Artificial Intelligence, Blockchain, e Criptovalute nello Sviluppo Software

Lezioni 13 e 14: Inferences, Non Parametric Approaches, and Logistic Regression

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Content

- More on the correlation coefficient
- Parametric and non parametric tests
- Non parametric correlations
- Logistic regression

More on the correlation coefficient

Status

- Now we know that the means of samples of a population tend to be distributed normally.
- This is an essential assumption to perform several numeric operations, like Montecarlo simulations, Bootstrap, etc.
- We would like now to understand the distribution of the Pearson momentum correlation coefficient of the sample
- Moreover, we have an open infinite issue on what to do if the data is NOT on a ratio scale

Modeling with linear models (1/2)

Linear regression is dependent on 4 hypothesis:

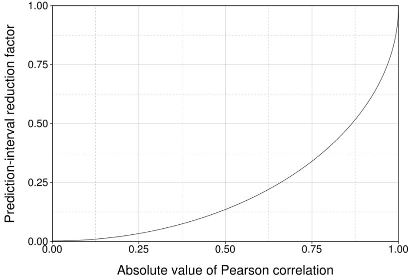
1. Normality
   The dependent variable is normally distributed at each value of the independent variables.
   How to check: histogram of standardized residuals, Q-Q plot
2. Homoscedasticity
   The variability of the standardized residuals is constant and does not depend on dependent variable.
   How to check: plotting the residuals over the mean value of dependent variable

Modeling with linear models (2/2)

1. Independence of error
   Each value of the residual does not depend in some way from the preceding value.
   How to check: Durbin-Watson statistic
2. Linearity
   There is linear dependency between regressors and response
   How to check: linear correlation coefficient

Is the correlation enough for predicting?

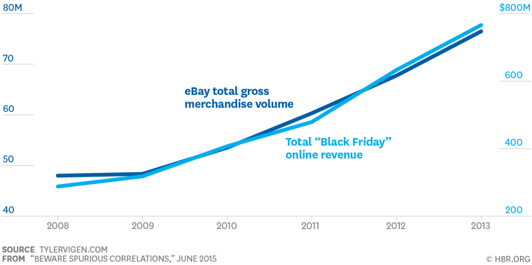
- The size of an acceptable correlation depends on the context
- A key question is what is the additional explanation that I get from analysing X vs just using Y
- The following diagram for instance shows how the 95% confidence interval is reduced for increasing values of the correlation



Source with modifications: *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*

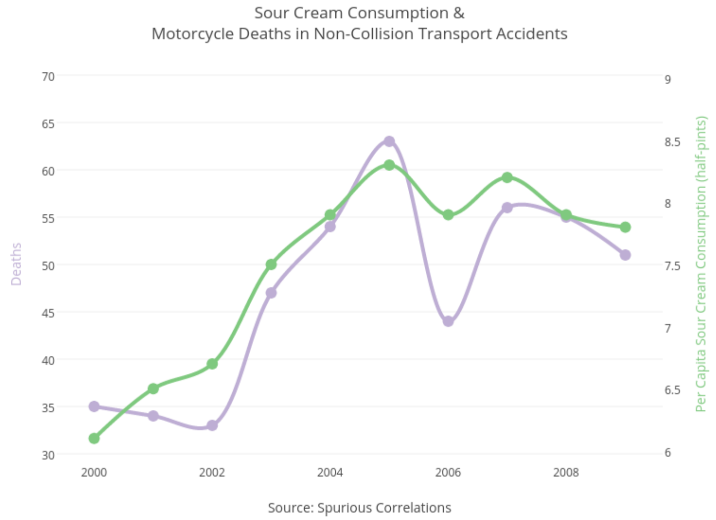
Spurious correlations. Why?

Comparing "Apples and Oranges"
Y axis scales that measure different values may show similar curves that shouldn't be paired. This becomes pernicious when the values appear to be related but aren't.
Example.

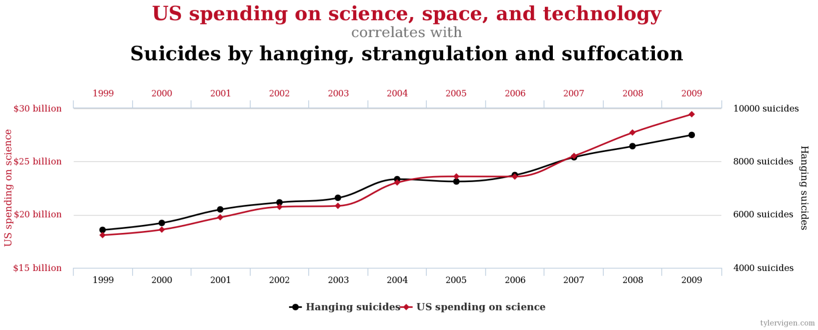


Correlation does not imply causation.

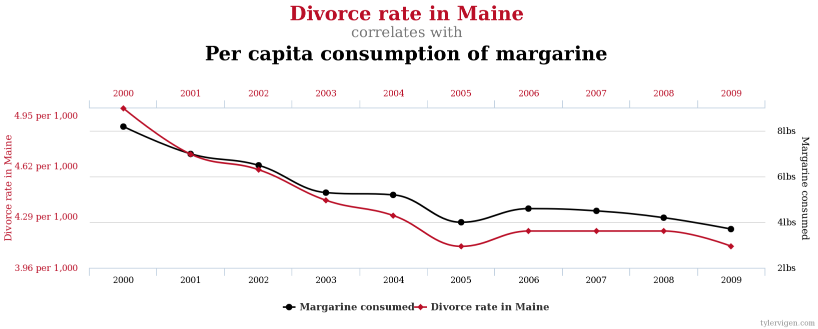
Spurious correlations



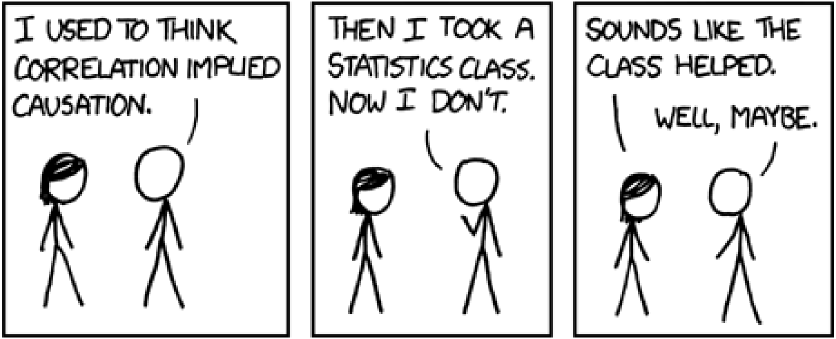
Spurious correlations



Spurious correlations



Spurious correlations



Distribution of r_{XY} of the sample

- Suppose, as usual, that we have two phenomena that we want to measure, X and Y and let us assume:
  - that there is a linear relationship between them
  - that I can express the data I collect as:

    $$\hat{Y} = \theta_0 + \theta_1 X + \epsilon$$

    where ϵ is a stationary Gaussian process $N(0, \sigma^2)$
  - that I have n samples, that is $\mathfrak{n}$ set of random pairs $\mathfrak{S}_\mathfrak{j} = \{(\mathfrak{X}_{\mathfrak{i}_\mathfrak{j}}, \mathfrak{Y}_{\mathfrak{i}_\mathfrak{j}})\}$, with:
    – $\mathfrak{j} \in [1 \dots \mathfrak{n}]$,
    – $\mathfrak{i}_\mathfrak{j} \in [1 \dots \mathfrak{m}_\mathfrak{j}]$,
    – $(\forall\ \mathfrak{j})\ \ \mathfrak{m}_\mathfrak{j} \in \mathbb{N}^+$
  - for each $\mathfrak{S}_\mathfrak{j}$ I can compute the Pearson correlation coefficient $\mathfrak{r}_{\mathfrak{X}_\mathfrak{j}\mathfrak{Y}_\mathfrak{j}}$
- What is the distribution of $\mathfrak{r}_{\mathfrak{X}_\mathfrak{j}\mathfrak{Y}_\mathfrak{j}}$?

The Student t (1/3)

- Used to determine the distribution of $\frac{\mathfrak{D}\mathfrak{n}}{\mathfrak{s}_\mathfrak{n}}$ – We know that $\frac{\mathfrak{D}\mathfrak{n}}{\sigma} \xrightarrow{d} N(0,1)$
- Apparently, started in the brewery of Guinness
- The pdf is:

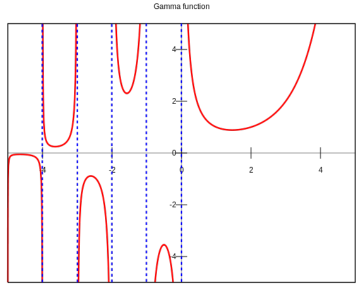
$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- We use the Γ function

Source with modifications: *https://en.wikipedia.org/wiki/Student%27s_t-distribution*

- The Γ function, an extension of the factorial to the whole $\mathbb{C}$ set apart from negative integers, that it, it is defined on $(\mathbb{R} - \mathbb{N}^-, \mathbb{R})$
- Formally:

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$$



The Student t (2/3)

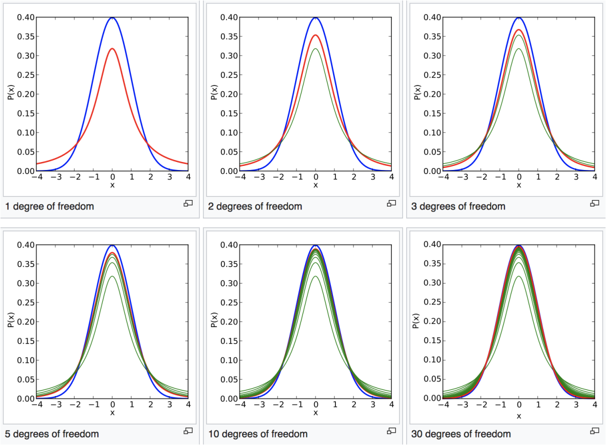
- Recall the pdf:

$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- The Student t is symmetric
- ν is the degree of freedom, as it increases the function becomes similar to the Gaussian (in the figure in Slide 18 the t is in red, the Gaussian is in blue, and the previous ts are in green)

Source with modifications: `https://en.wikipedia.org/wiki/Student%27s_t-distribution`

The Student t (3/3)



Source with modifications: *https://en.wikipedia.org/wiki/Student%27s_t-distribution*

Distribution of $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ (1/2)

- The $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ are approximated by a Student t distribution with $(n-2)$ degrees of freedom, under "good" assumptions
- Under such assumptions **and** the one that we have mentioned before, we have:

$$t = \mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}\sqrt{\frac{n-2}{1-{\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}}^2}}$$

- and conversely:

$$\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j} = \frac{t}{\sqrt{n-2+t^2}}$$

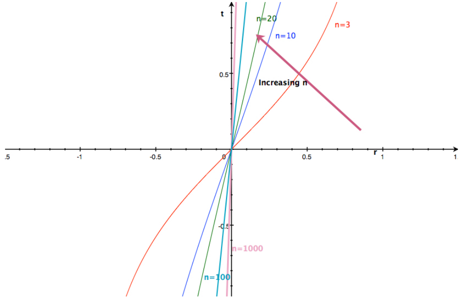
Source with modifications: *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*

Distribution of $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ (2/2)

- If we concentrate on:

$$t = \mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}\sqrt{\frac{n-2}{1-\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}{}^2}}$$

- we notice that for the same value of $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ we obtain higher values of t, with increasing values of n



Distribution of $\mathfrak{r}_{\mathfrak{X}_j \mathfrak{Y}_j}$

Claim:

$$t_{n-2} \sim \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Assamptions and Facts:

- $Y = \beta_0 + \beta_1 X + \epsilon$.
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Where $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R} \setminus \{0\}$ and $\epsilon \sim N(0, \sigma^2)$.

Plan:

- $\hat{\beta}_1 \sim \mathcal{N}$
- $RSS \sim \chi^2$
- $t \sim \frac{\hat{\beta}_1}{RSS}$

Given model $y = \beta_0 + \beta_1 x_i + \epsilon$, β_1 estimator $(\hat{\beta}_1)$ can be derived as follows :

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

Remember :

- $s_{xx} = \sum (x_i - \bar{x})^2$
- $s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \sum \frac{(x_i - \bar{x})}{s_{xx}} (y_i - \bar{y}) = \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i - \sum \frac{(x_i - \bar{x})}{s_{xx}} \bar{y}$$

Taken with modifications from https://math.stackexchange.com/questions/787939/show-that-the-least-squares-estimator-of-the-slope-is-an-unbiased-estimator-of-t/788010#788010

Therefore :

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \sum \frac{(x_i - \bar{x})}{s_{xx}} (y_i - \bar{y}) = \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i - 0 * \bar{y}$$
$$= \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i = \sum \frac{(x_i - \bar{x})}{s_{xx}} (\beta_0 + \beta_1 x_i + \epsilon_i)$$
$$= \beta_0 \sum \frac{(x_i - \bar{x})}{s_{xx}} + \beta_1 \sum \frac{(x_i - \bar{x})}{s_{xx}} x_i + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i$$

Note 1:

$$\sum \frac{(x_i - \bar{x})}{s_{xx}} = \frac{1}{s_{xx}} \sum (x_i - \bar{x}) = \frac{1}{s_{xx}} \left(\sum x_i - \sum \bar{x} \right) = 0$$

Taken with modifications from `https://math.stackexchange.com/questions/787939/show-that-the-least-squares-estimator-of-the-slope-is-an-unbiased-estimator-of-t/788010#788010`

Note 2:

$$\sum \frac{(x_i - \bar{x})}{s_{xx}} x_i = \sum \frac{(x_i - \bar{x})}{s_{xx}} (x_i - \bar{x} + \bar{x})$$
$$= \sum \frac{(x_i - \bar{x})}{s_{xx}} (x_i - \bar{x}) + \sum \frac{(x_i - \bar{x})}{s_{xx}} \bar{x}$$
$$= \frac{1}{s_{xx}} \sum (x_i - \bar{x})^2 = 1$$

Putting the simplifications into the original equation:

$$\hat{\beta}_1 = 0 \times \beta_0 + 1 \times \beta_1 + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i = \beta_1 + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i$$

- Remember that:
  - $\sum_i \frac{(x_i-\bar{x})^2}{s_{xx}} = 1$
  - $\hat{\beta}_1 = \beta_1 + \sum_i \frac{(x_i-\bar{x})}{s_{xx}} \epsilon_i$
  - $\forall i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - $\forall \lambda \neq 0, \quad \lambda\mathcal{N}(\mu, \sigma^2) = \mathcal{N}\left(\lambda\mu, (\lambda\sigma)^2\right)$
  - $\sum_i \mathcal{N}(\mu_i, \sigma_i^2) = \mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$

- Therefore:
  - $\sum \frac{(x_i-\bar{x})}{s_{xx}}\epsilon_i \sim \mathcal{N}\left(0, \sum_i \left(\frac{\sigma(x_i-\bar{x})}{s_{xx}}\right)^2\right) =$
    $= \mathcal{N}\left(0, \sigma^2 \sum_i \left(\frac{x_i-\bar{x}}{s_{xx}}\right)^2\right) = \mathcal{N}\left(0, \frac{\sigma^2}{s_{xx}}\sum_i \frac{(x_i-\bar{x})^2}{s_{xx}}\right) =$
    $\mathcal{N}\left(0, \frac{\sigma^2}{s_{xx}}\right)$
  - $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$
  - $\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{s_{xx}}} \sim \mathcal{N}(0,1)$

- Note the following:
  - We know that t_n can be rewritten using a χ^2_n distribution:

$$t_n \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/n}}$$

  - And also we can connect RSS, σ, and χ^2:

$$\frac{RSS}{\sigma^2} = \frac{\sum \epsilon_i^2}{\sigma^2} = \sum \left(\frac{\epsilon_i}{\sigma}\right)^2 = \sum \left(\mathcal{N}(0,1)\right)^2 \sim \chi^2_{n-2}$$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

- We will now proceed as follows:
  - $\hat{\beta}_1 \sim \mathcal{N}$
  - $RSS \sim \chi^2$
  - $t \sim \frac{\hat{\beta}_1}{RSS}$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

- $r^2 = 1 - \frac{RSS}{SST}$
- $RSS = (1 - r^2)s_{yy}$
- $SST = s_{yy} = \sum(y_i - \bar{y})^2 \frac{RSS}{SST}$
- $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2+\sum(y_i-\bar{y})^2}}$
- $t_{n-2} \sim \frac{\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{s_{xx}}}}{\sqrt{RSS/(n-2)\sigma^2}} = \frac{r\sqrt{(n-1)}}{\sqrt{(1-r^2)}}$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

Under null hypothesis $H_0 : \beta_1 = 0$

$$\frac{\frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{s_{xx}}}}{\sqrt{\frac{RSS}{\sigma^2(n-2)}}} = \frac{\hat{\beta}_1\sqrt{s_{xx}}}{\sigma} . \frac{\sqrt{\sigma^2(n-2)}}{\sqrt{RSS}} = \frac{\hat{\beta}_1\sqrt{s_{xx}}}{1} . \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)s_{yy}}}$$

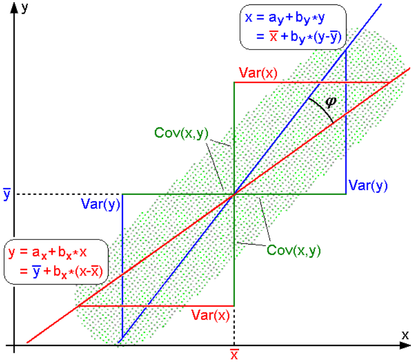
$$= \frac{\hat{\beta}_1\sqrt{s_{xx}}}{\sqrt{s_{yy}}} . \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{\frac{s_{xy}}{s_{xx}}\sqrt{s_{xx}}}{\sqrt{s_{yy}}} . \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

$$= \frac{s_{xy}}{\sqrt{s_{yy}s_{xx}}} . \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

QED

Reasoning on θ_1 (1/2)

$$\theta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_Y}{\sigma_X} r_{X,Y}$$



Source with modifications: *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*

Reasoning on θ_1 (2/2)

- Remember that:
$$\theta_1 = \frac{\sigma_Y}{\sigma_X} r_{X,Y}$$
- If $r_{X,Y} > 0$, we define the *p*-value of our correlation the $P(\theta_1 < 0)$, conversely, if $r_{X,Y} < 0$ we define the *p*-value of our correlation the $P(\theta_1 > 0)$
- In other terms, the *p*-value of a correlation is the probability that a slope change direction

Source with modifications: `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

Parametric and non parametric tests

Parametric and non parametric tests (1/2)

- We have already discussed that in the framework of empirical research we may want to know if two datasets have a probability of not being different smaller than a given α
- The probability refers to the populations from which the datasets have been taken
- The question is then rephrased if two dataset have the probability of being drawn from the same population smaller than a given α
- Typically, this α is 0.1, 0.05, or 0.01
- There are two main situations to consider based on our knowledge of the datasets,
  - when we know something about the dataset or
  - when we cannot make any assumption on it

Parametric and non parametric tests (2/2)

We can distinguish two major classes of tests:

- When we **can make assumptions** on the distributions of the two datasets;
  - for this case, we have the *parametric* tests, since we can assume parameters of the underlying distribution
- When we **cannot**
  - for this case, we have the *non parametric* tests, since we cannot make any assumption on any kind of parameter of the underlying distribution

Taken with modifications from `https://en.wikipedia.org/wiki/Power_(statistics)`

Parametric tests: Z Test

- The first test that we consider is the Z Test
- It is based on the analysis of the normal distribution
- It assumes that the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution
- Remember the CLT: there is a large number of datasets distributed *somehow* normally, such as the means of samples …

Taken with modifications from `https://en.wikipedia.org/wiki/Z-test`

Considerations on the Z-test

- A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution.
- The term "Z-test" is often used to refer specifically to the one-sample location test comparing the mean of a set of measurements to a given constant when the sample variance is known.

Hypothesis for the Z test

The fundamental hypotheses to have an exact application of the Z test are:

- Known standard deviation
- *to be precise, we need to know the nuisance parameter*
  - *a parameter that is not object of study*
  - *but which is needed to determine the object of study*
  - *like the standard deviation is for the mean, if we are analysing the mean value of a distribution*
- The test statistic should follow a normal distribution.

Taken with modifications from `https://en.wikipedia.org/wiki/Z-test` and `https://en.wikipedia.org/wiki/Nuisance_parameter`

Z test for the evaluation of the mean (1/2)

- Let as assume to have a sample $\{X_1 \dots X_n\}$ with average $\overline{X}$
- We assume that this sample comes from a population with mean μ
- We want to determine whether this sample does not belong to a population distributed as $N(\mu_0, \sigma)$ with an acceptable probability of error α
- We can build the null hypothesis as: $H_0 : \mu = \mu_0$
- Against the alternate hypothesis: $H_1 : \mu \neq \mu_0$

Taken with modifications from `https://it.wikipedia.org/wiki/Test_Z`

Z test for the evaluation of the mean (2/2)

- We need to compute the Z statistics:

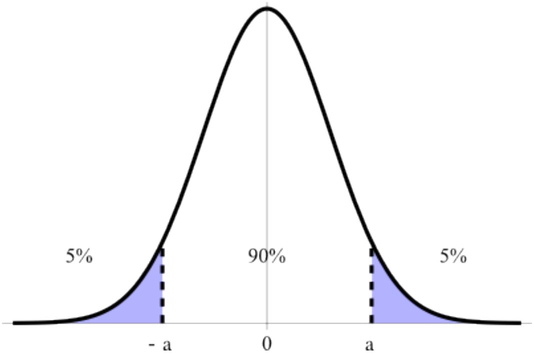
$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

- We need then to look up in the table the probability of Z, considering that in this case we accept a variation with respect to μ_0 in both directions (lower or larger)

Taken with modifications from `https://it.wikipedia.org/wiki/Test_Z`

Graphical interpretation

Assuming $\alpha = 0.1$:



Taken with modifications from `https://it.wikipedia.org/wiki/Test_Z`

The problem of multiple testing

Consider a case where you have **20** hypotheses to test, and a significance level of **0.05**.

The probability of observing at least one significant result just due to chance?

$$\mathbb{P}(at\_least\_1\_signif.\_results) = 1 - \mathbb{P}(no\_signif.\_results) =$$
$$= 1 - (1 - 0.05)^{20} \approx 0.64$$

The Bonferroni correction

So, with 20 tests being considered, we have a **64%** chance of observing at least one significant result, even if all of the tests are actually not significant.

The Bonferroni correction sets the significance cut-off at α/n.

For example, with **20** tests and $\alpha = 0.05$, you'd only reject a null hypothesis if the p-value is less than **0.0025**.

Toward the Bonferroni inequality (1/2)

Claim (Boole Inequality): Let A_1, A_2, …, A_n be n events, then:

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

Proof: *By induction*

Base

For n=1 it is trivially verified.

For n=2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$$

since $P(A_1 \cap A_2) \geq 0$.

Toward the Bonferroni inequality (2/2)

Step

Assuming it true for $n \geq 2$, we prove it for $n+1$.

Given the associativity of the $\cup$:

$$P(\bigcup_{i=1}^{n+1} A_i) = P(\bigcup_{i=1}^{n} A_i \cup A_{n+1})$$

Calling B the set $\bigcup_{i=1}^{n} A_i$ and C the set A_{n+1} we can write:

$$P(B \cup C) = P(B) + P(C) - P(B \cap C) \leq P(B) + P(C)$$

Which means:

$$P(\bigcup_{i=1}^{n+1} A_i) \leq P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) \leq \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) = \sum_{i=1}^{n+1} P(A_i)$$

QED

The Bonferroni correction - Proof

Claim (Bonferroni correction): In the case of m null hypotheses $H0_1 \cdots H0_m$ sufficient condition to have a probability than a given α of wrongly rejecting a null hypothesis is that $\forall i \in [1 \cdots m], p_i \leq \frac{\alpha}{m}$.

Proof:

From the Boole inequality:

$$P\big(\bigcup_{i=1}^{m}(p_i \leq \frac{\alpha}{m})\big) \leq \sum_{i=1}^{m}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m\frac{\alpha}{m} = \alpha$$

QED

The power function of a test

Remember that:

- A type 1 error is when we reject the null hypothesis when the null hypothesis is true, that is we think that something is going on, but nothing is really there. The probability of committing a type 1 error is typically referred to as α.
- A type 2 error is when we fail to reject the null hypothesis when actually we should reject it, that is, we fail to perceive a phenomena. The probability of committing a type 2 error is typically referred to as β.

The power function of a test informally is the probability of not committing a type 2 error, that is, $(1-\beta)$

Power of a test

- In general the non rejection of the null hypothesis H0 does not mean that H0 holds
- The power of a binary test is the probability that the tests rejects the null hypothesis when the alternate hypothesis is true

$$\text{Power(Test)} = \text{P(reject H0} \mid \text{ H1 is true)}$$

- If a test has power of 0.99 in a given situation, it means that the non rejections of H0 means that H0 holds with a p(error) ≤ 0.01
- The power of a test has an essential role in determining the test to select and in interpreting its results

Taken with modifications from https://en.wikipedia.org/wiki/Power_(statistics)

What influences the power of a test

The power of a test is influenced by a variety of factors, such as:

- the size of the datasets
- the magnitude of the effect
- the level of statistical significance
- the intrinsic structure of a test
  - we can use a test only if its hypotheses are all verified
  - informally, the more stringent the hypotheses, the higher the power of the test, since …
  - … *we know better the population*

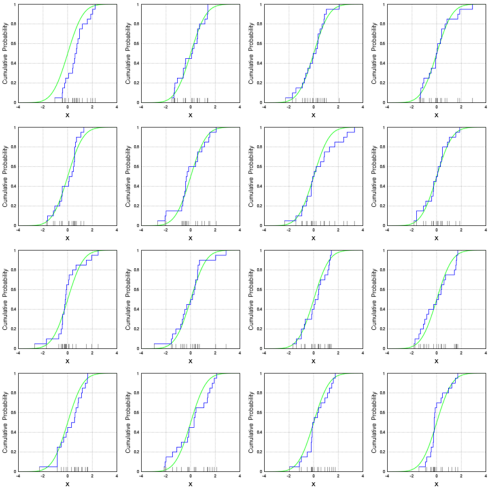
Taken with modifications from `https://en.wikipedia.org/wiki/Power_(statistics)`

The Empirical CDF

- Typically, in experimental situations, we recreate the CFD and the PDF from a set of datapoints
- After this we recreate and we assume some sort of regularity between such points and we interpolate
- Indeed, the ability to reconstruct a random process based on as set of observations depends on:
  - its "variability"
  - the number of such points
- Anyway, from a given set of point I can build an Empirical CDF (ECDF)

Taken with modifications from https://en.wikipedia.org/wiki/Empirical_distribution_function

Multiple ECDF from the same process



Taken with modifications from https://en.wikipedia.org/wiki/Empirical_distribution_function

ECDF (1/2)

- Let $X_1 \dots X_n$ be a set of n iid random variables coming from a CDF $F_X(t)$
- We define the empirical distribution function $\widehat{F}_n(t)$:

$$\widehat{F}_n(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq t}$$

- Where $\mathbf{1}_A$ is the indicator of event A, that is,
  - For a fixed t, the indicator $\mathbf{1}_{X_i \leq t}$ is a Bernoulli random variable with probability of success $p = F_X(t)$
  - Therefore, $n \times \widehat{F}_n(t)$ is a binomial random variable
  - the mean of $n \times \widehat{F}_n(t)$ is $n \times F_n(t)$
  - the variance of $n \times \widehat{F}_n(t)$ is $n \times F_n(t) \times (1 - F_n(t))$
- Therefore $n \times \widehat{F}_n(t)$ is an unbiased estimator of $F_n(t)$

Taken with modifications from `https://en.wikipedia.org/wiki/Empirical_distribution_function`

ECDF (2/2)

The ECDF has interesting asymptotic properties (from the strong LLN etc – not to demonstrate):

- Convergence:

$$\widehat{F}_n(t) \xrightarrow{\text{a.s.}} F(t)$$

- Consistency:

$$\|\widehat{F}_n - F\|_\infty \equiv \sup_{t \in \mathbb{R}} \left| \widehat{F}_n(t) - F(t) \right| \xrightarrow{\text{a.s.}} 0$$

- Normality of error:

$$\sqrt{n}\big(\widehat{F}_n(t) - F(t)\big) \ \xrightarrow{d}\ \mathcal{N}\Big(0, F(t)\big(1 - F(t)\big)\Big).$$

All these properties allow us to perform experimentations in the "traditional" way.

Taken with modifications from `https://en.wikipedia.org/wiki/Empirical_distribution_function`

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) Goodness-of-Fit Test

- Purpose of the K-S test
- Characteristics and Limitations of the K-S test
- Definition

Kolmogorov distribution

Remember that the empirical distribution function for n iid ordered observation of random variables X_i is:

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,x]}(X_i)$$

The, the Kolmogorov Smirnov statistics for a global CDF $F(x)$ is:

$$D_n = \sup_x |\widehat{F}_n(x) - F(x)|$$

If the sample comes from the distribution $F(x)$ then D_n goes to 0. But the question is "how fast?"

Taken with modifications from `https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test`

$\widehat{F}_n$ is an unbiased estimator of F

Proof.

- Let $x \in \mathbb{R}$
  - Then, $n\widehat{F}_n(x) \sim B(n, F(x))$
- Remember that $E(B(n,p)) = np$
  - $E[n\widehat{F}_n(x)] = nF(x)$
  - $E[\widehat{F}_n(x)] = F(x)$

QED.

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

Proof.

- Let $x \in \mathbb{R}$
- Remember that $\mathrm{Var}(B(n,p)) = np(1-p)$
- Then, $\mathrm{Var}(n\widehat{F}_n(x)) = nF(x)(1-F(x))$
- If we consider $\mathrm{Var}(\widehat{F}_n(x)) = E[(\widehat{F}_n(x) - E[\widehat{F}_n(x)])^2]$
  - $\mathrm{Var}(\widehat{F}_n(x)) = E[(\widehat{F}_n(x) - F(x))^2] = \frac{n^2}{n^2}E[(\widehat{F}_n(x) - F(x))^2] =$ $= \frac{1}{n^2}E[n^2(\widehat{F}_n(x) - F(x))^2] = \frac{1}{n^2}E[(n\widehat{F}_n(x) - nF(x))^2]$ $= \frac{1}{n^2}\mathrm{Var}(n\widehat{F}_n(x))$
- Remember that $\mathrm{Var}(B(n,p)) = np(1-p)$
  - Then, $\mathrm{Var}(n\widehat{F}_n(x)) = nF(x)(1-F(x))$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf

$\widehat{F}_n$ is a consistent estimator of F (2/2)

- $\mathrm{Var}(\widehat{F}_n(x)) = \frac{1}{n^2}\mathrm{Var}(n\widehat{F}_n(x)) =$

$$\frac{1}{n^2} nF(x)(1 - F(x)) = \frac{nF(x)(1 - F(x))}{n^2} = \frac{F(x)(1 - F(x))}{n}$$

- Then we have:

$$\lim_{n\to\infty} \mathrm{Var}(\widehat{F}_n(x)) = \lim_{n\to\infty} \frac{F(x)(1 - F(x))}{n} = 0$$

- And we conclude:

$$(\forall x \in \mathbb{R}) \ \lim_{n\to\infty} \mathrm{Var}(\widehat{F}_n(x)) = 0$$

QED.

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf

Theorem of Glivenko–Cantelli

Let $\widehat{F}_n$ be the empirical distribution function for n iid ordered observation of random variables X_i coming from a population with a global CDF $F(x)$. Let D_n be the Kolmogorov-Smirnov statistics defined as:

$$D_n = \sup_{-\infty<x<+\infty} |\widehat{F}_n(x) - F(x)|$$

Then:

$$\overset{P}{\lim_{n\to\infty}} D_n = 0$$

Meaning:

$$(\forall\epsilon \in \mathbb{R}^+, \forall\eta \in \mathbb{R}^+)(\exists\widehat{n} \in \mathbb{N})|(\forall n \in \mathbb{N}, n \geq \widehat{n})P(|D_n| < \eta) > (1-\epsilon)$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

The Distribution-Free Property of D_n states that the distribution of D_n is the same for all underlying F.

Proof *assuming that F is monotonically increasing.*

- If F is monotonically increasing, then F^{-1} exists and is also monotonically increasing.
- Since x ranges in $(-\infty, +\infty)$, then $y = F(x)$:
  - y exists
  - is unique, and
  - is in the range $[0, 1]$
- So we can rewrite:

$$D_n = \sup_{-\infty < x < +\infty} |\widehat{F}_n(x) - F(x)| =$$

$$= \sup_{0 \le y \le 1} |\widehat{F}_n(F^{-1}(y)) - F(F^{-1}(y))| = \sup_{0 \le y \le 1} |\widehat{F}_n(F^{-1}(y)) - y|$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah,

The Distribution-Free Property of D_n (2/4)

- Now, remember that:

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,x]}(X_i)$$

- Therefore:

$$\widehat{F}_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,F^{-1}(y)]}(X_i)$$

- Also we remember the definition of indicator function:

$$\mathbf{1}_{X_i \leq x} = \mathbf{1}_{[-\infty,x]}(X_i) \begin{cases} 1, \text{ if the event } X_i \text{ is } \leq x \\ 0, \text{otherwise} \end{cases}$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

The Distribution-Free Property of D_n (3/4)

- F is strictly monotonically increasing:

$$a < b \iff F(a) < F(b)$$

- Therefore

$$\mathbf{1}_{[-\infty,x]}(X_i) = \mathbf{1}_{[0,F(x)]}F(X_i) = \mathbf{1}_{[0,F(F^{-1}(y))]}F(X_i) = \mathbf{1}_{[0,y]}(F(X_i))$$

- since:

$$x < X_i \iff F(x) < F(X_i)$$

- and we can conclude that:

$$\widehat{F}_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[0,F(F^{-1}(y))]}(F(X_i)) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[0,y]}(F(X_i))$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

The Distribution-Free Property of D_n (4/4)

- Consider:

$$\widehat{F}_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[0,y]}(F(X_i))$$

- This is the empirical distribution function for iid random sample formed by the $\{F(X_i)\}$.
- It is a sample from the uniform distribution on the range $[0, 1]$.
- Therefore, it does not depend on F, indeed, provided that F is strictly monotonically increasing.

QED

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf

Ordered statistics

- Let $X_1 \dots X_n$ be a set of n iid random variables coming from a CDF F
- Let $X_{1:n} \dots X_{n:n}$ be reordering of the variables so that $(\forall h, k \in \mathbb{N} | 0 < h < k \leq n)\ (X_{h:n} \leq X_{k:n})$
- The sequence $X_{1:n} \dots X_{n:n}$ is defined an *ordered statistics* of $X_1 \dots X_n$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf

- Let $X_1 \ldots X_n$ be a set of n iid random variables coming from a *continuous* CDF F
- Let $X_{1:n} \ldots X_{n:n}$ be an ordered statistics of $X_1 \ldots X_n$
- Then, we have that:

$$(\forall h, k \in \mathbb{N} | 0 < h < k \leq n)\ (P(X_{h:n} = X_{k:n}) = 0)$$

Proof (simplified)

- Let us consider $X_{k:n} \ldots X_{n:n}$
- We want to see if $(\exists j \in \mathbb{N}, j \in [1, n], j \neq k) | (X_{j:n} = X_{k:n})$
- Since F is continuous:

$$P(X = X_{k:n}) = \int_{X_{k:n}} F(z)dz = 0$$

Since it is an integral of a continuous function on a point.

QED

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah,

Rewriting of D_n (1/2)

- Let $X_1 \dots X_n$ be a set of n iid random variables coming from a *continuous* CDF F
  - So all X_i are different.
- Remember that D_n is distribution free
  - So for the purpose of this computation we can assume that $X_1 \dots X_n$ is uniform in the interval of our interest, in our case $[0, 1]$.
- So we can write:

$$D_n = \sup_{-\infty < x < +\infty} |\widehat{F}_n(x) - F(x)| = \sup_{-\infty < x < +\infty} |\widehat{F}_n(x) - x|$$

$$= \sup_{0 \le x \le 1} |\widehat{F}_n(x) - x| =$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf

Rewriting of D_n (2/2)

- Given the stepwise structure of $\widehat{F}_n$ the max can occur only at the points of "jump" so we have that

$$D_n = \sup_{0 \leq x \leq 1} \left\{ |X_{i:n} - \widehat{F}_n(X_{(i-1):n})|, |\widehat{F}_n(X_{i:n}) - X_{i:n}| \right\} =$$

- Lastly, it is evident that:

$$\widehat{F}_n(X_{i:n}) = \frac{i}{n}$$

- Therefore we have:

$$D_n = \sup_{0 \leq x \leq 1} \left\{ |X_{i:n} - \frac{i-1}{n}|, |\frac{i}{n} - X_{i:n}| \right\} =$$

QED

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

Original work by Kolmogorov

The original Kolmogorov statistics was defined as follows:

- Let $x_1 \ldots x_n$ be an **ordered** set of iid random variables
- We can then define the Kolmogorov statistics D_n as:

$$D_n = max(x_1 - \frac{0}{n}, x_2 - \frac{1}{n}, \ldots, x_n - \frac{n-1}{n},$$

$$\frac{1}{n} - x_1, \frac{2}{n} - x_2, \ldots, \frac{n}{n} - x_n)$$

Taken with modifications from: George Marsaglia, Wai Wan Tsang, Jingbo Wang, *Evaluating Kolmogorov's Distribution*, Journal of statistical software, 8:1-14, January 2003

Example of computation of D_n (1/3)

- Suppose we have the following datapoints d_i: 1.41, 0.26, 1.97, 0.33, 0.55, 0.77, 1.46, 1.18
- We want to determine if they are not randomly sampled from a uniform distribution ranging from 0 to 2.
- Our H_0 is therefore that the data come from such uniform distribution and our H_1 that it does not come from it
- Let od_i be the sequence of ordered datapoints: 0.26, 0.33, 0.55, 0.77, 1.18, 1.41, 1.46, 1.97.
- Notice that with respect to the formulation above we need to get the values from the underlying uniform distribution

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Example of computation of D_n (2/3)

i	od_i	$\frac{i-1}{n}$	$\frac{i}{n}$	$X_{i:n}$	$\lvert X_{i:n} - \frac{i-1}{n}\rvert$	$\lvert \frac{i}{n} - X_{i:n}\rvert$
1	0.26	0	0.125	0.13	0.13	0.05
2	0.33	0.125	0.25	0.65	0.04	0.085
3	0.55	0.250	0.375	0.275	0.025	0.1
4	0.77	0.375	0.5	0.385	0.01	0.115
5	1.18	0.5	0.625	0.590	0.09	0.035
6	1.41	0.625	0.75	0.705	0.08	0.045
7	1.46	0.75	0.875	0.730	0.02	**0.145**
8	1.97	0.875	1	0.985	0.09	0.015

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Example of computation of D_n (3/3)

$$D_n = \sup_x[|F_n(x) - F_0(x)|]$$
$$\alpha = 1 - P(D_n \leq d)$$

	α			
n	0.20	0.10	0.05	0.01
1	0.90	0.95	0.98	0.99
2	0.68	0.78	0.84	0.93
3	0.56	0.64	0.71	0.83
4	0.49	0.56	0.62	0.73
5	0.45	0.51	0.56	0.67
6	0.41	0.47	0.52	0.62
7	0.38	0.44	0.49	0.58
8	0.36	0.41	$\boxed{0.46}$	0.54
9	0.34	0.39	0.43	0.51
10	0.32	0.37	0.41	0.49

0.46 is larger than 0.145, so we cannot reject the null hypothesis.

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Another exercise (1/5)

- We have available the following datapoints d_i: 108, 112, 117, 130, 111, 131, 113, 105, 128
- We want to determine if they come from a normal distribution with $\mu = 100$ and $\sigma = 10$.
- Our H_0 is therefore that the data come from such uniform distribution and our H_1 that it does not come from it
- Let od_i be the sequence of ordered datapoints: 105, 108, 111, 112, 113, 117, 128, 130, 131.
- Our significance level is 0.05.
- Notice that with respect to the formulation above we need to get the values from the underlying normal distribution using the suitable table.
- Consider also what would happen if the significance level were 0.2

Taken with modifications from https://newonlinecourses.science.psu.edu/stat414/node/323/

Another exercise (2/5)

i	od_i	$\frac{i-1}{n}$	$\frac{i}{n}$	$X_{i:n}$	$\lvert X_{i:n} - \frac{i-1}{n}\rvert$	$\lvert \frac{i}{n} - X_{i:n}\rvert$
1	105	-	-	0.0668	-	-
2	108	-	-	0.1151	-	-
3	111	-	-	0.1841	-	-
4	112	-	-	0.2119	-	-
5	113	-	-	0.2420	-	-
6	117	-	-	0.3821	-	-
7	128	-	-	0.7881	-	-
8	130	-	-	0.8413	-	-
9	131	-	-	0.8643	-	-

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Another exercise (3/5)

n\α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

Taken with modifications from
https://i0.wp.com/www.real-statistics.com/wp-content/uploads/2012/11/one-sample-ks-table.png

Another exercise (4/5)

i	od_i	$\frac{i-1}{n}$	$\frac{i}{n}$	$X_{i:n}$	$\lvert X_{i:n} - \frac{i-1}{n}\rvert$	$\lvert \frac{i}{n} - X_{i:n}\rvert$
1	105	0	0.1111	0.0668	0.0668	0.0443
2	108	0.1111	0.2222	0.1151	0.00399	0.1071
3	111	0.2222	0.3333	0.1841	0.0381	0.1492
4	112	0.3333	0.4444	0.2119	0.1214	0.2325
5	113	0.4444	0.5555	0.2420	0.2024	**0.3136**
6	117	0.5555	0.6667	0.3821	0.1734	0.2846
7	128	0.6667	0.7778	0.7881	0.1214	0.0103
8	130	0.7778	0.8889	0.8413	0.0635	0.0476
9	131	0.8889	1	0.8643	0.0246	0.1357

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Another exercise (5/5)

- For $n = 9$ and for $\alpha = 0.05$ the threshold level for the maximum value of D_n is 0.43001 (see Slide 75)
- In our case the maximum value of D_n is 0.3136.
- Therefore, we cannot reject the null hypothesis (see Slide 76)
- Therefore, we cannot reject the null hypothesis
- If the α value were 0.2, then the threshold level would be 0.33907 so still we could not reject the null hypothesis

Taken with modifications from `https://newonlinecourses.science.psu.edu/stat414/node/323/`

Confidence interval of D_n

- So far we have seen the structure of D_n
- Still to use it we need to determine the confidence interval
- We have used above the tabled values
- But how such values were generated?

Remember the Binomial (1/2)

- Remember that a Binomial distribution can be approximated by a Normal distribution:

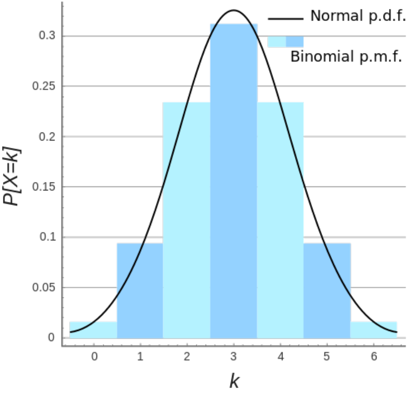
$$B(n, p) \sim N(np, np(1-p))$$

- The approximation works properly if:
  - n is large enough
  - p is far away from the extreme 0 and 1
- A possible strict rule is to have both np and $n(1-p)$ larger than 9

Taken with modifications from `https://en.wikipedia.org/wiki/Binomial_distribution`

Remember the Binomial (2/2)

This is the plot of a Binomial pmf when n is 0.6 and p 0.5, and its Normal approximation with $\mu = 3$ and $\sigma = 1.25$.



Taken with modifications from `https://en.wikipedia.org/wiki/Binomial_distribution`

Binomial and $\widehat{F}_n$

Remember that

- Let $x \in \mathbb{R}$
  - Then, $\widehat{F}_n(x) \sim \text{Bernoulli}(F(x))$
- Since $E(B(n,p)) = np$
  - Then, $n\widehat{F}_n(x) \sim B(n, F(x))$
- Given $Y \sim N(\mu, \sigma^2)$
  - Let $Z = \frac{Y-\mu}{\sigma}$
  - Then $Z \sim N(0,1)$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

From $\widehat{F}_n$ to N

- The above reasoning lead us to:

$$\overset{d}{\lim_{n\to\infty}} \frac{n\left[\widehat{F}_n(x) - F(x)\right]}{\sqrt{nF(x)\left[1 - F(x)\right]}} = N(0,1)$$

- Since we also know the following, we can conclude:

$$\overset{d}{\lim_{n\to\infty}} = F(x)$$

$$\overset{d}{\lim_{n\to\infty}} \frac{n\left[\widehat{F}_n(x) - F(x)\right]}{\sqrt{n\widehat{F}_n(x)\left[1 - \widehat{F}_n(x)\right]}} = \overset{d}{\lim_{n\to\infty}} \frac{\sqrt{n}\left[\widehat{F}_n(x) - F(x)\right]}{\sqrt{\widehat{F}_n(x)\left[1 - \widehat{F}_n(x)\right]}} =$$

$$= N(0,1)$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

Confidence interval (1/2)

- Remember that for $X = N(\mu, \sigma^2)$ we can compute the confidence interval as:

$$(\mu - \hat{z}\frac{\sigma}{\sqrt{n}}, \mu + \hat{z}\frac{\sigma}{\sqrt{n}})$$

Where $\hat{z}$ is the value of the standard normal distribution for the desired confidence interval

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

Confidence interval (2/2)

- Therefore, the asymptotic $(1-\alpha)$ confidence interval for $\widehat{F}_n$ is:

$$
\left(\widehat{F}_n(x) - z_{\alpha/2}\sqrt{\frac{\widehat{F}_n(x)\left[1-\widehat{F}_n(x)\right]}{n}},\right.
$$

$$
\left.\widehat{F}_n(x) + z_{\alpha/2}\sqrt{\frac{\widehat{F}_n(x)\left[1-\widehat{F}_n(x)\right]}{n}}\right)
$$

Taken with modifications from: Davar Khoshnevisan, *Empirical Processes, and the Kolmogorov–Smirnov Statistic*, Lectured notes of Math 6070, Spring 2014, University of Utah, `https://www.math.utah.edu/~davar/math6070/2014/Kolmogorov-Smirnov.pdf`

Kolmogorov distribution

CDF of Kolmogorov distribution:

$$F_X(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, & x > 0; \\ 0, & x \leqslant 0. \end{cases}$$

$X \sim \mathrm{K}$

Taken with modifications from https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test

Kolmogorov Theorem

Let $X_1, \ldots, X_n, \ldots$ is an infinite sample from a continuous distribution $F(x)$. Let $F_n(x)$ is a empirical CDF build on first n elements of the sample.

Then

$$\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to K$$

$n \to \infty$,
where K is a r.v. that has Kolmogorov distribution.

Purpose of the K-S test

The K-S test is used to decide if a sample comes from a population with a specific distribution.
The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given N ordered data points $Y_1, \ldots, Y_N$ the ECDF is defined as

$$E_N = n(i)/N$$

where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by $1/N$ at the value of each ordered data point.

Definition

The Kolmogorov-Smirnov test is defined by:

H_0 : The data follow a specified distribution

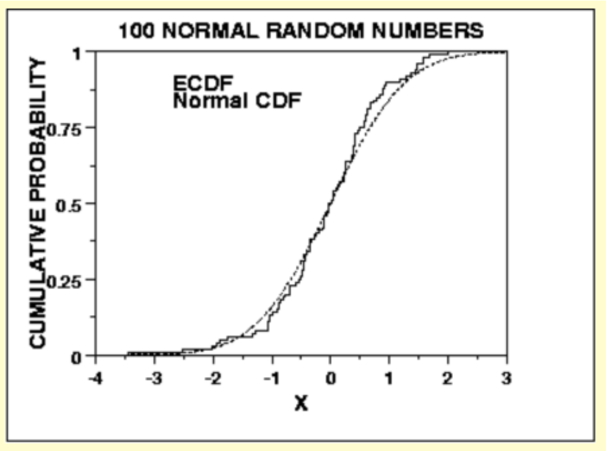
H_1 : The data do not follow the specified distribution

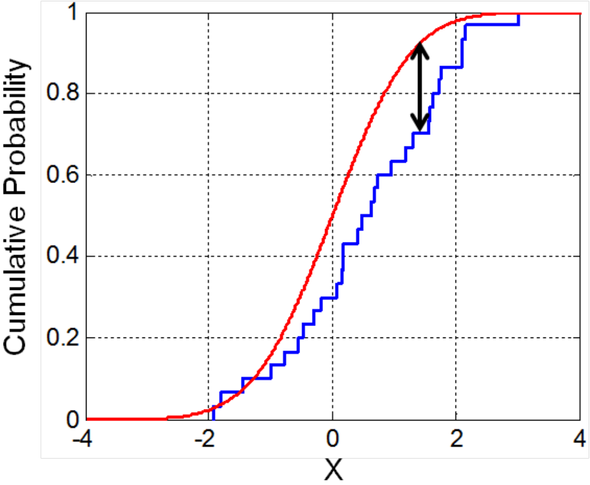
Test Statistic:

$$D = \max_{1 \leq i \leq N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i))$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution

is rejected if the test statistic, D, is greater than the critical value obtained from a table.


100 NORMAL RANDOM NUMBERS
ECDF
Normal CDF
CUMULATIVE PROBABILITY
1
0.75
0.5
0.25
0
-4
-3
-2
-1
0
1
2
3
X



Cumulative Probability
1
0.8
0.6
0.4
0.2
0
-4
-2
0
2
4
X

Characteristics and Limitations

Advantage: An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested.

The K-S test has several important **limitations**:

- It only applies to continuous distributions.
- It tends to be more sensitive near the center of the distribution than at the tails.
- The distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

A test with no restrictions

- We can consider to have m distributions of categorical variables, which can assume n values
- We want to assume as null hypothesis that they come from the same underlying distribution
- If the numbers of occurrences of such variables are "high enough" we can use the χ^2 test
- Otherwise?
  - We can resort to combinatorial calculus

Taken with modifications from `http://mathworld.wolfram.com/FishersExactTest.html`

Non parametric correlations

Spearman's Rank Correlation Coeff. (1/3)

- What can we do when the data is not normally distributed?
- Or even if the data is not on a ratio scale, just on an ordinal scale?

- *If the data is on a nominal scale, the concept of correlation looses interest; at most we can consider clustering.*

Spearman's Rank Correlation Coeff. (2/3)

Idea:

- Transform the data into ranks
- Apply the Pearson correlation coefficient to ranks
- Indeed, the values can be different, and also the significance and the mutual relationship
- Remember that:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- And also that:

$$\theta_1 = \frac{\sigma_X \sigma_Y}{Var(X)} r_{X,Y}$$

Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

Spearman's Rank Correlation Coeff. (3/3)

Definition:

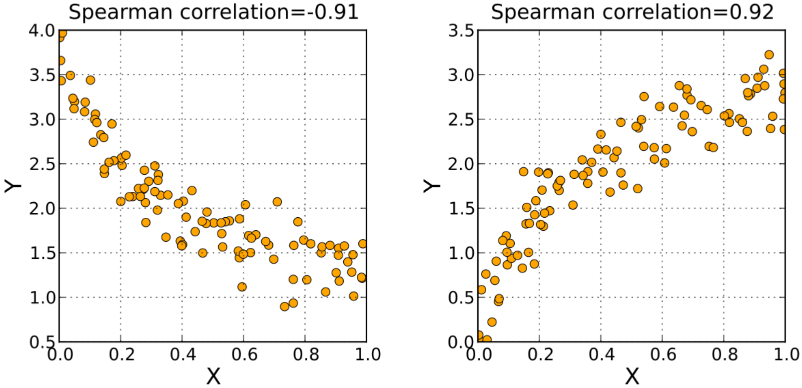
- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's define
  - $(\forall X_i \in X)\ rk_{X\,i} = \text{rank}(X_i, X),\ Rk_X = \{rk_{X_i}\}$
  - $(\forall Y_i \in Y)\ rk_{Y_i} = \text{rank}(Y_i, Y),\ Rk_Y = \{rk_{Y_i}\}$
- We define the Spearman's Rank Correlation Coefficient between X and Y, $r_S(X, Y)$ as:

$$r_S(X, Y) = r(Rk_X, Rk_Y) = \frac{Cov(Rk_X, Rk_Y)}{\sigma_{Rk_X}\sigma_{Rk_Y}}$$

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

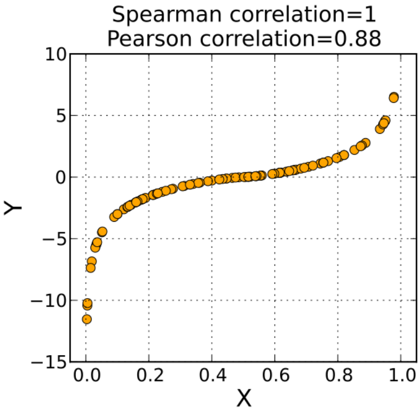
Visualization of r_S

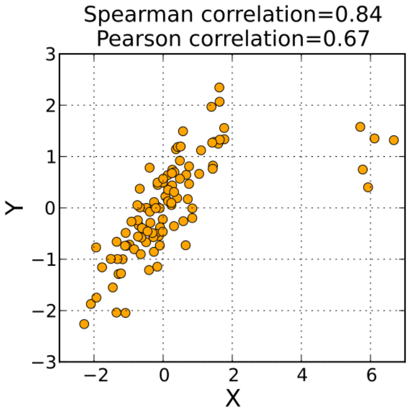
Spearman's Rank Correlation Coefficient is based on monotonicity:



Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

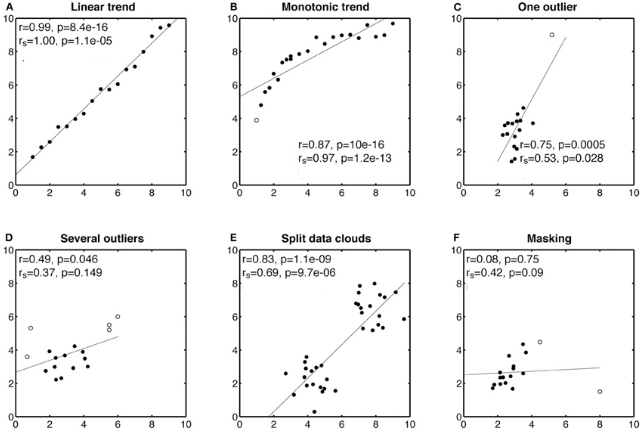
Indeed, the values of r and r_S can be different:





Source with modifications: *https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient*

r and r_S (2/2)



Source with modifications: *https://www.researchgate.net/figure/Examples-of-Pearson-and-Spearman-correlations-In-each-subplot-r-is-Pearson-correlation_fig7_224915794*

Notes about r_S

- If two identical values are assigned their fractional rank
  - So if we have 20, 20, 30, 35, 36, then their ranks should be 1.5 (the average between 1 and 2), 1.5, 3, 4, 5 respectively
- Taking into account that we are dealing with integer ranks, we can simplify the formula as follows if all values are different:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  - where n is the number of observations and each d_i is equal to the difference in rank between X_i and Y_i:

$$d_i = Rk_{X_i} - Rk_{Y_i}$$

Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

Significance of r_S (1/3)

- Being based on ordinals and non assuming anything on the distribution of the underlying populations, the computation of the significance of r_S is based on permutations
- This belong to the family of permutation tests
  - A permutation test (or exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points
- In our case, since I have sequence of ordinals, we can consider all possible pairs of mutual relationships and, based on this, determine if the monotonic relationship that we have obtained is significantly different from a random order

Source with modifications: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

Significance of r_S (2/3)

- Consider as an example the dataset $\{(X_i, Y_i)\} = \{(10, 2), (15, 0), (20, 4), (21, 50)\}$
- Does it have a significant positive correlation?
- We need to assign ranks the elements, leading to $\{(Rk_{X_i}, Rk_{Y_i})\} = \{(1, 2), (2, 1), (3, 3), (4, 4)\}$
- This leads to $r_S = 0.8$
- To compute the significance, I determine the number of times the comparison $Rk_{Y_i} \leq Rk_{Y_j}$ are true when $i < j$
- These are sequences of Bernoulli trials …

Source with modifications: `https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests`

Significance of r_S (3/3)

- It is possible to test for significance also using:

$$w = r\sqrt{\frac{n-2}{1-r^2}}$$

- w follows a t distribution

$$w \sim t$$

Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

Kendall's τ (1/2)

An alternative non parametric correlation coefficient is the Kendall's τ

- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's assume that the two sets X and Y do not contain duplicates
- Let's define
    - a concordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i > Y_j)$ or $(X_i < X_j$ and $Y_i < Y_j)$
    - a discordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i < Y_j)$ or $(X_i > X_j$ and $Y_i < Y_j)$

Source with modifications: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

Kendall's τ (2/2)

- We can define the Kendall's τ as:

$$\tau = \frac{(\#\text{ concordant pairs}) - (\#\text{ discordant pairs})}{n(n-1)/2}$$

Source with modifications: *https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient*

Logistic regression

Outline

- Likelihood function, definition
- Maximum likelihood
- Log likelihood
- Logistic regression

Some slides are take from:
`https://www.cs.ox.ac.uk/people/nando.defreitas/`

Likelihood function

Let $X_1, X_2, ..., X_n$ denote a random sample from p.d.f.

$$X_i \sim f_\theta(x),$$

where θ represents one ore more unknown parameters of the distribution.
The joint p.d.f. of $X_1, X_2, ..., X_n$ is $f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n)$.
If we consider this joint p.d.f. as a function of θ it is called *likelihood function* of a random sample:

$$L_{x_1,x_2,...,x_n}(\theta) = f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n).$$

Let's consider an estimator of θ:

$$\hat{\theta} = u(X_1, X_2, ..., X_n).$$

If for every possible θ $L_{x_1,x_2,...,x_n}(\hat{\theta})$ is at least as great as $L_{x_1,x_2,...,x_n}(\theta)$ then $\hat{\theta}$ is called *maximum likelihood estimator*.
Finally:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(L_{x_1,x_2,...,x_n}(\theta))$$

Maximum loglikelihood (2/2)

Note that, since the likelihood function $L_{x_1,x_2,\ldots,x_n}(\theta)$ and its logarithm $ln(L_{x_1,x_2,\ldots,x_n}(\theta))$, are maximized for the same value θ, either likelihood or its logarithm can be used to find maximum likelihood estimator:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(ln(L_{x_1,x_2,\ldots,x_n}(\theta)))$$

The concept of regression

Regressions can be of multiple types, so far we have analysed the so called OLS regression:

- quadratic cost function of the kind $\sum_i (\hat{y}_i - y_i)^2$
- linear model of the kind $\hat{y} = \mathbf{Ax} + \eta$

What if:

- we use a different objective function, or
- we use a different model

?

Remember that model is called "the mean function" and its inverse "the link function."

Posing a different problem

Let's suppose to have:

- three iid random variables y_i with $i \in [1 \dots 3]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, 1)$
- θ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining y_1 and y_2 and y_3.

In other terms our objective function is the probability of occurrence of y_1 and y_2 and y_3.

We are looking for a maximum likelihood estimator!

Computing the highest probability

Our objective function is therefore:

$$P(y_1, y_2, y_3|\theta) = P(y_1|\theta) \times P(y_2|\theta) \times P(y_3|\theta)$$

We can rewrite this problem as:

$$\max_\theta(\prod_{i=1}^{3} P(y_i|\theta))$$

Note that since θ is a *crisp* value:

$$y_i \sim N(\theta, 1) = \text{a shift of } \theta \text{ of } N(0, 1)$$

Using concrete numbers (1/2)

Let us assume that:

- $y_1 = 1$
- $y_2 = 0.5$
- $y_3 = 1.5$

Remember that $N(\theta, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(y-\theta)^2}{2\sigma^2}}$

Therefore, we want to maximize:

$$\prod_{i=1}^{3} P(y_i|\theta) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(y_i-\theta)^2}{2}} =$$

$$= \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(1-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(0.5-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(1.5-\theta)^2}{2}}$$

Using concrete numbers (2/2)

This is like maximizing:

$$e^{-\frac{(1-\theta)^2}{2}} \times e^{-\frac{(0.5-\theta)^2}{2}} \times e^{-\frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2}{2}-\frac{(0.5-\theta)^2}{2}-\frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2+(0.5-\theta)^2+(1.5-\theta)^2}{2}} = e^{-\frac{3.5-6\theta+3\theta^2}{2}}$$

This is like minimizing $g(\theta) = 3.5 - 6\theta + 3\theta^2$.

$$\frac{dg(\theta)}{d\theta} = \frac{d3.5 - 6\theta + 3\theta^2}{d\theta} = -6 + 6\theta$$

Which becomes 0 for $\theta = 1$

What we have discovered

Our solution is therefore $\theta = 1$ and the desired pdf is $N(1, 1)$. But ...

$$mean(1, 0, 5, 1.5) = 1$$

We can try to generalize it...

Generalizing ...

Let's suppose to have:

- n iid random variables y_i with $i \in [1 \dots n]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, \sigma)$
- θ and σ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining $(\forall i)\ y_i$.

In other terms our objective is to maximize the probability of occurrence of all y_i, that is a maximum likelihood estimation.

Typically, we would perform a least square estimation, and we know that optimal least square estimator is the Gaussian centered in the average of the points, with their standard deviation.

Maximum likelihood estimator (again)

Let' look for a maximum likelihood estimator!

$$\max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(y_i-\theta)^2}{2\sigma^2}}) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{i=1}^{n} e^{-\dfrac{(y_i-\theta)^2}{2\sigma^2}}\right) = \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}\dfrac{(y_i-\theta)^2}{2\sigma^2}}\right) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\right)$$

At this point we can take the log of the expression, knowing that the log function is differentiable and monotonically increasing on all $\mathbb{R}$.

Computing the ml estimator

$$log\Big((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\Big) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) + log(e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2$$

Taking the partial derivative over θ we obtain:

$$\frac{\partial\Big(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\Big)}{\partial\theta} =$$

$$= -\frac{\partial\left(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\theta} = -\frac{1}{\sigma^2} \times \left(\sum_{i=1}^{n}(y_i - \theta)\right)$$

And equating it to 0:

$$-\frac{1}{\sigma^2} \times \left(\sum_{i=1}^{n}(y_i - \theta)\right) = 0 \Rightarrow \sum_{i=1}^{n} y_i = n \times \theta \Rightarrow \theta = \frac{\sum_{i=1}^{n} y_i}{n}$$

Oh! θ is the average of the observed y_i!

Computing the ml estimator - σ (1/2)

$$\frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= \frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}})\right)}{\partial\sigma} - \frac{\partial\left(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2$$

And equating it to 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2 = 0 \Rightarrow \left(\sum_{i=1}^{n}(y_i - \theta)^2\right) \times \frac{1}{\sigma^3} = \frac{n}{\sigma}$$

Computing the ml estimator - σ (2/2)

Assuming $\sigma \neq 0$:

$$\Rightarrow \left(\sum_{i=1}^{n} (y_i - \theta)^2 \right) = n \times \sigma^2 \Rightarrow$$

But we know $\theta = \overline{y_i}$, therefore:

$$\Rightarrow \sigma^2 = \frac{1}{n} \times \left(\sum_{i=1}^{n} (y_i - \overline{y_i})^2 \right)$$

Oh! σ is the standard deviation of the observed y_i!

What we have found

We have determined that the maximum likelihood estimator for a sequence of points assumed to be distributed normally is formed by a normal distribution with:

- average equal to the average of the sample,
- standard deviation equal to the standard derivation of the sample.

This coincides with the best quadratic estimator!

We now move forward considering the maximum likelihood estimator for a regression line, meaning, what happens if now we want to model an interdependencies using as objective function the maximum likelihood.

Let's suppose to have:

- $n \times m$ values $x_{i,j}$ with $i \in [1 \dots n]$, $j \in [1 \dots m]$ represented in short by a matrix $\boldsymbol{X}$ or a vector $\mathbf{x_i}$, $n > m$ *(why?)*
- n iid random variables y_i with $i \in [1 \dots n]$ represented in short by a vector $\boldsymbol{y}$
- a linear relationships $\boldsymbol{\theta}$ between $\boldsymbol{X}$ and $\boldsymbol{y}$, *that is, we use the usual link / mean functions*
- each y_i distributed normally with mean $\boldsymbol{x_i^T \theta}$ and standard deviation σ (the same σ for all y_i), that is
- $(\forall i)\ y_i \sim N(\boldsymbol{x_i^T \theta}, \sigma)$
- $\boldsymbol{\theta}$ and σ to be determined.

We want to determine the value of $\boldsymbol{\theta}$ and σ that maximizes the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{\theta,\sigma}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta},\sigma)) = \max_{\boldsymbol{\theta},\sigma}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i},\boldsymbol{\theta},\sigma))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Computing the optimal θ (1/3)

We can express for simplicity our equation in vectorial form:

$$\max_{\sigma,\boldsymbol{\theta}} \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right)$$

As mentioned, this is equivalent to maximizing the log:

$$\max_{\sigma,\boldsymbol{\theta}} \left(log \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right) \right)$$

Which becomes:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) + log \left(e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}} \right) \right)$$

Computing the optimal θ (2/3)

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta})}{2\sigma^2}\right)$$

And now we take the partial derivative over $\boldsymbol{\theta}$:

$$\frac{\partial \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta})}{2\sigma^2}\right)}{\partial \boldsymbol{\theta}} =$$

$$= -\frac{1}{2\sigma^2}\frac{\partial\left((\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta})\right)}{\partial\boldsymbol{\theta}} = -\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X\theta})$$

And equating it to 0 we obtain:

$$-\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X\theta}) = 0 \quad \Rightarrow \quad \boldsymbol{y} = \boldsymbol{X\theta}$$

If $\boldsymbol{X}$ were square, then the solution would be:

$$\boldsymbol{\theta} = \boldsymbol{X}^{-1}\boldsymbol{y}$$

But, as we said, $n > m$, therefore the solution is given by:

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

What a surprise, isn't it?

Computing the optimal σ

Starting from:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log \left(\frac{1}{\sqrt{2\pi}} \right) + n \times log \left(\frac{1}{\sigma} \right) - \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T (\boldsymbol{y} - \boldsymbol{X\theta})}{2\sigma^2} \right)$$

And now we take the partial derivative over σ:

$$\frac{\partial \left(n \times log \left(\frac{1}{\sqrt{2\pi}} \right) + n \times log \left(\frac{1}{\sigma} \right) - \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T (\boldsymbol{y} - \boldsymbol{X\theta})}{2\sigma^2} \right)}{\partial \sigma} =$$

$$= -\frac{n}{\sigma} + \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T (\boldsymbol{y} - \boldsymbol{X\theta})}{\sigma^3}$$

And equating it to 0, assuming as usual $\sigma \neq 0$ we obtain:

$$\sigma^2 = \frac{(\boldsymbol{y} - \boldsymbol{X\theta})^T (\boldsymbol{y} - \boldsymbol{X\theta})}{n}$$

Maximum likelihood estimator - properties

Claim 1: The maximum likelihood estimator of a Gaussian distribution over a set of points coincides with the OLS estimator.
Proof: See above.
QED

Claim 2: The maximum likelihood linear regression coincides with the OLS linear regression.
Proof: See above.
QED

Bernoulli and maximum likelihood

The pdf of a Bernulli distribution can be represented in terms of conditional probability as:

$$P(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

where clearly x can only be 0 or 1.

We can now introduce the concept of entropy, already hinted in class. Entropy represents the level of uncertainty of a variable.

Entropy (and Bernulli and ml)

Definition (Entropy): Given a random vectorial variable x of n components and a parameter θ, we define entropy of x, $H(x)$ as:

$$H(x) = \sum_{i=1}^{n} p(x_i|\theta) \times log(p(x_i|\theta))$$

We notice that for a Bernulli distribution:

$$H(x) = (1-\theta)log(1-\theta) + \theta log(\theta)$$

Indeed, as θ tends to 0 or to 1 the uncertainty tends to 0, since the likely value of x tend to be 0 or 1 respectively.

From B&B plus ml to LR

We are now ready to move to study a radically different form or regression, the so-called logistic regression.

Our goal is to have a regression that not only represents a relationship between two variables, but is also possible to capture a prediction of probability.

However, the value of a probability is from 0 to 1, so we need a “good” function that can translate any value in such range.

We use often as such function the so-called “sigmoid function.” To introduce the sigmoid we start with the definition of a “logistic function.”

Definition (Logistic function): Given $L, x_0 \in \mathbb{R}$, $k \in \mathbb{R}^+$ a logistic function $f(x)$ is defined as:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Properties (of the logistic function:)

- the domain is all $\mathbb{R}$
- the range is $[0 \ldots L]$ if L is positive and $[L \ldots 0]$ if L is negative
- $f(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $f(x)$ is symmetric over x_0
- k is the rate of growth of $f(x)$ and for $k \to +\infty$ $f(x)$ tends to become the step function in x_0

Sigmoid

Definition (Sigmoid): Given $k \in \mathbb{R}^+$, a sigmoid function $sigm(x)$ is defined as a logistic function with $L = 1$ and $x_0 = 0$:

$$sigm(x) = \frac{1}{1 + e^{-kx}}$$

Properties (of the sigmoid function):

- the domain is all $\mathbb{R}$
- the range is $[0 \dots 1]$
- $sigm(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $sigm(x)$ is symmetric over 0
- k is the rate of growth of $sigm(x)$ and for $k \to +\infty$ $sigm(x)$ tends to become the step function

Toward a logistic regression (1/2)

Suppose that we want to determine if a given event is going to happen based on a series of n predictors $x_1 \ldots x_n$. We can model the probability of occurrence of the event with a random variable y.

It is as if we have a sequence of flipping of coins each with different values of the possible variables that affect the result, for instance the intensity of the flipping, the temperature, the wind, etc.

Based on such set we want to predict what will be the result of the next flipping, given a set of values assigned to the covariates.

Our question is what is:

$$\mathrm{P}(\text{Head} \mid \text{strong toss, strong wind, 60 degrees})$$

?

Toward a logistic regression (2/2)

Let's try to build a regression line.

As we mentioned, any time we compute a regression we need to determine:

- the function to use as a model, and in this case a linear function would not be suitable, since probabilities range from 0 to 1, for this reason we select a sigmoid function;
- the objective function, and in this case the least square would be inappropriate because it is not a proper metrics space, so we opt for maximizing the conditional probability, that is, we aim at a maximum likelihood estimation.

Logistic regression - HPs

Let

- (y_i, x_i) be a collection of pairs with:
  - $i \in [1 \ldots n]$
  - $y_i \in \{0, 1\}$
  - $x_i \in \mathbb{R}^m$
  - $n > m$
- assume that the y_i are iid random variables
- consider as target mean function the sigmoid
- consider as optimality criteria the maximum likelihood

Logistic regression - goals

We want to determine the values of the parameters that maximize the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{Parameters}(P(\mathbf{y}|\boldsymbol{X}, Parameters) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i}, Parameters))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Given our link/mean:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\theta})) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|sigm(\boldsymbol{x_i}^T\boldsymbol{\theta}))$$

Logistic regression - structure

Since the pdf of a Bernulli distribution is:

$$P(z|k) = k^z(1-k)^{(1-z)}$$

For us the probability k of each event is "approximated" by the sigmoid function (our mean function):

$$\boldsymbol{k} = \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}$$

And this lead us to

$$P(y_i|\boldsymbol{x_i},\boldsymbol{\theta}) = \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{y_i} \times \left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{1-y_i}$$

Logistic regression - the problem

Our problem has therefore the form of:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta})) = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{y_i} \times \left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{1-y_i}$$

It is like finding an n-dimensional hyperplane dividing the n-dimensional hyperspace in 2 parts, those leading to y being 0 and those leading to y being 1.

Logistic regression - solution (1/3)

Since the log function is continuous, differentiable and monotonically increasing in all $\mathbb{R}^+$, our problem is equivalent to:

$$\max_{\boldsymbol{\theta}} \left(log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{y_i} \times \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{1-y_i} \right) \right)$$

And, given the property of logs, this is like maximizing:

$$log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{y_i} \right) + log \left(\prod_{i=1}^{n} \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{1-y_i} \right) =$$

$$= \sum_{i=1}^{n} log \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{y_i} + \sum_{i=1}^{n} log \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{1-y_i} =$$

$$= \sum_{i=1}^{n} y_i \times log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) + \sum_{i=1}^{n}(1-y_i) \times log\left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = \cdots$$

A bit of logarithms...

$$log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log(1) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = -log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

$$log\left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}-1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) =$$

$$= log\left(e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = \boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) =$$

Logistic regression - solution (3/3)

$$= -\sum_{i=1}^{n} y_i \times log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) + \sum_{i=1}^{n} (1 - y_i) \times \left(\boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)\right) =$$

For simplicity let w_i be $log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$.

$$= -\sum_{i=1}^{n} y_i \times w_i + \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} + \sum_{i=1}^{n} y_i \times w_i =$$

$$= \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} =$$

$$= \sum_{i=1}^{n} (1 + y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i$$

Logistic regression - comments

Let $f(\theta)$ be:

$$\sum_{i=1}^{n}(1+y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

Claim: $f(\theta)$ is convex.

Proof: Omitted

Consequence: Optimization algorithms can easily find the maximum.