# Statistica appunti

## statistica descrittiva & statistica inferenziale

Per studiare un fenomeno statistico si può decidere di interpellare l'intera popolazione, oppure di analizzare solo una parte della popolazione, un campione, e di estendere poi i risultati ottenuti all'intera popolazione. Nel primo caso si parla di **statistica descrittiva**, mentre nel secondo caso di **statistica inferenziale**.

## Approccio Frequentista e Bayesiano

Nello studio di un fenomeno in generale si cerca di formalizzare la realtà nella quale ci si trova ad operare tramite un modello che permetta la piena conoscenza del fenomeno stesso. Per fare ciò si ricorre in ambito statistico all'**inferenza** cioè ad una particolare procedura che permetta di ottenere, da dati raccolti tramite un campione, informazioni generalizzabili alla popolazione dalla quale questi dati sono stati estratti. L'inferenza, però, dipende dal campione estratto, dal modello probabilistico sottostante che genera tale campione ma anche dall'approccio che si sceglie per ricondurre alla popolazione incognita ciò che si viene a conoscere dal campione noto.

Si distinguono infatti due importanti approcci quello **classico frequentista** e quello **bayesiano** ( fondato sull'uso del risultato del teorema di Bayes ai fini dell'inferenza statistica), entrambi caratterizzati da procedure inferenziali che rispettano loro specifici criteri sottostanti.

In sintesi quello che caratterizza l'**approccio classico** è che la probabilità è vista come un *valore oggettivo*, i parametri sono fissi e incogniti e le procedure inferenziali sono basate su un campionamento ripetuto nelle stesse condizioni. Quello che invece sottostà all'**approccio bayesiano** è la probabilità intesa come *valore soggettivo*, parametri come variabili casuali e la procedura inferenziale è basata sulla distribuzione di probabilità dei parametri osservando il campione estratto e avendo a disposizione ulteriori informazioni.

Si aggiunga il fatto che l'approccio bayesiano è in grado di utilizzare informazioni già in possesso, modificando la probabilità a priori e ottenendo così delle probabilità a posteriori diverse.

# Inferenza Statistica

Si definisce **inferenza statistica** il procedimento mediante il quale, dall'analisi dei dati osservati sul campione, si traggono informazioni relative all'intera popolazione. Si possono distinguere due aspetti principali dell'inferenza statistica:

- la stima campionaria, quando dallo studio dei parametri del campione, quali la media, la varianza, etc si può stimare il corrispondente parametro della popolazione che non è noto;
- la verifica delle ipotesi, quando dall'esame del campione si vuole decidere se un'ipotesi fatta su una data popolazione è accettabile o rifiutabile; la decisione è presa ad un dato livello di probabilità di commettere un errore nell'accettare l'ipotesi quando questa è falsa o nel rifiutarla quando questa è vera.

# Sampling

Sampling is **a process used in statistical analysis in which a predetermined number of observations are taken from a larger population**.

$z \sim D$ sampling $z$ according to $D$

# Sample

A sample is a subset of the entire population. In inferential statistics, the goal is to use the sample to learn about the population. Consequently, the sample typically is selected in a manner that allows it to be an unbiased representation of the entire population. Drawing a random sample is a common method for achieving this unbiased representation. In a simple random sample, each member of the population has an equal probability of being included in the sample. However, different modifications of simple random samples can be used to meet specific research needs.

# Inferential statistics

Inferential statistics use a random sample to draw conclusions about the population. Typically, it is not practical to obtain data from every member of a population. Instead, we collect a random sample from a small proportion of the population. From the sample, statistical procedures can infer the likely properties of the population.

For example, it is impractical to measure the height of every adult woman, but you can measure the heights of a random sample and use that information to make generalizations about the heights of all women. For example, a confidence interval provides a range that the population mean height is likely to fall in.

# Parameter

Parameters are the unknown values of an entire population, such as the mean and standard deviation. Samples can estimate population parameters but their exact values are usually unknowable.

Parameters are also the constant values that appear in probability functions. These parameters define the shape of probability distributions. Parameters are typically denoted using Greek symbols to distinguish them from sample statistics.

For example, the parameters of the normal distribution are $\mu$ ( mu = population mean) and $\sigma$ (sigma = population standard deviation).

# Estimator

A sample statistic that estimates a population parameter. The value of the estimator is referred to as a point estimate. There are several different types of estimators.

- If the expected value of the estimator equals the population parameter, the estimator is an **unbiased estimator**.
- If the expected value of the estimator does not equal the population parameter, it is a **biased estimator**.
- If the expected value of the estimator approaches the population value as the sample size increases, it is an **asymptotically unbiased estimator**.

# Hypothesis tests

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. These two statements are called the null hypothesis and the alternative hypothesis.

Hypothesis tests are not 100% accurate because they use a random sample to draw conclusions about entire populations. When you perform a hypothesis test, there are two types of errors related to drawing an incorrect conclusion.

- Type I error: The rejects a null hypothesis that is true. You can think of this as a false positive.
- Type II error: The test fails to reject a null hypothesis that is false. You can think of this as a false negative.

A test result is statistically significant when the sample statistic is unusual enough relative to the null hypothesis that you can reject the null hypothesis for the entire population. "Unusual enough" in a hypothesis test is defined by how unlikely the effect observed in your sample is if the null hypothesis is true.

If your sample data provide sufficient evidence, you can reject the null hypothesis for the entire population. Your data favor the alternative hypothesis.

## Null hypothesis

The null hypothesis is one of two mutually exclusive hypotheses in a hypothesis test. The null hypothesis states that a population parameter equals a specified value. If your sample contains sufficient evidence, you can reject the null hypothesis and conclude that the effect is statistically significant. The null hypothesis is often displayed as H0.

In every experiment, there is an effect or difference between groups that the researchers are testing. It could be the effectiveness of a new drug, building material, or other intervention that has benefits. Typically, the null hypothesis states that the true effect size equals zero—that there is no difference between the groups. Therefore, if you can reject the null hypothesis, you can favor the alternative hypothesis, which states that the effect exists (doesn't equal zero) at the population level.

## Alternative hypothesis

The alternative hypothesis is one of two mutually exclusive hypotheses in a hypothesis test. The alternative hypothesis states that a population parameter does not equal a specified value. Typically, this value is the null hypothesis value associated with no effect, such as zero. If your sample contains sufficient evidence, you can reject the null hypothesis and favor the alternative hypothesis. The alternative hypothesis is often denoted as H1 or HA.

If you are performing a two-tailed hypothesis test, the alternative hypothesis states that the population parameter does not equal the null hypothesis value. For example, when the alternative hypothesis is HA: $\mu \neq 0$, the test can detect differences both greater than and less than the null value.

A one-tailed alternative hypothesis can test for a difference only in one direction. For example, HA: $\mu > 0$ can only test for differences that are greater than zero.

# Standardized variable

A standardized variable (sometimes called a z-score or a standard score) is **a variable that has been rescaled to have a mean of zero and a standard deviation of one**.

# Critical Value

A critical value is the value of the test statistic which **defines the upper and lower bounds of a** confidence interval, or which defines the threshold of statistical significance in a statistical test.

# Hypothesis testing

A representative sample is one which is drawn without bias from the population of interest. Representativeness is not so much about the sample size but depending on the right composition of the sample.

we might ask: that sample is well done? if you want to be completely sure about that, the technique for checking this representativeness is called **hypothesis testing**.

# Central Limit Theorem

The **Central Limit Theorem** establishes that if we gather together a set of independent random variables, they should follow a normal distribution even if the original variables themselves are not normally distributed.

# Differenza tra media e valore atteso

Se hai un esperimento statistico:  Un approccio pratico produce una distribuzione di frequenza e un valore medio; un approccio teorico produce una distribuzione di probabilità e un valore atteso.  Se lo spazio campione è infinitamente grande, il valore medio dovrebbe avvicinarsi al valore atteso.  Tuttavia, per la maggior parte del tempo, questi due termini sono usati in modo intercambiabile.

# Qual è la differenza tra il metodo analitico e il metodo numerico?

Il metodo **analitico è** un metodo diretto perché si trattano direttamente le equazioni, ma non sempre tutte le equazioni sono integrabili.

Con il metodo **numerico** si approssimano le derivate **e** gli integrali con dei piccoli rapporti incrementali **e** trapezi **e** risolvi matematicamente tramite un computer.

# Stima puntuale e stima intervallare

Una **STIMA PUNTUALE** è il risultato di un procedimento che, attraverso le informazioni tratte dal campione osservato, genera un singolo valore numerico usato per stimare il corrispondente parametro della popolazione. Una **STIMA INTERVALLARE** è il risultato di un procedimento che, attraverso le informazione tratte dal campione osservato, genera un intervallo di valori che, con un dato grado di fiducia, conterrà il parametro da stimare.

In molte situazioni è preferibile una **stima intervallare** (cioè è preferibile indicare come stima del parametro un intervallo al posto di un singolo punto sull'asse dei valori, ovvero la **stima puntuale**) che esprima anche l'errore associato alla stima (precisione).

# Indici di posizione

## Mediana

La **mediana** $M$ di un insieme di nn dati ordinati in modo crescente, è il valore centrale dei dati se il numero di dati è dispari, o la media aritmetica dei due valori centrali se il numero dei dati è pari.

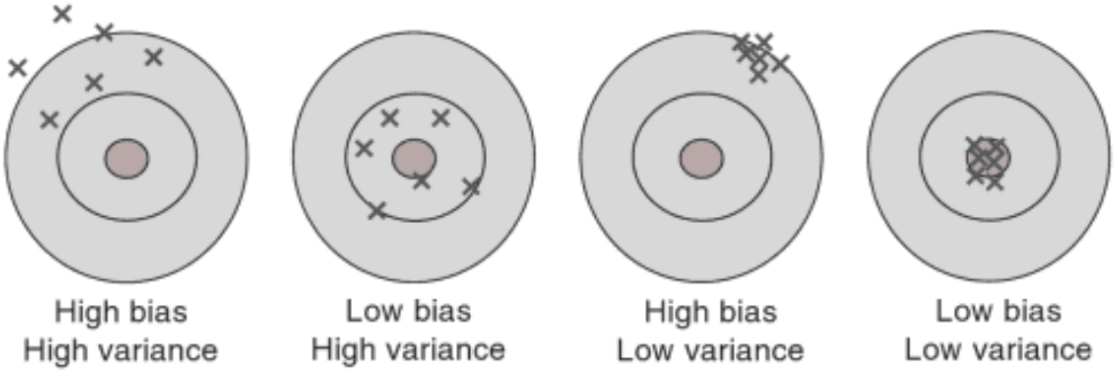
## Quantili : quartili e percentili

Quantili e' la "categoria". I **quantili** sono utilizzati in statistica per frazionare in N parti uguali un insieme di dati.

Oltre alla mediana, che divide a metà un insieme di dati ordinati, vengono usati anche altri indici di posizione che dividono le distribuzioni in determinate percentuali detti **quartili**, **quantili** e **percentili**. Questi sono detti **indici di posizione non centrale** e vengono usati soprattutto per ampi insiemi di dati.

- I **quartili** sono un caso particolare dei quantili e, come dice la parola stessa, si ottengono dividendo l'insieme di dati ordinati in 4 parti uguali ed esattamente:
- I **percentili**, invece, dividono la distribuzione in 100 parti.

# Differenza Bias e varianza

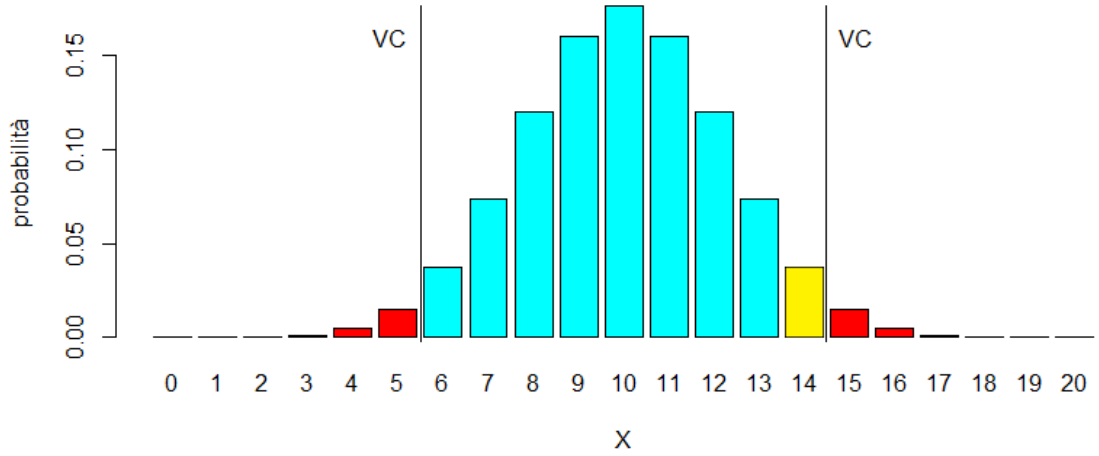



# Test d'ipotesi

prima di calcolare il valore sperimentale della statistica test, si consiglia in genere di scegliere il livello di significatività, indicato convenzionalmente col simbolo $\alpha$. Il suo impiego è quello di discrimine per il valore $p$: il risultato del test si dice significativo se $p < \alpha$, altrimenti si considera non significativo. $H_0$ è rifiutata se il risultato è significativo.

Maggiore è la fiducia riposta nell'ipotesi nulla, maggiore l'evidenza richiesta per smentirla, e minore deve essere $\alpha$.

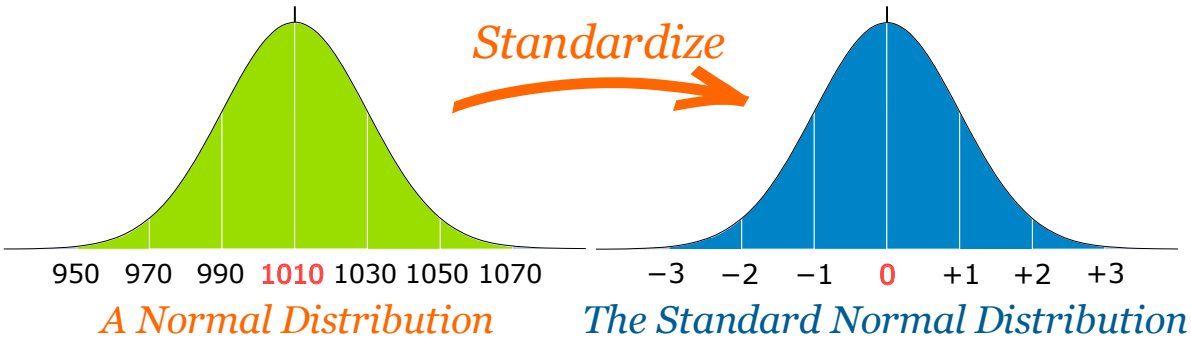
Dato un certo livello di significatività, l'insieme dei valori della statistica test a cui corrisponde un p minore di α si chiama *regione di rifiuto*. Abbiamo poi la cosiddetta *regione di accettazione*. Si chiamano invece *valori critici* i punti che separano le regioni di rifiuto ed accettazione.




Distribuzione della statistica test binomiale X dell'esempio della moneta; la regione di rifiuto è evidenziata in rosso e sono segnalati i valori critici. Nel caso di test a una coda destro, la coda sinistra esce dalla regione di rifiuto e il punto 14, evidenziato in giallo, vi entra.

# Standardizing

$$Z = \frac{X - \mu}{\sigma}$$




## Joint, Marginal and Conditional distributions

Given random variables $X, Y, \ldots$, that are defined on a probability space, the **joint probability distribution** for $X, Y, \ldots$ is a probability distribution that gives the probability that each of $X, Y, \ldots$ falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a bivariate distribution, but the concept generalizes to any number of random variables, giving a multivariate distribution.

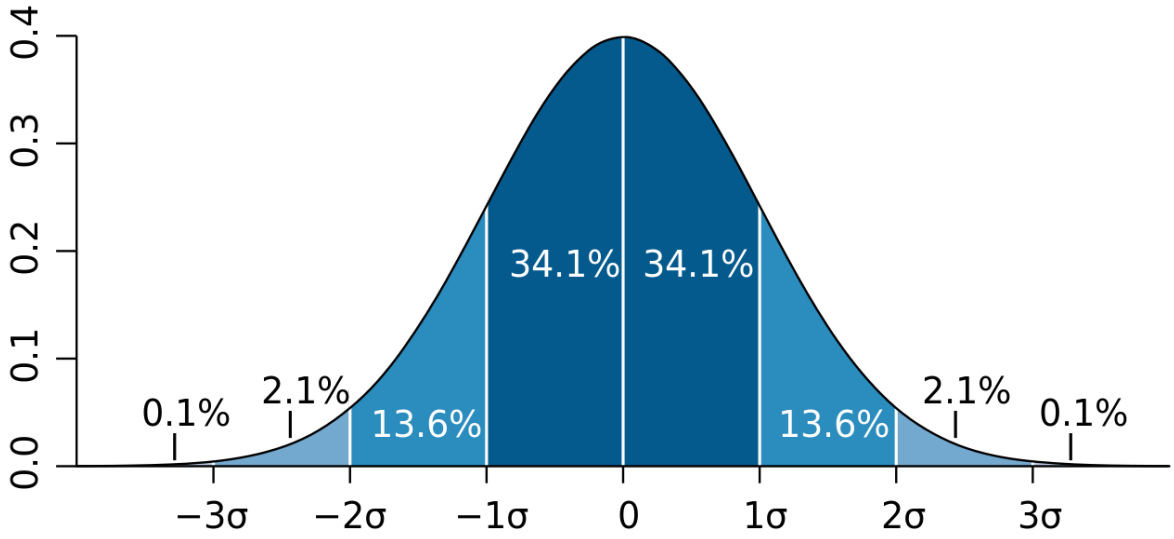
The **marginal distribution** of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. This contrasts with a conditional distribution, which gives the probabilities contingent upon the values of the other variables.

*Attenzione*: the concept of joint distribution can be generalized to **vectors** → multivariate distribution

## Univariate, Bivariate and Multivariate

All three analyses are very important in any analytical project. However, most of the analysis that we end up doing are multivariate due to complexity of the world that we are living in.

- **Univariate** - One variable is analyzed at a time. Objective is to describe the variable. Example- How many students are graduating with "Analytics" degree?
- **Bivariate** - Two variables are analyzed together for any possible association or empirical relationship. Example- What is the correlation between "Gender" and graduation with "Analytics" degree?
- **Multivariate** - More than two variables are analyzed together for any possible association or interactions. Example – What is correlation between "Gender", "Country of Residence" and graduation with " Analytics" degree? Any statistical modeling exercise such as Regression, Decision Tree, SVM, Clustering are multivariate in nature.



Univariate Gaussian distribution

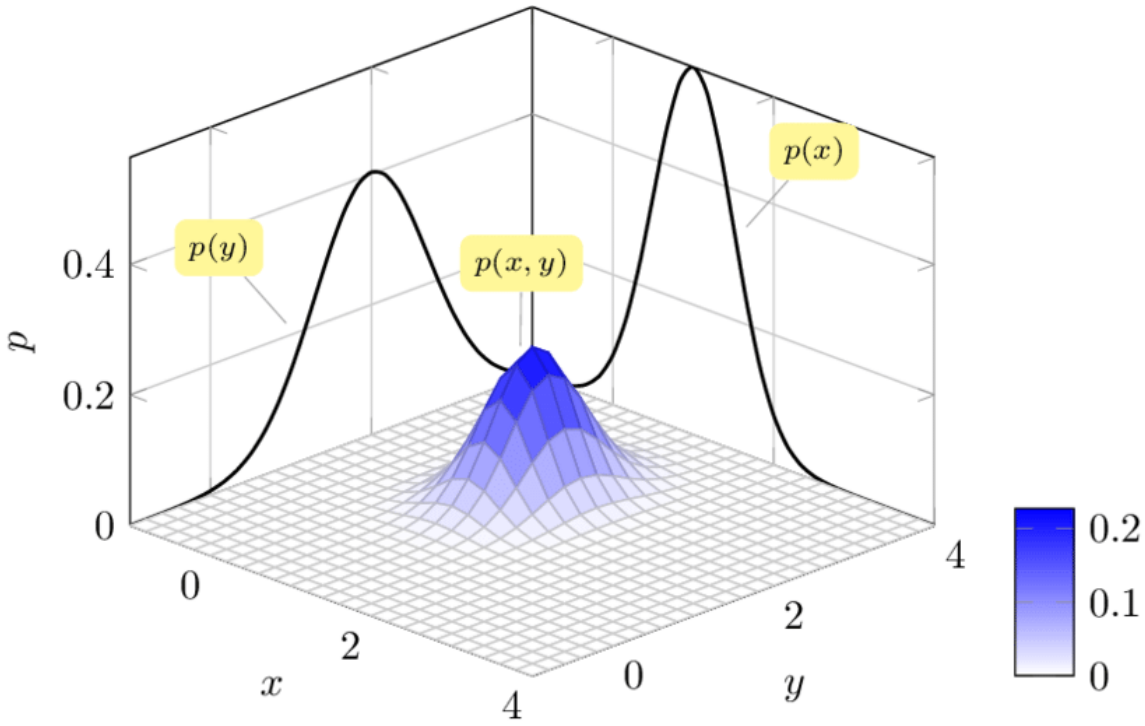



Illustration of a Bivariate Gaussian distribution. The marginal and joint probability

## Moments (of a statistical distribution)

The shape of any distribution can be described by its various 'moments'. The first four are:

1. The **Mean**, which indicates the central tendency of a distribution.

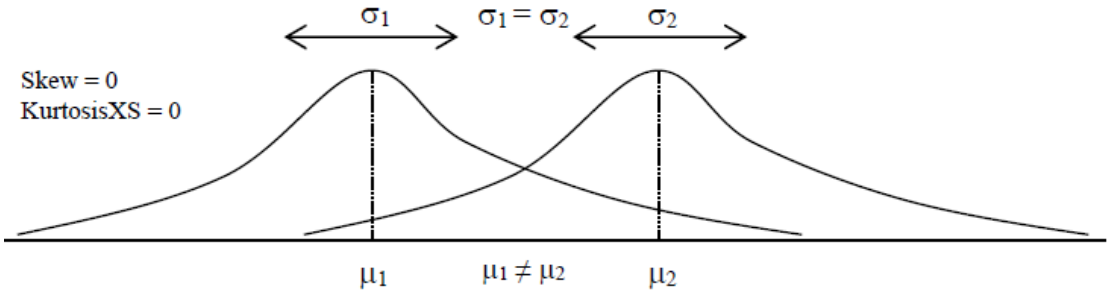
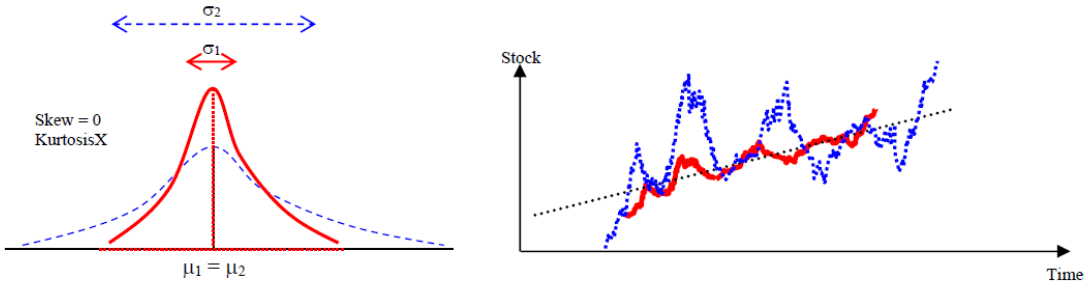



**Figure 1.** Visualizing the First Moment

2. The second moment is the **Variance** (*central moment*), which indicates the width or deviation




**Figures 2 and 3.** Second Moment and Stock Price Fluctuations

3. The third moment is the **Skewness**, which indicates any asymmetric 'leaning' to either left or right.

4. 


5. The fourth moment is the **Kurtosis**, which indicates the degree of central 'peakedness' or, equivalently, the 'fatness' of the outer tails.

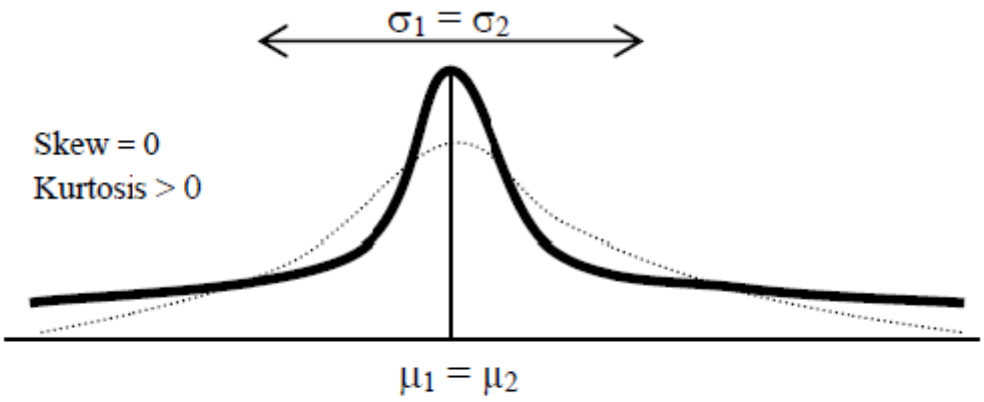



**Figure 6.** The Fourth Moment

## Difference between Raw Moment and Central Moment

The $n-th$ moment about zero of a probability density function $f(x)$ is the expected value $E[X]$ of $X^n$ and is called a **raw moment**. The moments about its mean $\mu$ are called **central moments**; these describe the shape of the function, independently of translation.

$$\mathbb{E}(X^n) = \text{raw moment}$$
$$\mathbb{E}\left[(X - \mathbb{E}(X))^n\right] = \text{central moment}$$

where the 2nd central moments represents the variance.

only equal when $\mathbb{E}(X) = 0$ as with $\mathcal{N}(0,1)$.

## Binomial Coefficient

Esistono due formule equivalenti

**Homework 8 #3:** Define the generalized binomial coefficients for $\alpha \in \mathbb{C}$ and $n \in \mathbb{N}$,

$$\binom{\alpha}{n} = \begin{cases} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}, & n \neq 0 \\ 1, & if\ n = 0. \end{cases}$$

Prove that the definition above agrees with original definition of the binomial coefficient for $\alpha \in \mathbb{N}$. In particular, prove

$$\binom{\alpha}{n} = \begin{cases} 1, & \text{if } n = 0, \\ \frac{\alpha!}{n!(\alpha-n)!} & \text{if } \alpha \in \mathbb{N} \text{ and } n \leq \alpha, \\ 0 & \text{if } \alpha \in \mathbb{N} \text{ and } n > \alpha. \end{cases}$$

# Why we consider log likelihood instead of Likelihood

1. It is extremely useful for example when you want to calculate the *joint likelihood* for a set of independent and identically distributed points. Assuming that you have your points:

   $$X = \{x_1, x_2, \ldots, x_N\}$$

   The total likelihood is the product of the likelihood for each point, i.e.:

   $$p(X \mid \Theta) = \prod_{i=1}^{N} p(x_i \mid \Theta)$$

   where $\Theta$ are the model parameters: vector of means $\mu$ and covariance matrix $\Sigma$. If you use the log-likelihood you will end up with sum instead of product:

   $$\ln p(X \mid \Theta) = \sum_{i=1}^{N} \ln p(x_i \mid \Theta)$$

2. Also in the case of Gaussian, it allows you to avoid computation of the exponential:

   $$p(x \mid \Theta) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

   Which becomes:

   $$\ln p(x \mid \Theta) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \Sigma) - \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)$$

3. Like you mentioned $\ln x$ is a monotonically increasing function, thus log-likelihoods have the same relations of order as the likelihoods:

   $$p(x \mid \Theta_1) > p(x \mid \Theta_2) \Leftrightarrow \ln p(x \mid \Theta_1) > \ln p(x \mid \Theta_2)$$