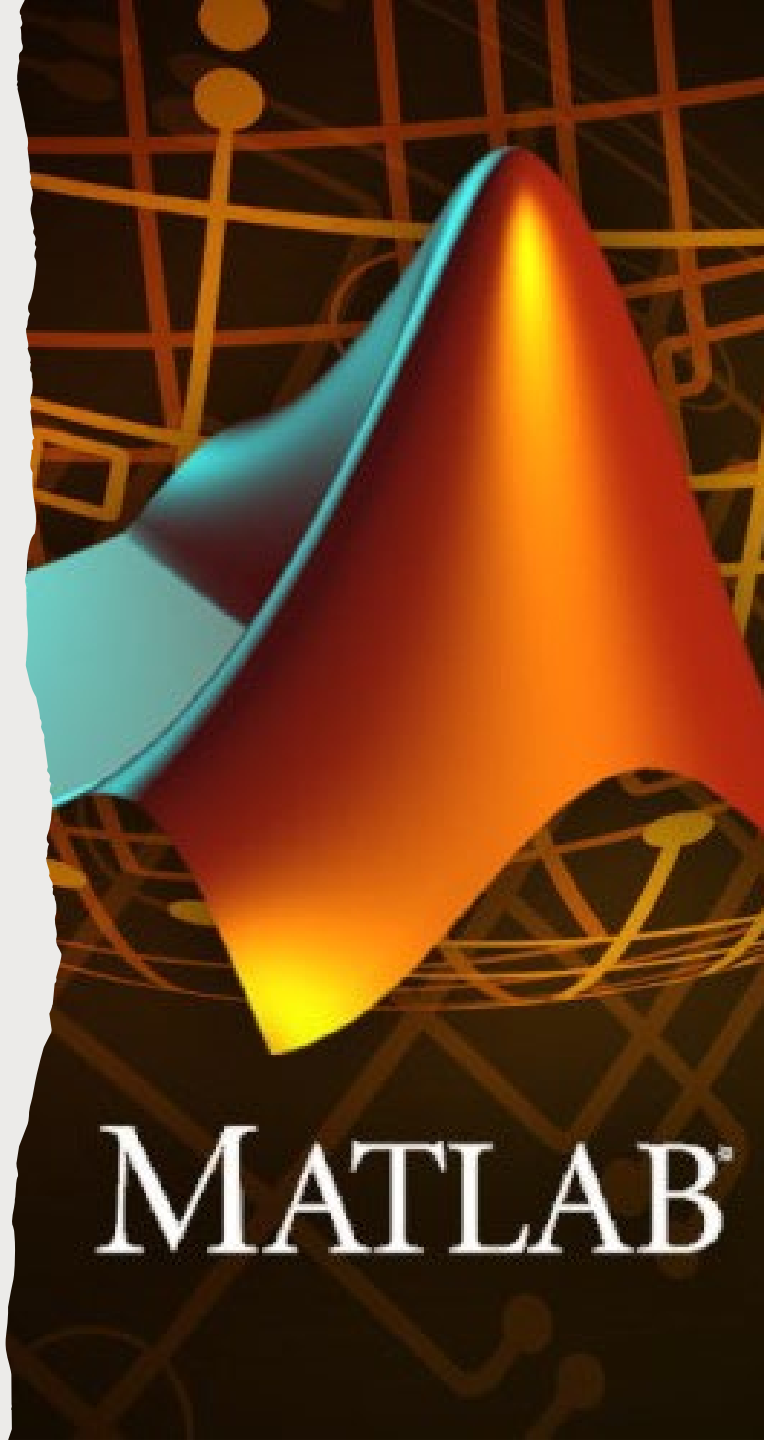
Data analysis for Geoscience

GEO8026

Dr. Matt Perks



Univariate analysis

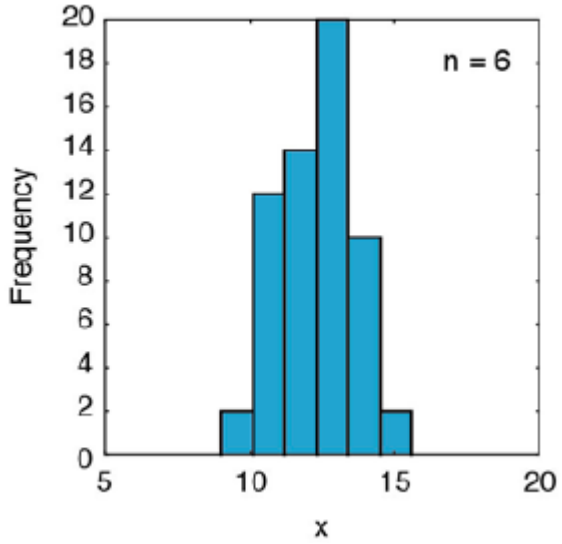
Often it is useful to understand the distribution of the continuous or discrete data that we have.

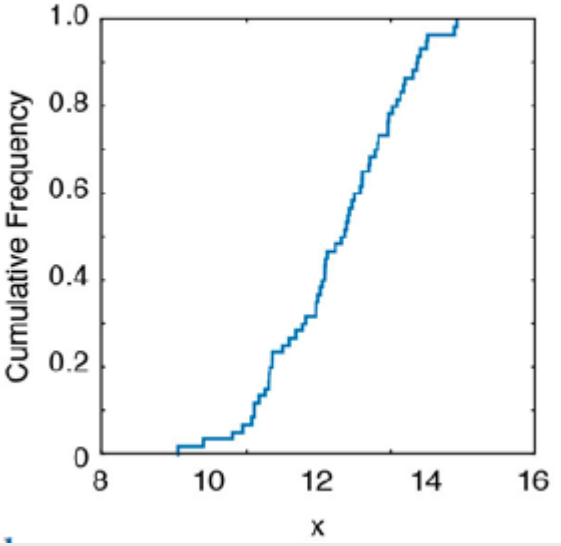
This can be useful to understand the properties of what we have measured, or it can be used to inform the kind of statistical tests that are most suitable given the distribution.

Typically, we would be interested in the following:

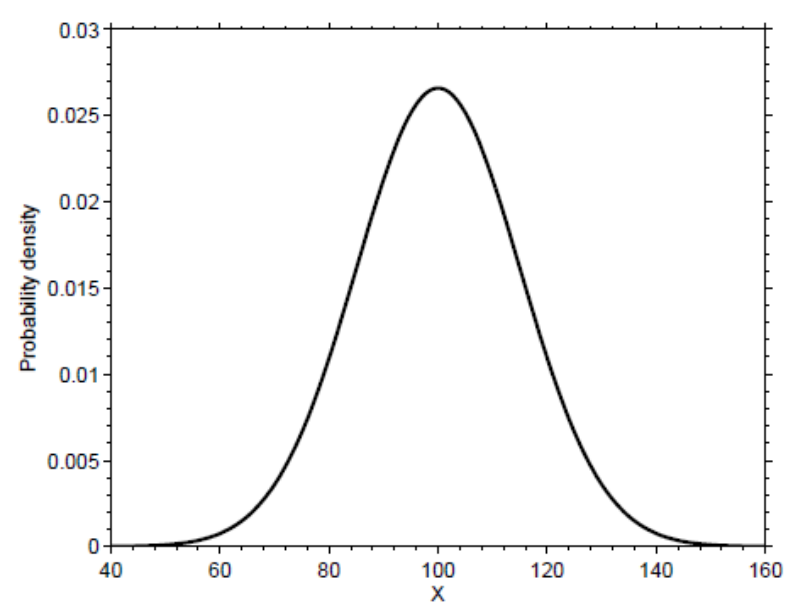
- Sample size: length(x), or numel(x)
- Mean: mean(x) – sum of values divided by sample number
- Mode: mode(x) – most frequent
- Median: median(x) – middle value when ranked
- Percentiles: iqr(x), prctile(x,p)
- Standard deviation: std(x)
- Variance: var(x)
- Skewness: skewness(x) – is the x-axis distribution skewed?
- Kurtosis: kurtosis(x) - how outlier-prone a distribution is

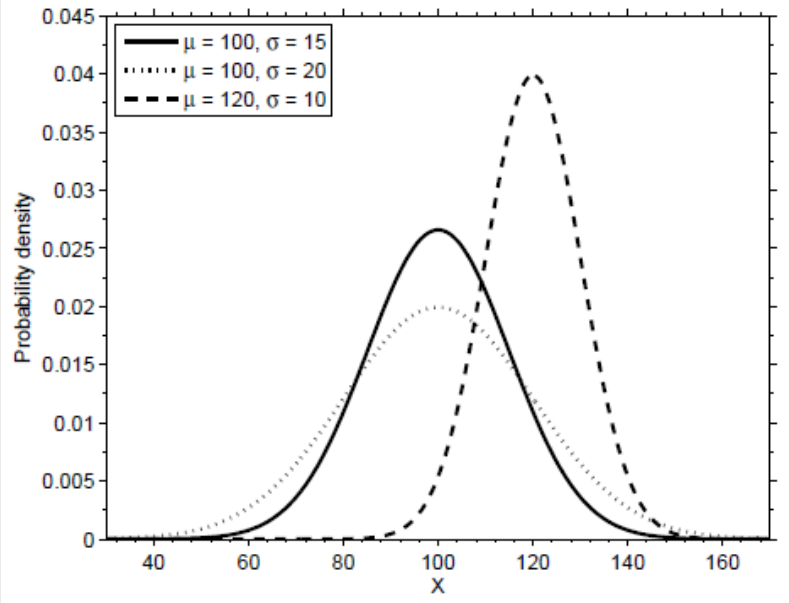
$$\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$$





Univariate analysis





Many statistical tests operate under the assumption that the data are normally distributed.

This is because they are seeking to test the mean of the distribution, which is influenced by outliers.

We can test whether the data distribution meets the assumptions of parametric tests using the:

- Anderson-Darling test: adtest(x)
- Lilliefors test: lillietest(x)

Or alternatively the: One-sample Kolmogorov-Smirnov test to see if data is from a standard normal distribution (mean 0, std 1).

Output of 0 → fails to reject the hypothesis that the data are normally distributed i.e., normal at the 95% ci

Output of 1 → data are not consistent with the null hypothesis i.e. not normal

Hypothesis testing

Parametric tests e.g. t-tests, ANOVA

Operate under the assumption that the data are normally distributed.

Focus on assessing the distribution around the mean

Influenced by outliers and extreme values

Used for continuous datasets

More powerful when mean accurately represents the center of your distribution and your sample size is large enough

Non-parametric tests e.g. Mann-Whitney test

Do not work on the assumption of normality

Focus on assessing the distribution around the median

Assumption that the spread between groups is similar (as this is not assessed in the test)

Used for continuous or discrete data

Hypothesis testing

General Procedure:

1. Formulate the null (H_0) and alternative (H_1) hypotheses.
2. Choose the significance level at which the test will be performed.
3. Calculate the test statistic
4. Compare the test statistic to a critical value or values
5. Reject the H_0 if the p-value is less than the level of significance (α)

Guidance:

An example null hypothesis could be that both samples come from a normal distribution

95% is a common choice for testing of significance. Here we can say that the null hypothesis can be rejected at the 0.05 level of significance.

If $p \leq \alpha$ we can reject the null hypothesis at the defined level of significance
If $p > \alpha$ we can not reject the null hypothesis at the defined level of significance

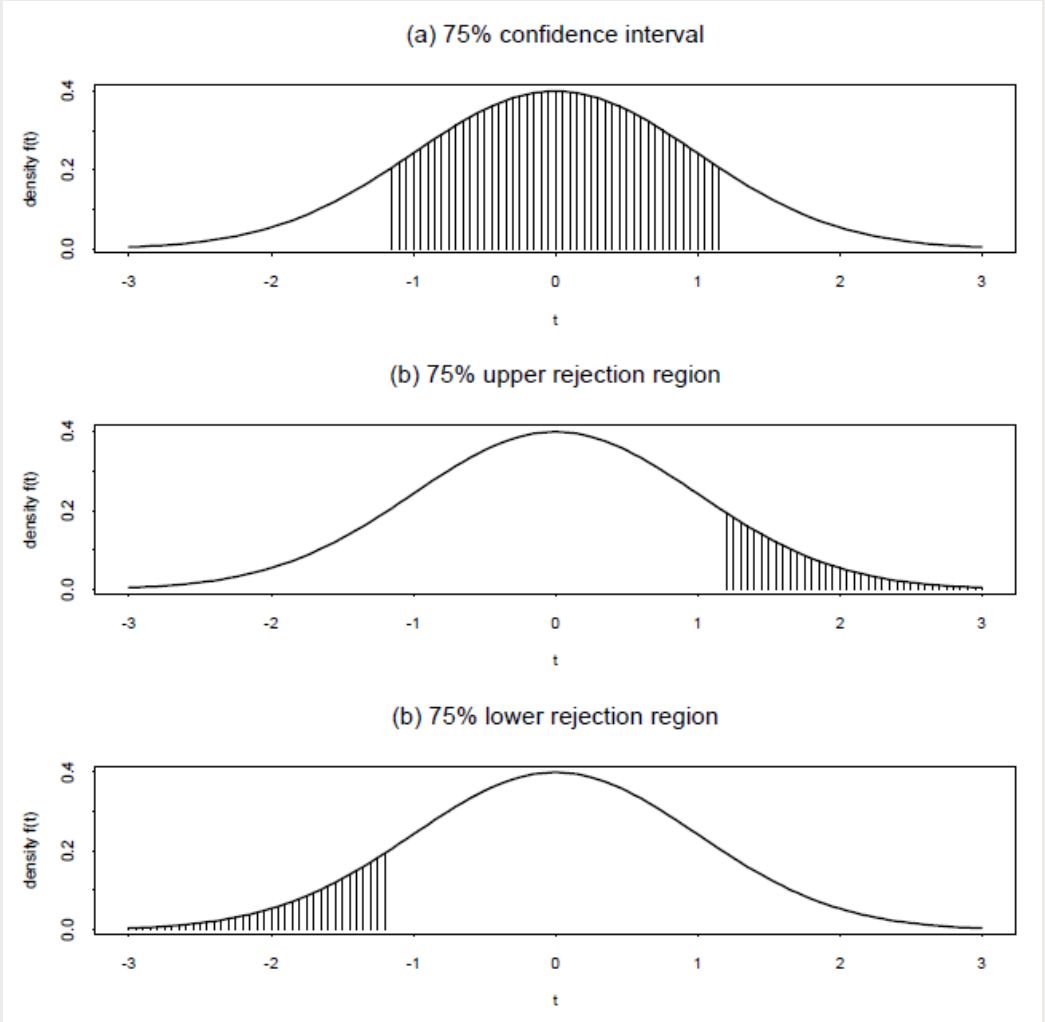
Hypothesis testing

Hypothesis testing can be single-sided, or double sided.

Double tends to be the most widely used. The alternate hypothesis is not direction specific, only states that it is different

Conversely, single sided tests are directional. e.g. a test for whether the mean is greater

In this example, the null hypothesis is only rejected in favor of the alternative hypothesis if the sample mean was significantly greater than the population mean.



One-sample Tests

One sample tests are used to test whether a particular sample could have been drawn from a population with known parameters. The tests compare an observed sample statistic with a given population parameter.

T test:

h = ttest(x,m)

Returns a test decision for the null hypothesis that the data in the vector x comes from a normal distribution with mean m and unknown variance

P > 0.05 = unable to reject null hypothesis

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Z test:

h = ztest(x,m,sigma)

Returns a test decision for the null hypothesis that the data in the vector x comes from a normal distribution with mean m and a standard deviation sigma

$$H_0 : \mu = \mu_0$$
$$\sigma = \sigma_0$$
$$H_1 : \mu \neq \mu_0$$

χ2-Test: **h = fitdist(x,'Weibull');**

h = chi2gof(x,'CDF',pd)

Returns a test decision for the null hypothesis that the data in the vectors x come from a population with a defined distribution. In the above case, a Weibull distribution.

Two-sample Tests

Two sample tests are used to test whether two samples of data could have come from the same population e.g. environmental studies before vs after intervention

T test: h = ttest(x,y)

Test the null hypothesis that the pairwise difference between data vectors x and y has a mean equal to zero.

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

F test: h = vartest2(x,y)

Returns a test decision for the null hypothesis that the data in the vectors x and y come from a normal distribution with the same variance

$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 \neq \sigma_2$$

Two-sample Tests

Non-parametric tests do not assume the data is normally distributed

Mann-Whitney U test: `p = ranksum(x,y)`

Returns a test decision for null hypothesis that data in x and y are samples from continuous distributions with equal medians

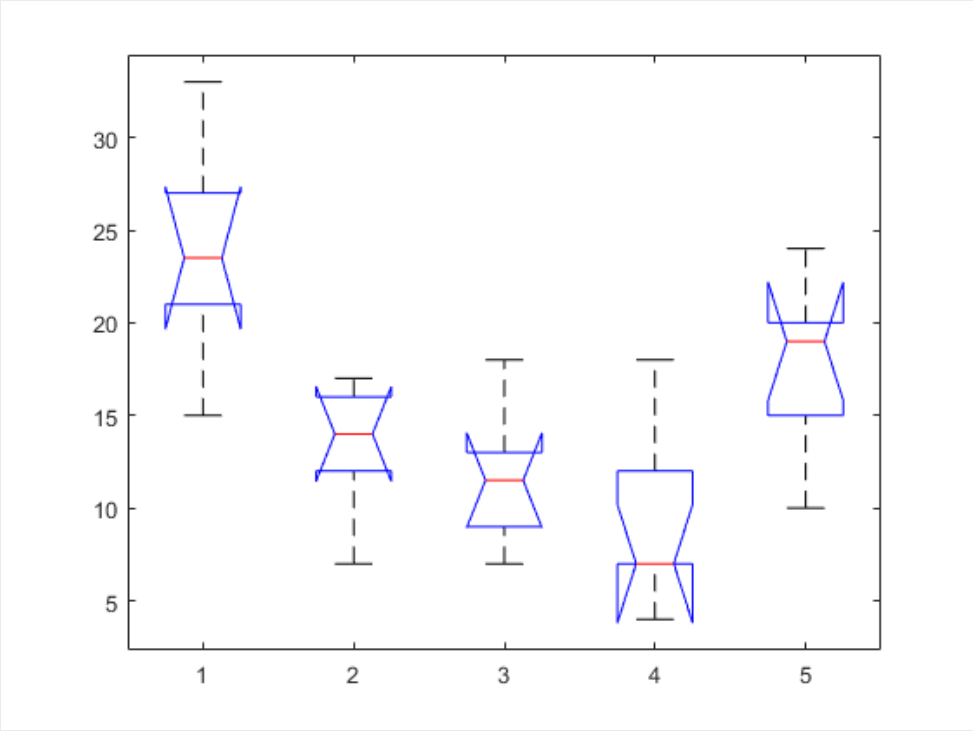
P > 0.05 = they are the same

Ansari-Bradley Test: `[h,p,stats] = ansaribradley(x,y)`

Returns a test decision for the null hypothesis that the data in vectors x and y come from the same distribution. Thee test requires that the samples have similar medians, which can be achieved by subtracting the medians from the samples

n-sample Tests

When more than two samples need to be tested, rather than using multiples of a two-sample test it is better to use an integrated multiple testing approach such as ANOVA.



[p,tbl,stats] = anova1(x);
Returns a test decision for the null hypothesis that all samples stored in variable x are drawn from populations with the same mean

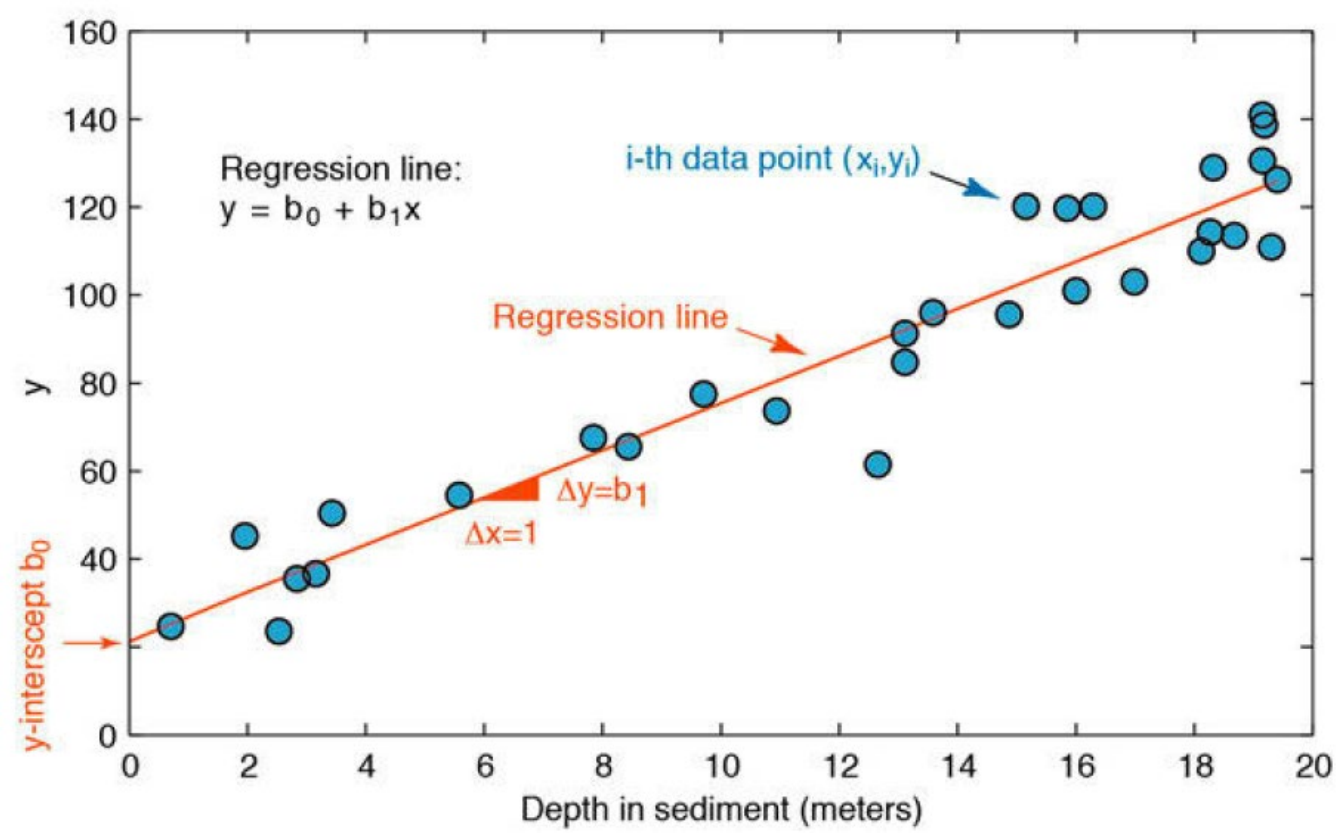
Where x is a 2D array with columns representing different sample types/conditions.

P value under the null hypothesis that all samples are drawn from populations with the same mean

$P \leq \alpha$ = At least one group mean is different

multcompare(stats)
Uses the post-hoc Tukey (or other specified) test to interactively assess which groups have means that are significantly different from other groups.

Linear regression



Principles:

$$E(Y|x) = \beta_0 + \beta_1 x$$

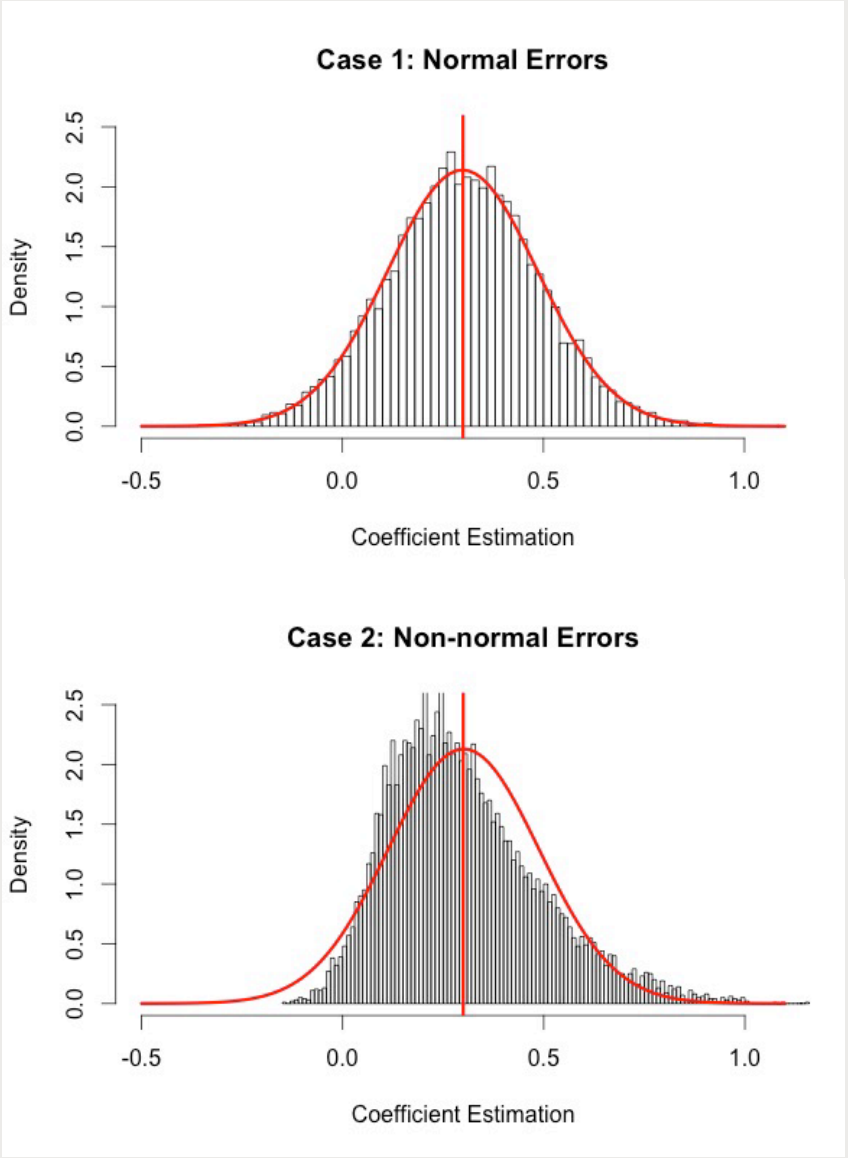
Is there a causal and linear trend between the (transformed) variables?

R2 = a ratio of explained to total variation

P value is < 0.05 we can reject the null hypothesis (i.e. $\beta_1 \neq 0$)

Normality, equal variance, and independent errors

Linear regression



Principles:

$$E(Y|x) = \beta_0 + \beta_1 x$$

Are the distributions homoscedastic?

Are transformations required?

Common types of transformations
Independent variable:

- Square root transformation of the independent variable
- Inverse transformation (X' = 1 / X)
- Log transform

Dependent and independent variable:

- Log-log transform

Uncertainty

Two broad ways of assessing confidence intervals: using bootstrapping techniques:

Calculation coefficient confidence intervals using the Wald method which assumes that errors are normally distributed

Bootstrapping confidence intervals based on resampling of cases, or resampling of residuals