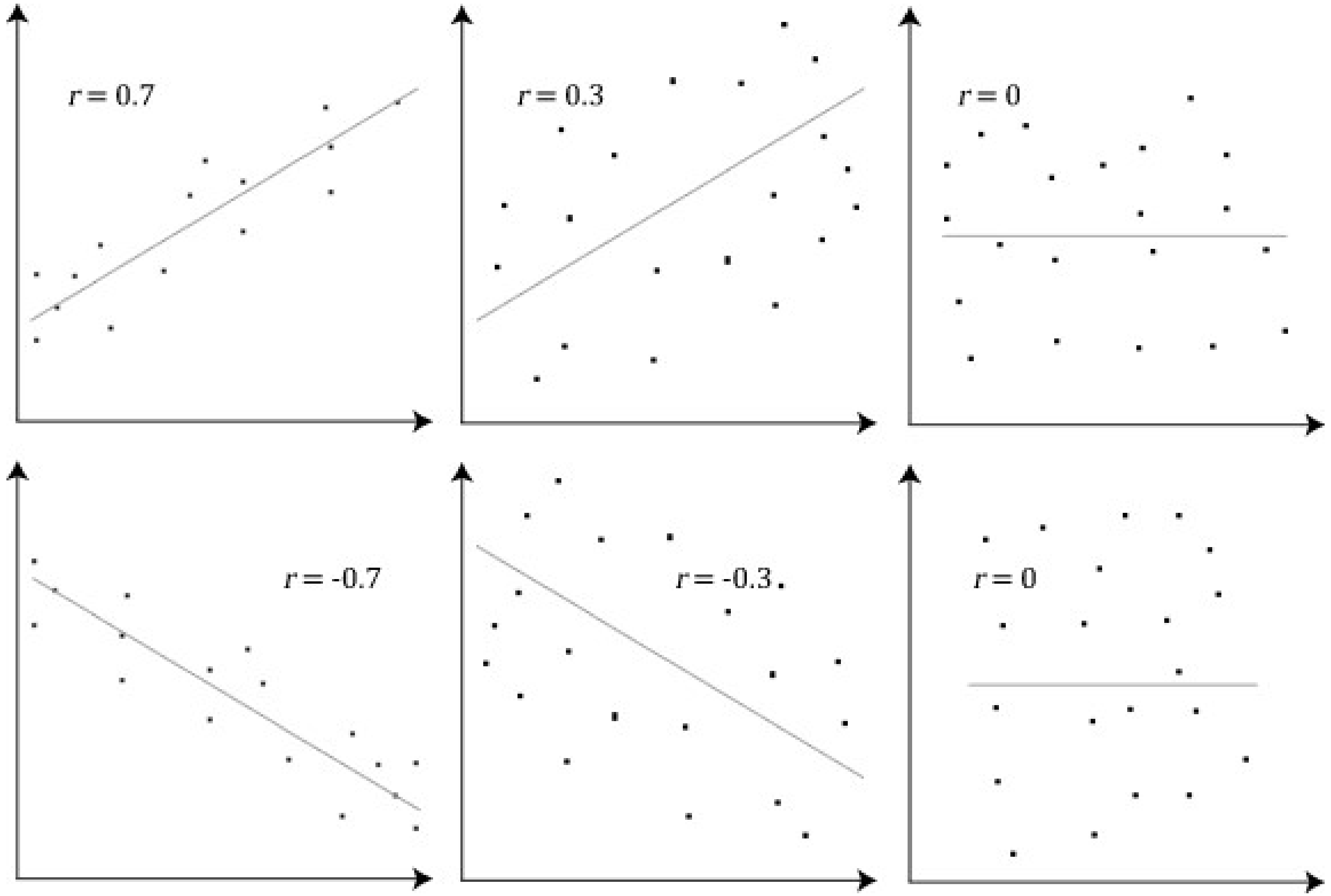
2. Rejecting the null hypothesis simply means the probability is pretty high that the observed correlation is due to a relationship between the dependent and independent variables.

3. Correlation does not mean causation!

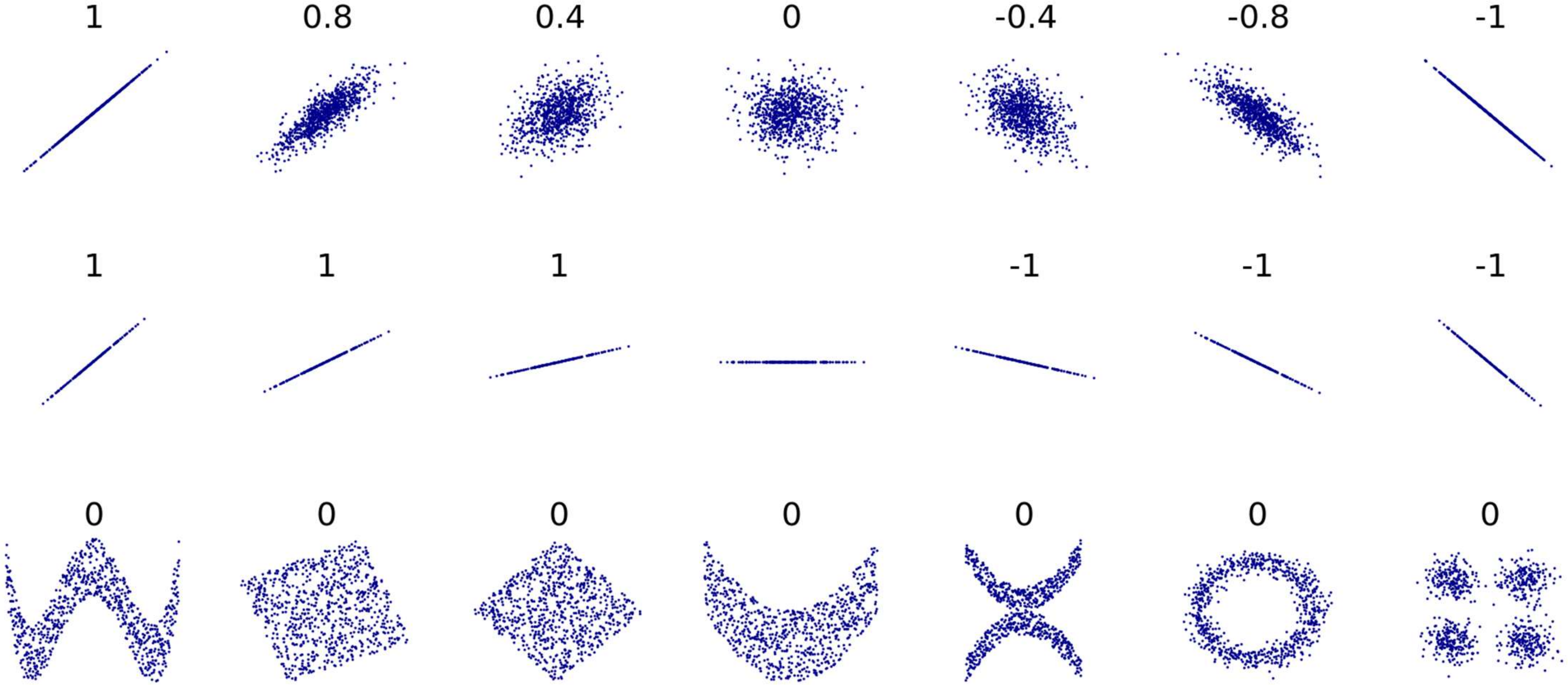
Correlation Coefficient

- Correlation coefficients are used to measure the strength of the relationship between two variables.
- Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.
- Values always range between -1 (strong negative relationship) and +1 (strong positive relationship). Values at or close to zero imply a weak or no linear relationship.
- Correlation coefficient values less than +0.8 or greater than -0.8 are not considered strong. Although it also depends on the subject of study. A 0.9 correlation may be weak in physics, but a 0.7 is considered strong in social sciences.

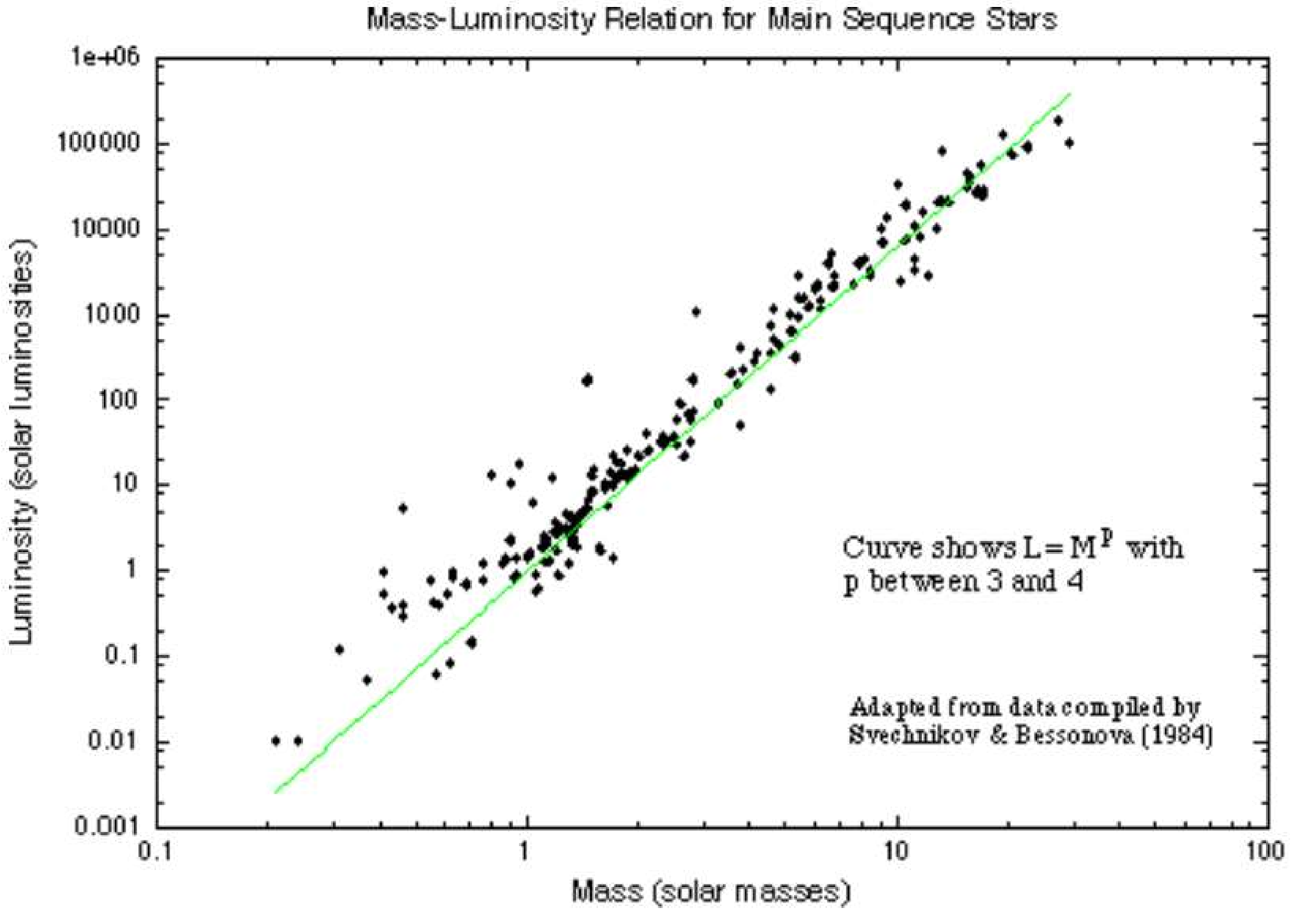
Scatterplot



Scatterplot



Star Mass vs Luminosity



Tests of Correlation

Three statistics are used to assess correlation:

1. Pearson's r, for normally distributed data.
2. Kendall's τ, for non-parametric (not normally distributed) data.
3. Spearman's ρ, also for non-parametric data.

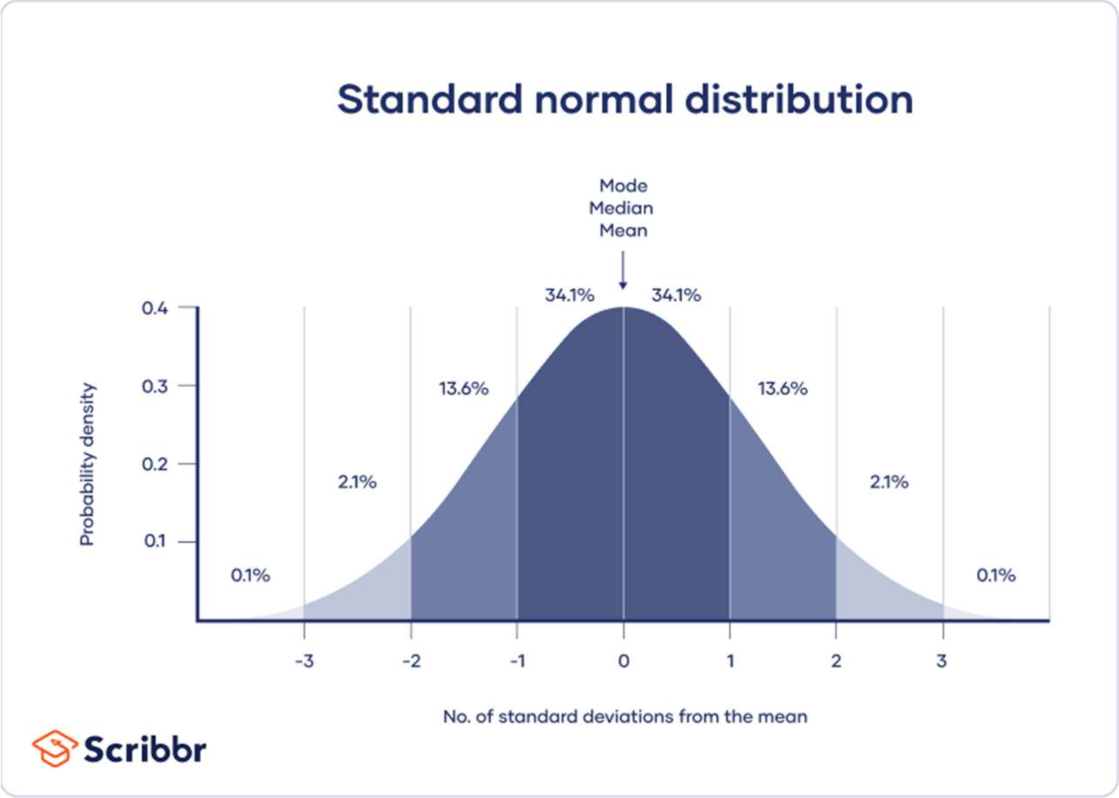
Normal Distribution

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

Height, IQ, Income, Shoe Size, Birth Weight, blood Pressure, etc.

μ= average

σ=standard deviation



Exponential Distribution

$$f(x) = \frac{1}{\lambda} e^{-(x-\mu)/\lambda} \qquad x \geq \mu; \lambda > 0$$

Examples:

The amount of time until an earthquake occurs

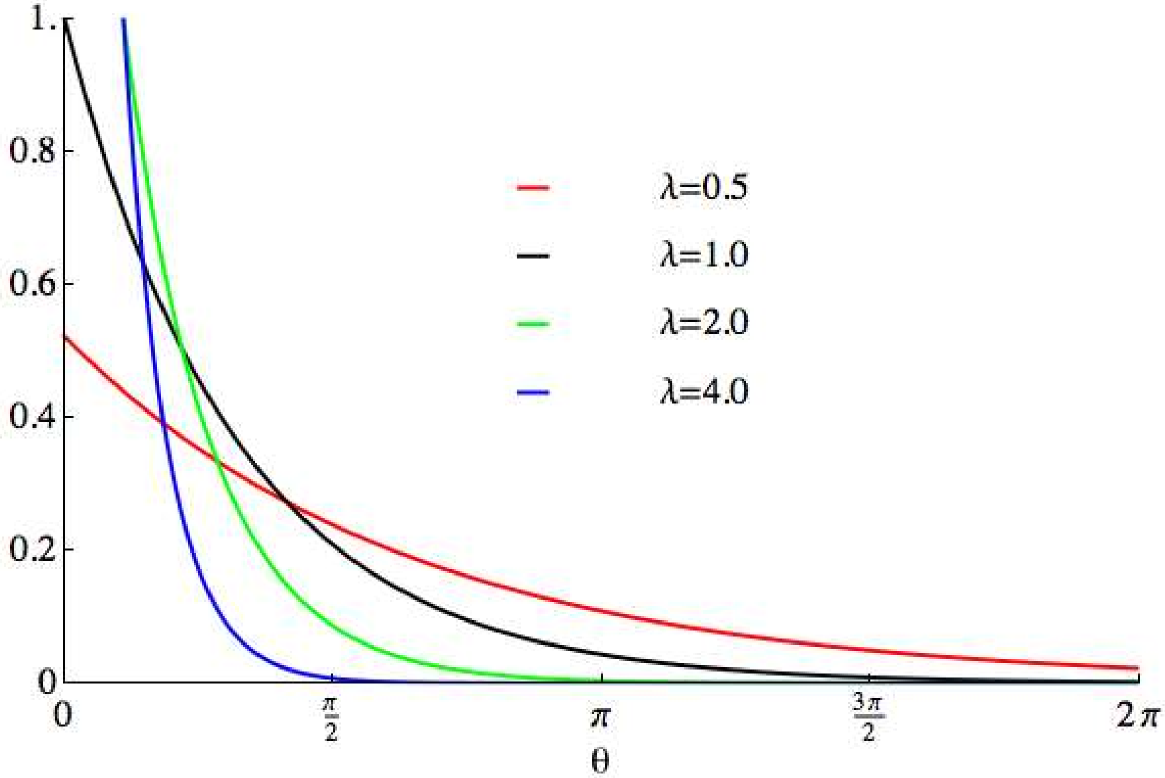
Length, in minutes, of long distance business telephone calls

The amount of time, in months, a car battery lasts

The value of the change that you have in your pocket

The amount of money customers spend in one trip to the supermarket

λ=1/average



Gamma Distribution

$$f(x) = \frac{(\frac{x-\mu}{\beta})^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta\Gamma(\gamma)} \qquad x \geq \mu; \gamma, \beta > 0$$

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

Examples:

The amount of rainfall accumulated in a reservoir

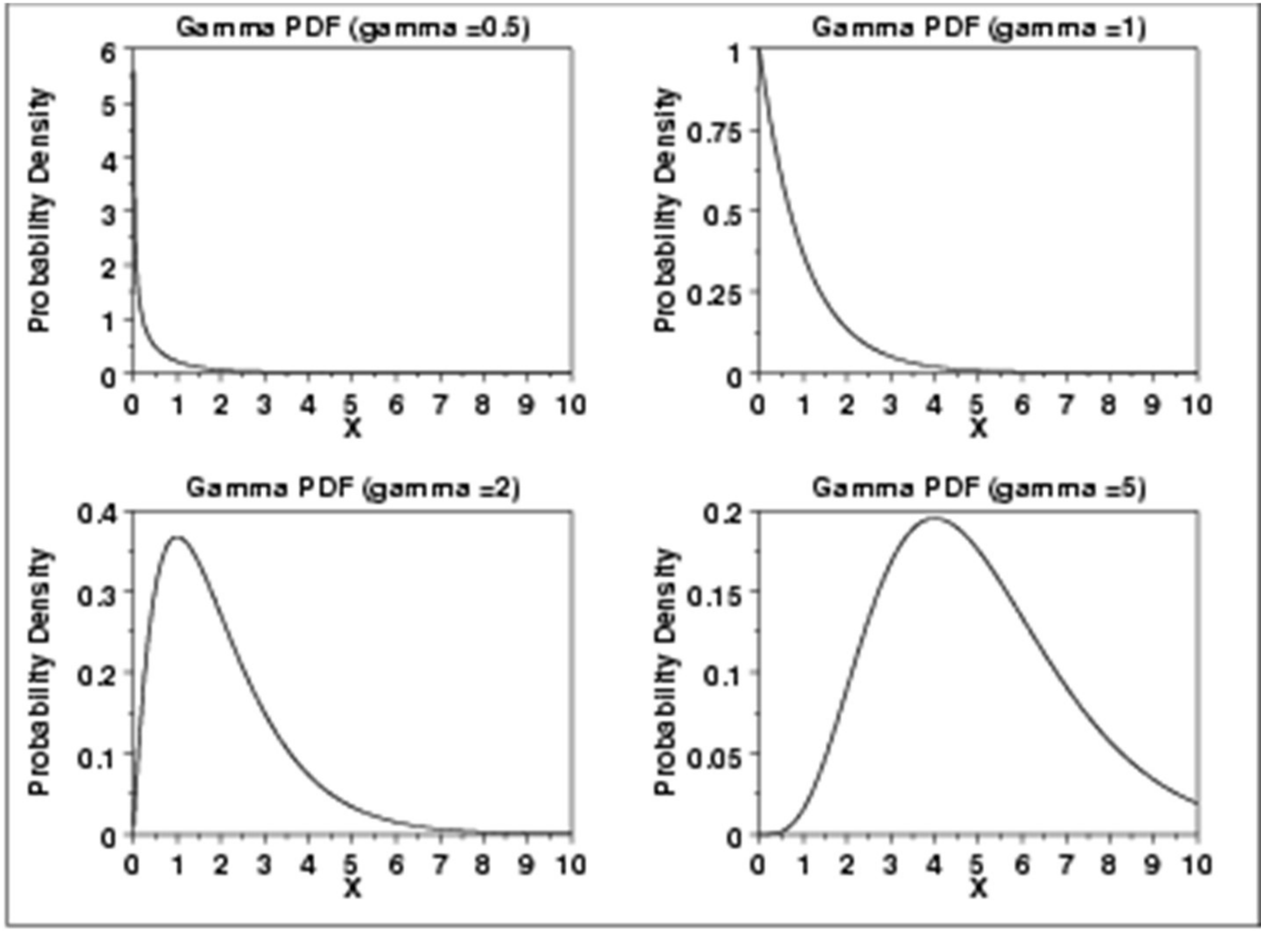
The size of loan defaults or aggregate insurance claims

The flow of items through manufacturing and distribution processes

The load on web servers

The many and varied forms of telecom exchange

γ, shape parameter=average



Pearson's r

- Suitable for normally distributed interval data:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- r = 1 is perfect positive correlation (all points fall on a line with positive slope).
- r = −1 is perfect negative correlation (all points fall on a line with negative slope).
- r = 0 indicates no correlation.

Pearson's r

- Pearson's correlation remains a consistent estimator of the population correlation even when bivariate normality is not present.
- When the variables are not bivariate normal inferential tests that assumes a normal sampling distribution (e.g., via a Fisher transformation, or a t-distribution) may not be trustworthy. One of these methods is usually used to determine statistical significance

Non-parametric statistics: Spearman's ρ

Spearman's rank correlation coefficient is also appropriate for non-parametric data:

Let $rank_{vi}$ be the "rank" of value v_i when all the values are ordered and assigned an integer i corresponding to their position in the ordering. Then,

$r = \rho(rank_X, rank_Y)$

where $rank_V$ is the ranking of the independent variables, and $rank_Y$ is the ranking of the dependent variables. In other words, we just compute Spearman's ρ on the rankings rather than the actual values.

Non-parametric statistics: Kendall’s τ

Formally called the Kendall rank correlation coefficient, τ also measures correlation of non-parametric data:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

where $\operatorname{sgn}(x)$ is 0, -1, or 1 depending on whether x is zero, less than one, or greater than one.

Non-parametric (rank) statistics: Intuition

- Let (x_i, y_i) and (x_j , y_j) be two data points in our sample. If y is
- positively correlated with x, we would expect $rank(y_j) > rank(y_i)$
- if $rank(x_j) > rank(x_i)$.
- In other words, higher values of x should result in higher values of y.

Summary

1. Use Pearson's r for normally distributed data.

2. Use Spearman's ρ or Kendall's τ for any data, but especially non-parametric data.

3. Correlation does not mean significance.

4. Significant does not mean substantial.

5. Significant correlation does not mean causation.

6. Null hypothesis (H_0) is either rejected, or not. Nothing, including H_1, is ever proven with statistics.

7. Know your sample!