Statistics and it's applications

Ali Akbari

Sharif University of Technology
EE department
AIRLab

Outline

Motivations for this talk

- Statistics and related fields, are often used(even without knowing) in most of the research, in which we have to extract some data from a population.
- The population can be anything or of any type, which sometimes make it rather difficult to rely on the extracted data.
- Fortunately, statistics offer a wide variety of tools to analyze our data. These tools make sure that the risk of any mistake or flaw in the data and the way it is treated, is fairly low to the point of insignificance.

What will be covered in this presentation ?

- First we review some the basics that we encounter during this presentation.
- The second part, includes some advanced topics and tools to analyze our data more efficiently.

Sampling basics

- So why we use sampling ? Main reasons may be :
  - We have limited resources.
  - We have limited data available.
  - A destructive testing.
- Limited resources almost always play some part. For example, if we wish to extract some data from patients who suffer from AD, we can not access all of the patients in the world who have this disease.
- Sometimes there is only a small sample available, no matter what cost may be incurred.
- Sampling may involve destructive testing. We can not wait to see how long do patients with AD will survive !

How to sample then ?

Warning !

It is essential to distinguish between bad luck and bad management. A poorly managed sampling may alter the course of a research !

- We must note that the sampling, if done poorly, may result in acceptable results by good chance. It is up to us if we decide to leave things to destiny !
- A good sampling method decrease the chance of bias in the data, in favor of some particular group.

What are measures?

- Statistical measures are a descriptive analysis technique used to summarise the characteristics of a data set.
- This data set can represent the whole population or a sample of it.
- Statistical measures can be classified as measures of central tendency and measures of spread.

- Measures of Central Tendency describe some key characteristics of the data set based on the average or middle values, as they describe the centre of the data.
- The measures of central tendency that we will be looking at are the mean, mode, and median.

The mean

The mean

Assuming that we have n data points of a chosen data set, the estimation of the population mean will be :

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- This is the most common central measure.
- It shows a rather good estimation of where the centre is.
- Note that we used the term *estimation* to emphasise the fact that we are dealing with some chosen points in the main unknown data set.

The Mode

The Mode

This mode defined as the most frequent value in our data set.

- Generally, the mode is not a good measure of central tendency, since it often depends on the arbitrary grouping of the data.
- is also possible to draw a sample where the largest frequency (highest bar in the group) occurs at two (or even more) heights ; this unfortunate ambiguity is left unresolved, and the distribution is "bimodal."

The Median

The Median

This is the $50'$th percentile ; i.e., the value below which half the values in the sample fall.

- Since it splits the observations into two halves, it is sometimes called the middle value.
- In cases where the midpoint values are two (when the number of data points is even), you need to find the average of both middle values.

Introduction

- Assume that we have a constant population, so that it's mean μ and variance σ^2 are constants (though generally unknown). These are called population parameters.
- By contrast, the sample mean $\bar{X}$ and sample variance s^2 are random variables, varying from sample to sample, with a certain probability distribution.

Introduction cont.

- Suppose that we want to know the mean firing rate of cells in a certain part of the brain, during an event.
- After sampling from a certain group of cells(not all of them), we obtain a number which we call *fr*.
- Unless we are very lucky, this number differs from the real mean a little bit. So we define an interval, called the **confidence interval**.

Confidence Interval

The Confidence Interval

Assuming the distribution of the mean is normal, a confidence interval for the mean can be defined as :

$$\mu = \bar{x} \pm z_p \frac{\sigma}{\sqrt{n}}$$

In which p denotes the interval we desire.

95% Confidence Interval

One important confidence interval is the 95% which is defined as :

$$\mu = \bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}}$$

- Now let's see what z_p denotes.
- z_p is the critical value leaving $p \times 100\%$ probability in the upper tail of the standard normal distribution.

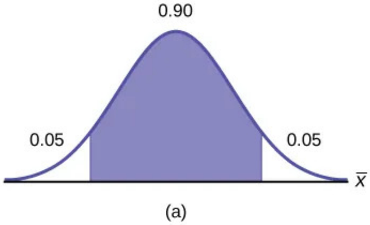


Figure – A demonstration of the above

Estimators

- Now let's get more general. We consider any population parameter θ, and denote an estimator for it by *Hat*θ.
- We would like the random variable $\hat{\theta}$ to vary within only a narrow range around its fixed target θ.

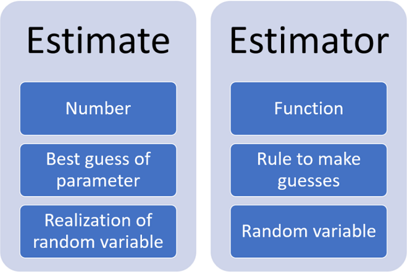


Figure – Difference between estimation and estimator

Estimators Cont.

Bias of an Estimator

Suppose that we have an estimator $\hat{\theta}$. The bias of this estimator is defined as :

$$B[\hat{\theta}] = E[\hat{\theta}] - \theta$$

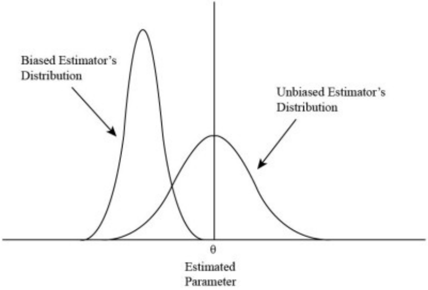


Figure – Unbiased and biased estimators

Biased and Unbiased Estimators

- An example of a biased estimator is :

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  Which on the average underestimate σ^2, the population variance.

- The unbiased version can be defined as :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Efficiency

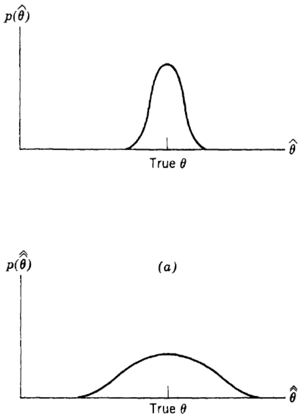
The Efficiency

The Efficiency of 2 estimators compared to each other is defined as :

$$Eff = \frac{Var(\hat{\hat{\theta}})}{Var(\hat{\theta})}$$

- The point is that a good estimator, has relatively lower variance compared to other estimators.

Efficiency


$p(\hat{\theta})$
$\hat{\theta}$
True θ
(a)
$p(\hat{\hat{\theta}})$
$\hat{\hat{\theta}}$
True θ

Estimators

- Now let's look at some of the more popular estimators :
  - Method of Moments (MOM)
  - Maximum Likelihood Estimator (MLE)
  - Maximum a Posteriori (MAP)
- Our goal is to compare them in action and investigate their use cases.
- But first, let's look up their defenitions.

Maximum Likelihood Estimator (MLE)

Maximum Likelihood Estimator

Let $x = (x1, \ldots, xn)$ be *iid* samples from probability mass function $p_X(t|\theta)$ (if X is discrete), or from density $f_X(t|\theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x given θ to be the "probability" of observing x if the true parameter is θ. If X is discrete,

$$L(x|\theta) = \prod_{n}^{i=1} p_X(x_i|\theta)$$

If X is continuous,

$$L(x|\theta) = \prod_{n}^{i=1} f_X(xi|\theta)$$

Moments

Moments

Let X be a random variable and $c \in R$ a scalar. Then The k'th moment of X is :

$$E[X^k]$$

and the k'th moment of X (about c) is :

$$E[(X-c)^k]$$

Usually, we are interested in the first moment of $X := E[X]$, and the second moment of X about $\mu : E[(X-\mu)^k]$

Sample Moments

Sample Moments

Let X be a random variable, and $c \in R$ a scalar. Let $x_1, \ldots, x_n$ be *iid* realizations (samples) from X. The $k'th$ sample moment of X is :

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

The $k'th$ sample moment of X (about c) is :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k$$

For example, the first sample moment is just the sample mean, and the second sample moment about the sample mean is the sample variance.

Method of Moments

Method of Moments

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be *iid* realizations (samples) from probability mass function$p_X(t;\theta)$ (if X is discrete), or from density $f_X(t;\theta)$ (if X is continuous), where θ is a parameter (or vector of parameters).
We then define the method of moments (MoM) estimator $\hat{\theta}_{MoM}$ of $\theta = (1, \ldots, k)$ to be a solution (if it exists) to the k simultaneous equations where, for $j = 1, \ldots, k$, we set the $j'th$ (true) moment equal to the $j'th$ sample moment :

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\ldots$$

$$E[X^k] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Maximum a Posteriori (MAP)

- MAP estimation is quite different from the other estimators that we saw.
- Previously, we assumed that the parameters are unknown yet fixed, which is called the frequentist view.
- Now, we assume that our parameter *theta*, is a $R.V$.

Comparison

- Which do you think is “better”, MLE or MAP ?
- There is no right answer. There are two main schools in statistics : Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed quantity.
- On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a random variable, and we seek the mode - the value with the highest probability or density.

Comparison Cont.

- An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not ?
- If so, what distribution should we put on the parameter space ?
- Anyway, in the long run, the prior "washes out", and the only thing that matters is the likelihood ; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate.
- However, as the number of samples goes to infinity, the MAP and MLE are equal.

Comparison Cont.

- MAP is appropriate for those problems where there is some prior information, e.g. where a meaningful prior can be set to weigh the choice of different distributions and parameters or model parameters. MLE is more appropriate where there is no such prior.
- MAP is much better than MLE If our dataset is small, so make sure to use MAP if you have information about prior probability.
- As a quick review of MoM, we can say that MoM relies on specific equation of the moments, and if we pick up the wrong density, we do totally wrong, while MLE is more resilient, as we in all case minimise the KL divergence.

Intuition

- Assume that you have a coin. To test if it is fair or not, you flip it 100 times.
- Now you set a critical value. If you see more than 60 heads, you reject the assumption of the coin being fair.
- You observe 95 heads and 5 tails.
- Now, we want to see if it is just bad luck, or the coin is really modified.

Hypothesis

- In hypothesis testing you are a judge. You judge the hypothesis based on numerical values.
- You should remember that you never accept or prove any hypothesis.
- You just reject the hypothesis, or fail to reject it !

Hypothesis Cont.

- The first thing that we have to prepare for hypothesis testing is a good hypothesis.
- The hypothesis that we wish to test, is called the null hypothesis or H_0.
- The alternate of the hypothesis, is called H_A.
- Now we have to determine what level of significance fits our needs. A popular choice is $\alpha = .05$ which shows we can be about 95% confident about our results.

Hypothesis Cont.

- Now we calculate a numerical value, called the $p - value$.
- A little intuition about this number is that it shows how lucky we were to achieve the results.
- The α parameter that we introduced earlier, shows how much luck we tolerate!
- If $p - value < \alpha$, we reject the null hypothesis. Otherwise, we state that we failed to reject the null hypothesis.

Hypothesis Cont.

- The null hypothesis, always comes with an equality, unlike the alternate.
- By equality, we mean :

$$\{=, \leq, \geq\}$$

- On the other hand, the alternate hypothesis comes with an non-equality :

$$\{\neq, <, >\}$$

Hypothesis Cont.

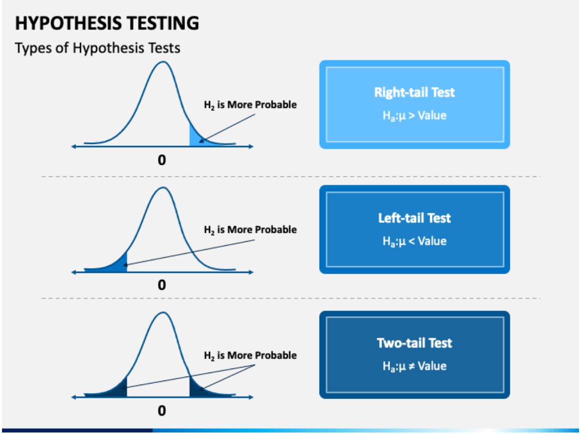


Figure – α

Hypothesis Cont.

- Now we investigate the errors that we may encounter.

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

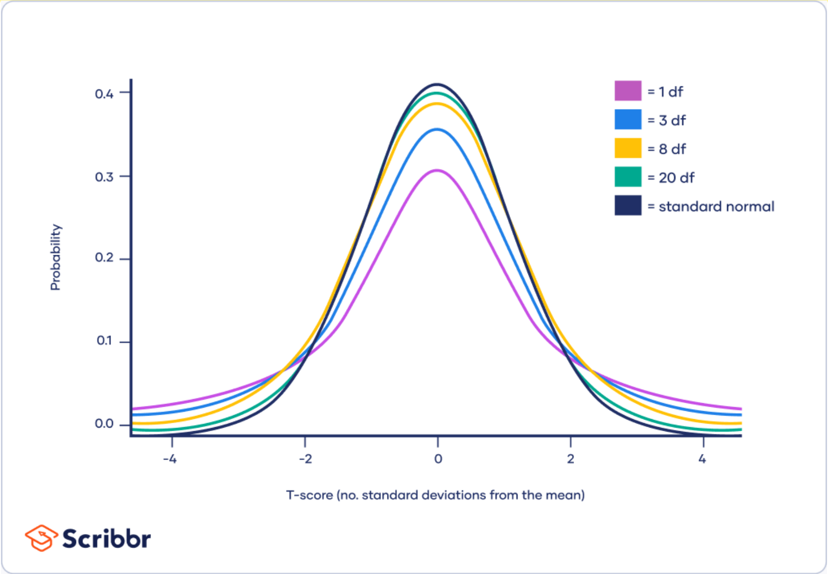
Scribbr

Figure – Types of errors in hypothesis testing

Hypothesis Cont.

- Another detail is that in action, we don't know the true variance of the population, σ^2
- In this situation, we should use Student's t-distribution instead of Normal distribution.
- Although, if the number of samples is sufficiently high($n \geq 30$),Student's t-distribution converges to Normal distribution.

Hypothesis Cont.



Hypothesis Testing in Action

- Now let's see how it is used.
- Suppose we have some patients with hardship in reading some paragraph due to an unknown(but same) disease.
- It is previously determined that in a 100-word paragraph, the patients, on average, understand 67 words.
- Recently, another study showed that a group of 16 patients understand 71 words in the same paragraph with $\sigma = 3$.
- We want to know, with 95% certainty, that the results do not contradict each other.

Hypothesis Testing in Action Cont.

- With previous assumptions, we set the parameters :
  - $H_0 : \mu = 67$
  - $H_1 : \mu \geq 67$
  - $\alpha = 0.05$
- Now we have :

$$\frac{71 - 67}{\frac{3}{\sqrt{16}}} = 2.33$$

- Now we have to calculate :

$$P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq 2.33) = 0.01833$$

Hypothesis Testing in Action Cont.

- We ended up with $p - value = 0.018$, which is less than our alpha level.
- So indeed, the mean is not 67.

Main Idea

- Suppose, we have n different groups.
- We wish to test, if these groups are in fact different.
- Let's see this as a hypothesis :
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$
  - $H_A : \mu_i \neq \mu_j$ for some i, j
- To develop a plausible test of this hypothesis we first require a numerical measure of the degree to which the sample means differ.

Some Measures

- $$s_{\bar{X}}^2 = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X} - \bar{\bar{X}}^2)$$
- $$s_p^2 = \frac{1}{r}\sum_{i=1}^{r}(s_i^2)$$
- Now the $F - score$ is :

$$F = \frac{ns_{\bar{X}}^2}{s_p^2}$$

Some Measures Cont.

- Now let's see what these measures really tell us.
- We are trying to see if the groups are really different.
- We note that if the $F - score$ is near 1, then we fail to reject the null hypothesis.
- If $F - score$ is relativly high, then we can reject the null hypothesis.