Assumptions

Required to generate conclusions about statistical populations.

Must be made with care: inappropriate assumptions can generate wildly inaccurate conclusions.

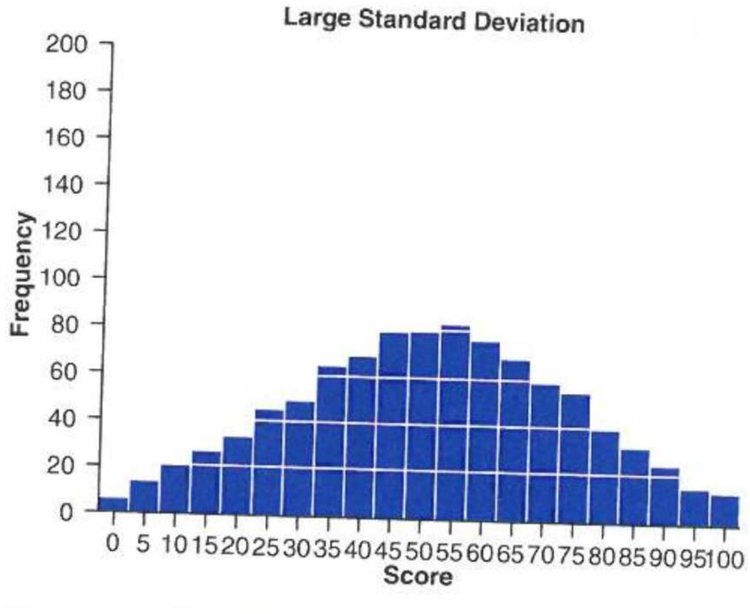
Why to bother with assumptions?

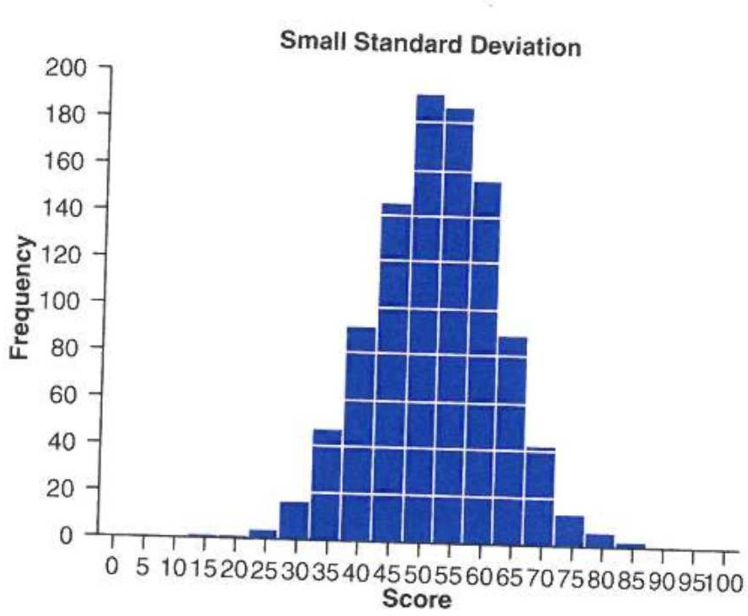
- Why don’t you just perform a non parametric test rather than a parametric test?
- Parametric data have more assumptions, therefore they are more powerful
- Powerful: tests are better at picking up differences in variables in the population.
- For the same number of observations, they are more likely to lead to the rejection of a false hull hypothesis.

Normality

Used to determine whether a data set is well-modeled by a normal distribution or not.

You can use a statistical test and or statistical plots to check the sample distribution is normal.





Two most often used normality tests:

Kolmogorov-Smirnov test

- It has poor power to detect non-normality compared to the Shapiro-Wilk.
- Some statisticians say it is now really only of historical interest.

Shapiro-Wilk test

- A regression-type test that uses the correlation of sample order statistics (the sample values arranged in ascending order) with those of a normal distribution.
- Most powerful normality test available and is able to detect small departures from normality.
- Limitation: is it's not suitable for very large sample sizes. Samples up to 5,000 observations, but some software limits use to 2,000, or as few as 50, observations.

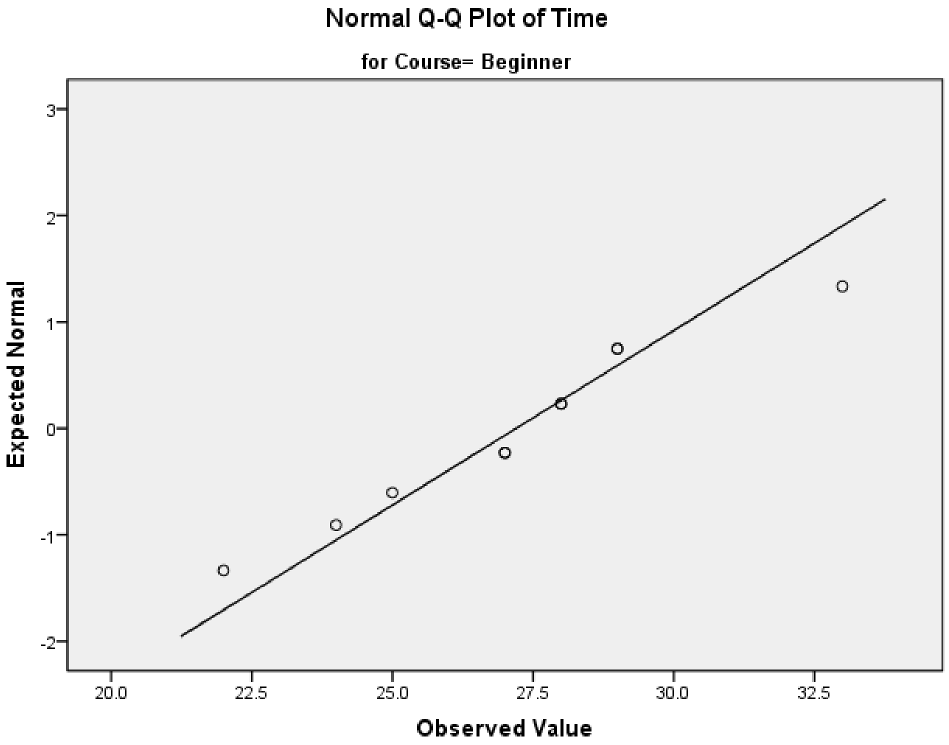
Normality test is a hypothesis test!!!
Test a null against alternative hypothesis

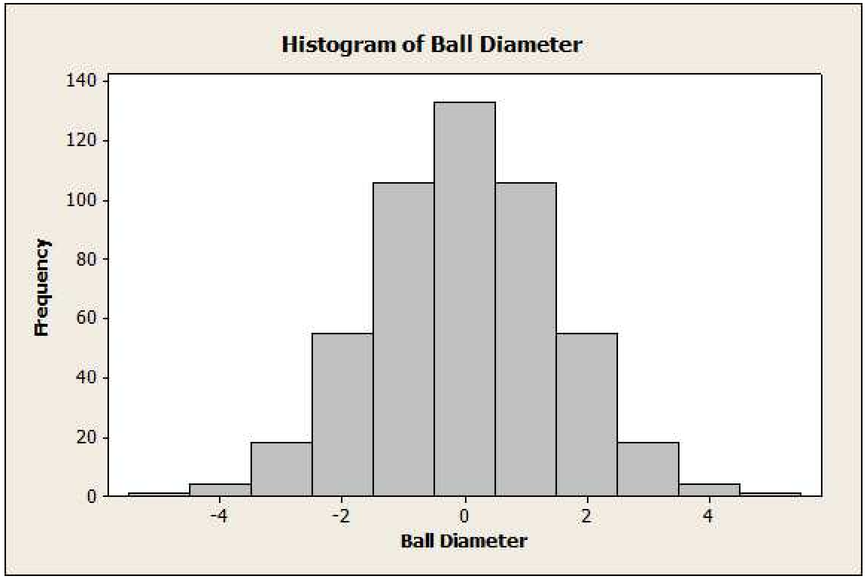
Ho: sample data are normally distributed.
H_1: sample data are not normally distributed.

***p*-value tells you if you should you reject the null hypothesis or you fail to reject the null hypothesis.**

- **When it's significant ($p < 0.05$) you should reject the null hypothesis and conclude the sample is not normally distributed.**
- **When it is not significant (> 0.05), do not reject the null hypothesis and you can only assume the sample is normally distributed.**

***always double-check the distribution is normal using the Normal Q-Q plot and Frequency histogram.**





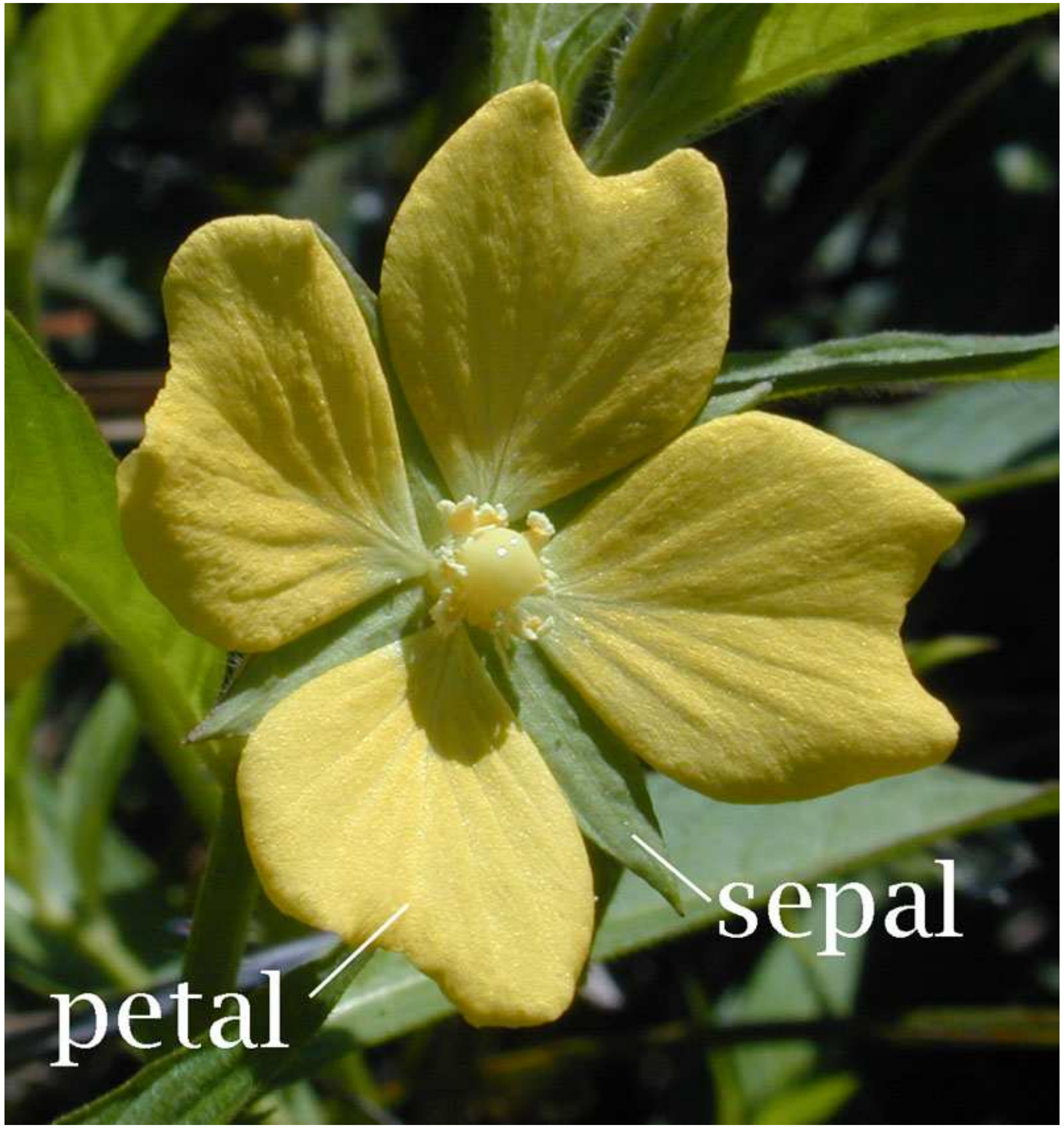
Ex.: Beauty factor under alcohol effect.

- Your turn!

The famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

→ Test if the 4 variables are normally distributed.

** remember to set your null and alternative hypothesis*

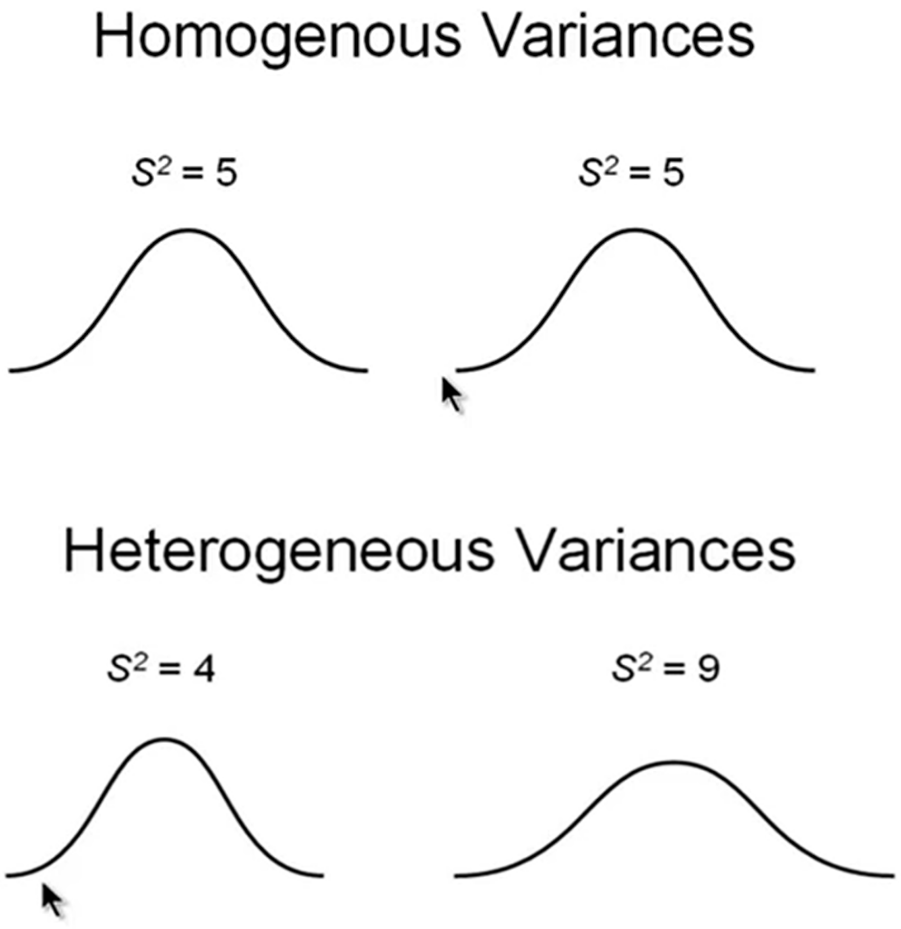

sepal
petal

What if distribution is not normal?

- Transform your data:
- http://pareonline.net/getvn.asp?v=8&n=6
- How to do in SPSS → https://statistics.laerd.com/spss-tutorials/transforming-data-in-spss-statistics.php
- OR collect more data.
- OR… use non-parametric tests.

Variance

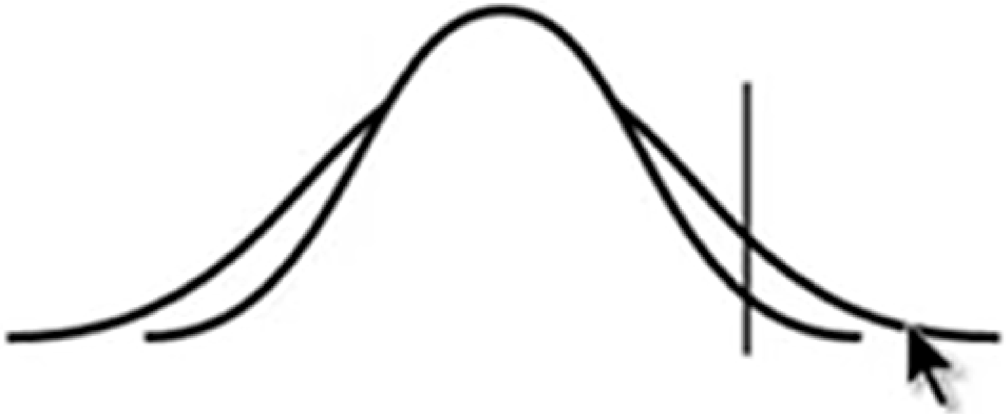
- **Describes how far the numbers lie from the mean.**



Heterogeneous Variances

$S^2 = 169$

$S^2 = 289$

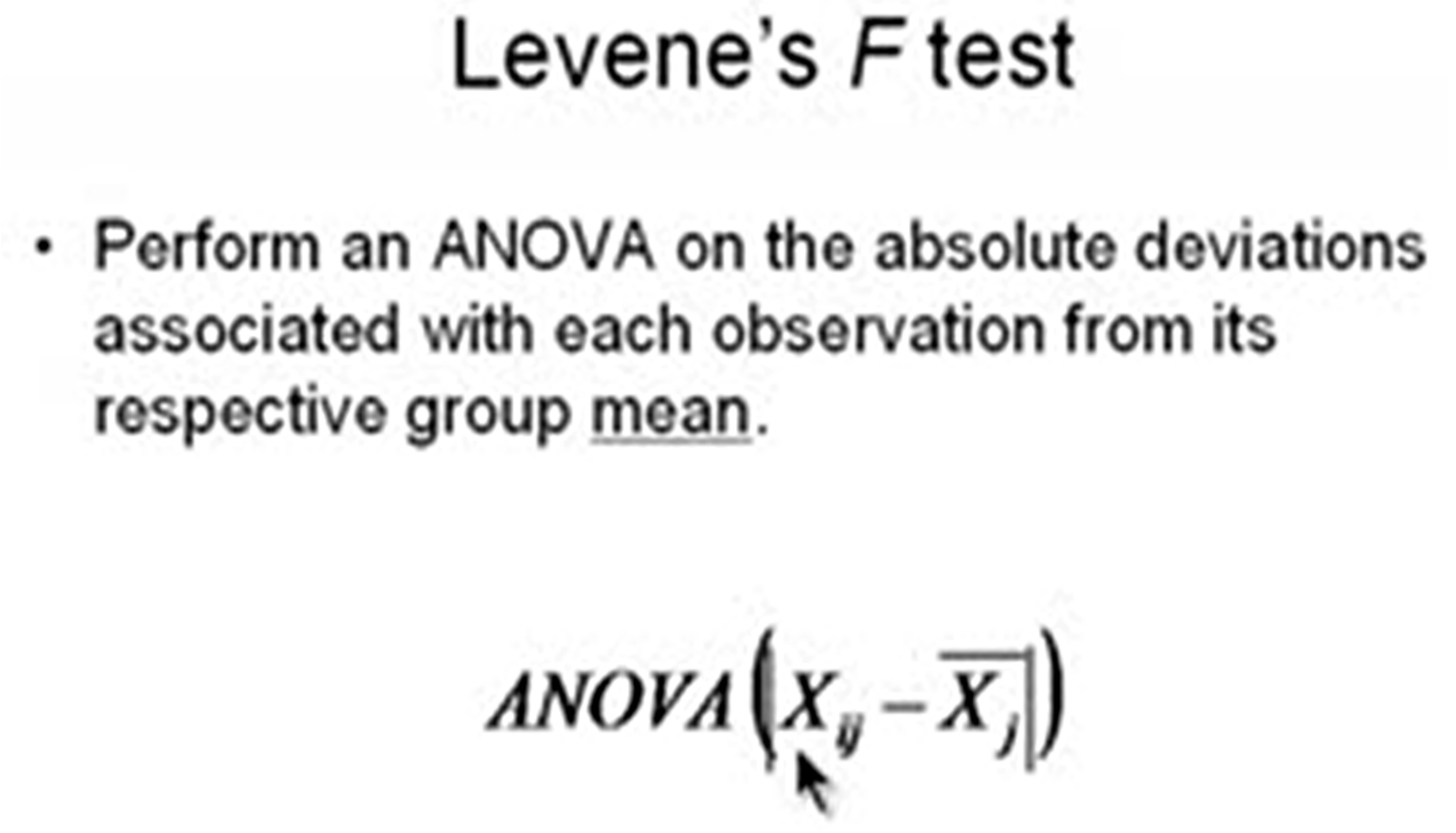


Mean = 100

IQ >130 = 2.5%

Mean = 100

IQ >130 = 7.5%



Levene's *F* test

- Perform an ANOVA on the absolute deviations associated with each observation from its respective group mean.

$$ANOVA\left(\left|X_{ij}-\overline{X}_j\right|\right)$$

*** Chicken wt example**

Test a null against alternative hypothesis

- H_0: $s_1^2 = s_2^2$
- H_1: $s_1^2 \neq s_2^2$

- $p > 0.05$ variances are homogeneous
- $p < 0.05$ variances are heterogeneous

- T test and ANOVA can handle differences in variances up to about 4 times between smallest and largest (Howell, 2007)

Exercise

- The dataset UCBA aggregates data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

→ Test if the variance of the variable “frequency” is homogeneous between rejected and approved students.

What if variances are heterogeneous?

- Assess why and if it is the case, transform your data:http://www3.nd.edu/~rwilliam/stats2/l25.pdf
- Collect more data.
- Non-parametric test.

Independence of observations

- **Two groups consist of different individuals, not the same individuals measured twice or specially matched individuals (such as siblings).**

Most used Parametric tests

→T Test

- used to compare two means or proportions.

Assumptions:

- Independence.
- Equal Variances
- Normality

→The analysis of variance (ANOVA)

- ANOVA can be used both for comparing several means and in more complex situations.

Assumptions:

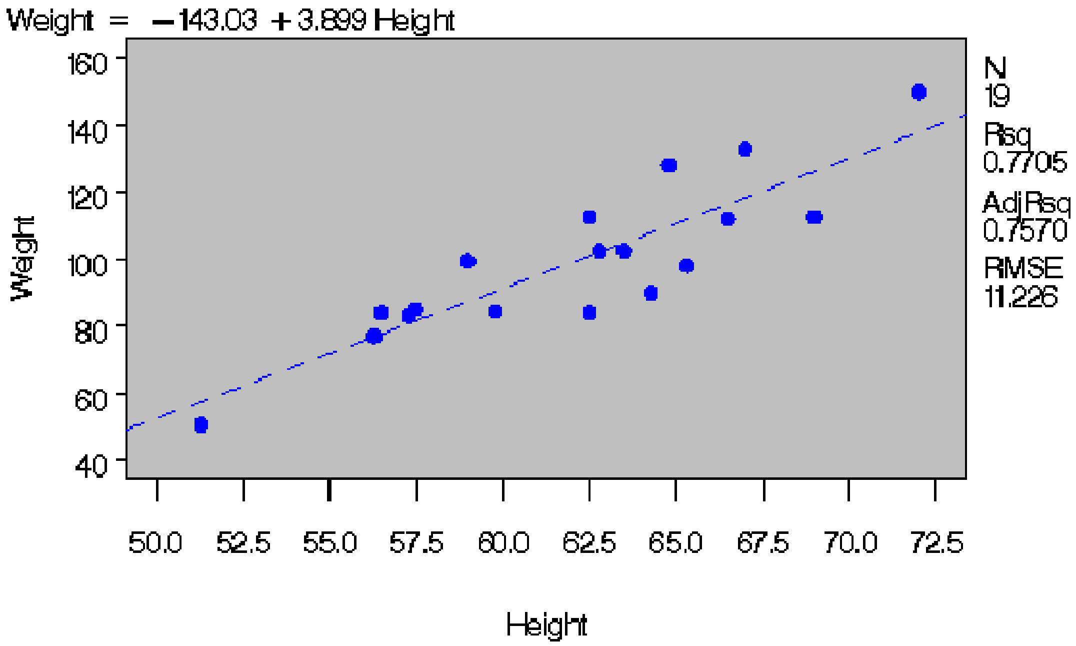
- Independence.
- Equal Variances
- Normality

→Regression

Estimates the relationships among variables.

Assumptions

- **Linearity**
- **Independence**
- **Equal variance**
- **Normality**



* **Independent X Dependent variables**

→Correlation

- Correlation quantifies the extent to which two quantitative variables, X and Y, “go together.” When high values of X.
- are associated with high values of Y, a positive correlation exists.

Assumptions

- Linearity
- Independence
- Equal variance
- Normality