# Contents

# 1 Statistical Inference

What is Statistical Inference?

- Hyptothesis testing

- Confidence Intervals

- Sample size estimation.

## 1.1 Hypothesis testing: introduction

The objective of hypothesis testing is to access the validity of a claim against a counterclaim using sample data

- The claim to be proved is the alternative hypothesis($H_1$).

- The competing claim is called the null hypothesis($H_0$).

- One begins by assuming that $H_0$ is true.

If the data fails to contradict $H_0$ beyond a reasonable doubt, then $H_0$ is not rejected. However, failing to reject $H_0$ does not mean that we accept it as true. It simply means that $H_0$ cannot be ruled out as a possible explanation for the observed data. A proof by insufficient data is not a proof at all.

> The process by which we use data to answer questions about parameters is very similar to how juries evaluate evidence about a defendant. from Geoffrey Vining, Statistical Methods for Engineers, Duxbury, 1st edition, 1998.

# 2 Confidence Intervals

# 3 Hypothesis Testing

Hypothesis testing is a common practice in science that involves conducting tests and experiments to see if a proposed explanation for an observed phenomenon works in practice. A hypothesis is a tentative explanation for some kind of observed phenomenon, and is an important part of the scientific method. The scientific method is a set of steps that is commonly employed by those in scientific fields to give scientific explanations for various phenomena.

Any tentative explanation can be referred to as a hypothesis if it can be submitted to hypothesis testing. There are, however, a set of guidelines for an explanation to be considered a true scientific hypothesis. The first major point is testability; a scientific hypothesis must be able to proceed to the stage of hypothesis testing to be considered a scientifically legitimate hypothesis. It is generally suggested that a hypothesis be relatively simple, though this is not always possible. Hypotheses must also be able to explain the phenomena under any set of conditions; if a hypothesis can only explain a phenomenon in one set of conditions, it is generally considered unacceptable.

Hypotheses are generally considered useful only if they are likely to improve on the current body of knowledge on a subject and pave the way for greater knowledge to be acquired in the future. Also, a hypothesis is generally not acknowledged if it defies other commonly recognized knowledge. If a hypothesis meets all of these requirements, it will typically proceed to the hypothesis testing phase.

In hypothesis testing, the testers seek to discover evidence that either validates or disproves a given hypothesis. Usually, this involves a series of experiments being conducted in many different conditions. If the hypothesis does not stand up to the tests in all conditions, something is usually wrong with the hypothesis and a new one must be formed to take the new information into account. The new hypothesis is submitted to the same hypothesis testing. If it passes and is not proven wrong, it can eventually be considered a scientific theory or law, though nothing in science can be proven to be absolutely true.

One common method of hypothesis testing is known as statistical hypothesis testing, and typically deals with large quantities of data. Experiments and tests are conducted and the data is collected. If the data collected shows that it is unlikely that the results occurred by chance, it is considered statistically significant and can be used to support a hypothesis.

# 4    Degrees of Freedom

Degree of freedom (df) is a concept most used in statistics and physics. In both cases it tends to define limits of a system and position or size of what is being analyzed, so that it can be visually represented. Definition of df in both fields is related, but not quite the same.

In physics, degree of freedom positions objects or systems, and each degree references a position in time, space or in other measurements. Df could be used synonymously with the term coordinate, and it usually means independent coordinates of the fewest number. The actual degree of freedom is based on the system being described in phase space or in all the potential types of space a system inhabits simultaneously. Every single part of phase space the system takes up can be considered a df, which helps to define the full realities of the system being considered.

From a statistical standpoint, degree of freedom defines distributions of populations or samples and is encountered when people begin to study inferential statistics: hypothesis testing and confidence intervals. As with the scientific definition, df in statistics describes shape or aspects of sample or population depending on data. Not all drawn representations of distributions have a degree of freedom measurement. The common standard normal distribution is not defined by degrees; instead, it will be the same bell-shaped curve in all instances.

A similar distribution to standard normal is student-t. The student-t is defined in part by degree of freedom in the formula n-1, where n is sample size. This means that were variables from the distribution to be picked one by one, all but the last one could be chosen freely. There is no choice but to take the very last one and no freedom to choose any other variable at that point. Therefore one variable is not free; its like having to pick the last tile out of a bag during a Scrabble game where there is no choice but to choose that letter.

Different distributions like the F and the chi-square have different definitions of degree of freedom, and some even use more than one df in definition. The issue gets confusing because df definition is linked to type of test performed and isnt the same with the various parametric (based on parameters) and non-parametric (not based on parameters) tests. Essentially, it wont always be n-1. Goodness of fit or contingency table testing may use the chi-square distribution with different df than that which evaluates single variable hypothesis testing of the variance or standard deviation.

What is important to remember is that each time degree of freedom is used to define a distribution, it changes it. It still may have certain characteristics that are unchanging, but size and appearance vary. When people are drawing representations of distributions, particularly two of the same distributions that have a different df, theyre advised to make them look different in size to convey that df is not the same.

# 5  Statistical significance

Statistical significance is a mathematical tool used to determine whether the outcome of an experiment is the result of a relationship between specific factors or due to chance. Statistical significance is commonly used in the medical field to test drugs and vaccines and to determine causal factors of disease. Statistical significance is also used in the fields of psychology, environmental biology, and any other discipline that conducts research through experimentation.

Statistics are the mathematical calculations of numeric sets or populations that are manipulated to produce a probability of the occurrence of an event. Statistics use a numeric sample and apply that number to an entire population. For the sake of example, we might say that 80% of all Americans drive a car. It would be difficult to question every American about whether or not they drive a car, so a random number of people would be questioned and then the data would be statistically analyzed and generalized to account for everyone.

In a scientific study, a hypothesis is proposed, then data is collected and analyzed. The statistical analysis of the data will produce a number that is statistically significant if it falls below 5%, which is called the confidence level. In other words, if the likelihood of an event is statistically significant, the researcher can be 95% confident that the result did not happen by chance.

Sometimes, when the statistical significance of an experiment is very important, such as the safety of a drug meant for humans, the statistical significance must fall below 3%. In this case, a researcher could be 97% sure that a particular drug is safe for human use. This number can be lowered or raised to accommodate the importance and desired certainty of the result being correct.

Statistical significance is used to reject or accept what is called the null hypothesis. A hypothesis is an explanation that a researcher is trying to prove. The null hypothesis holds that the factors a researcher is looking at have no effect on differences in the data. Statistical significance is usually written, for example, $t = .02, p < .05$. Here, "t" stands for the statistic test score and "p¡.05" means that the probability of an event occurring by chance is less than 5%. These numbers would cause the null hypothesis to be rejected, therefore affirming that the alternative hypothesis is true.

Here is an example of a psychological hypothesis using statistical significance: It is hypothesized that baby girls smile more than baby boys. In order to test this hypothesis, a researcher would observe a certain number of baby girls and boys and count how many times they smile. At the end of the observation, the numbers of smiles would be statistically analyzed.

Every experiment comes with a certain degree of error. It is possible that on the day of observation all the boys were abnormally grumpy. The statistical significance found by the analysis of the data would rule out this possibility by 95% if t=.03. In this case, the null hypothesis that baby girls do not smile more than baby boys would be rejected, and with 95% certainty, the researcher could say

that girls smile more than boys.

# 6    Fisher's exact test

Fisher's exact test is a statistical significance test used for small sample sizes. It is one of a number of tests used to analyze contingency tables, which display the interaction of two or more variables. Fisher's exact test was invented by English scientist Ronald Fisher, and it is called exact because it calculates statistical significance exactly, rather than by using an approximation.

To understand how Fisher's exact test works, it is essential to understand what a contingency table is and how it is used. In the simplest example, there are only two variables to be compared in a contingency table. Usually, these are categorical variables. As an example, imagine you are conducting a study on whether gender correlates with owning pets. There are two categorical variables in this study: gender, either male or female, and pet ownership.

A contingency table is set up with one variable on the top, and the other on the left side, so that there is a box for each combination of variables. Totals are given on the bottom and at the far right. Here is what a contingency table would look like for the example study, assuming a survey of 24 individuals:

```
 Pet Owner Not a Pet Owner Total
Male 1 9 10
Female 11 3 14
Total 12 12 24
```

Fisher's exact test calculates deviance from the null hypothesis, which holds that there is no bias in the data, or that the two categorical variables have no correlation with each other. In the case of the present example, the null hypothesis is that men and women are equally likely to own pets. Fisher's exact test was designed for contingency tables with a small sample size, or large discrepancies between cell numbers, like the one shown above. For contingency tables with a large sample size and well-balanced numbers in each cell of the table, Fisher's exact test is not accurate, and the chi-square test is preferred.

In analyzing the data in the table above, Fisher's exact test serves to determine the probability that pet-ownership is unevenly distributed among men and women in the sample. We know that ten of the 24 people surveyed own pets, and that 12 of the 24 are female. The probability that ten people chosen at random from the sample will consist of nine women and one man will suggest the statistical significance of the distribution of pet owners in the sample.

Probability is denoted by the letter p. Fisher's exact test determines the p-value for the above

data by multiplying the factorials of each marginal total – in the table above, 10, 14, 12, and 12 – and dividing the result by the product of the factorials of each cell number and of the grand total. A factorial is the product of all positive integers less than or equal to a given number. 10!, pronounced "ten factorial," is therefore equal to $10 \times 9 \times 8 \times \ldots \times 3 \times 2 \times 1$, or 3,628,800.

For the table above, then, $p = (10!)(14!)(12!)(12!)/(1!)(9!)(11!)(3!)(24!)$. Using a calculator, one can determine that the probability of getting the numbers in the table above is under 2%, well below chance, if the null hypothesis is true. Therefore, it is very unlikely that there is no contingency, or significant relationship, between gender and pet-ownership in the study sample.

Continuous data protection is a type of backup system that allows for information on a computer system to be constantly backed up on a different server, possibly even at a different physical location. Many consider this to be one of the safest forms of backup protection. While some systems back up the information stored on the main computer every day or even every few days, continuous data protection ensures that not even a day's work is lost in the event of a catastrophic computer failure.

In most cases, continuous data protection works by saving an exact copy of a file a user is working with to a different location. Each time the user changes the file, a new file is saved remotely, usually overwriting the previous file saved under that same name. Thus, it is not real-time backup in that each keystroke or move of the mouse is recorded and saved. That may be a very important distinction for those who are expecting to recover lost work because of things like power failures, where a computer will shut down unexpectedly. Real-time backup products are also available, for those who want them.

Computer users can utilize continuous data backup in a number of different ways. Software products currently on the market allow users to save backup settings so that each file saved will go automatically to a backup system. Those who do not want the extra expense of more servers, can utilize an Internet service that offers continuous data protection. In such cases, each file saved is sent through an Internet connection to the service's servers.

If using an Internet service that provides continuous data protection, there is usually a monthly or yearly charge for the service. Also, some companies may only allow a certain amount of space, or at least have tiered pricing levels, so that customers can choose the amount of space they will need. That allows users to have a certain amount of control over their pricing, but that could also lead to higher charges or loss of data if users go beyond their maximum storage level allowed.

For those who want the ultimate in protection using a real-time process, it may be best to consider a continuous data protection system with a uninterruptible power supply. This allows the computer to maintain power in the event the regular power has gone off. While these power supplies generally only have a limit of a few minutes or hours, it should be enough time to save all critical files to the data protection system, and shut down the machines safely.

## 6.1 Hypothesis testing

The standard deviation of the life for a particular brand of ultraviolet tube is known to be $S = 500hr$, and the operating life of the tubes is normally distributed. The manufacturer claims that average tube life is at least 9,000hr. Test this claim at the 5 percent level of significance against the alternative hypothesis that the mean life is less than 9,000 hr, and given that for a sample of $n = 18$ tubes the mean operating life was $\bar{X} = 8,800hr$.

## 6.2 two populations

Two samples drawn from two populations are independent samples if the selection of the sample from population 1 does not affect the selection of the sample from population 2. The following notation will be used for the sample and population measurements:

- $p_1$ and $p_2$ = means of populations 1 and 2,

- $\sigma_1$ and $\sigma_2$ = standard deviations of populations 1 and 2,

- $n_l$ and $n_2$ = sizes of the samples drawn from populations 1 and 2 $(n_1 > 30, n_2 > 30)$,

- $x_1$ and $x_2$, = means of the samples selected from populations 1 and 2,

- $s_1$ and $s_2$ = standard deviations of the samples selected from populations 1 and 2.

## 6.3 Standard Error

$$S.E(\bar{X}_1 - \bar{X}_2) = \sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})} \tag{1}$$

## 6.4 Example

The mean height of adult males is 69 inches and the standard deviation is 2.5 inches. The mean height of adult females is 65 inches and the standard deviation is 2.5 inches. Let population 1 be the population of male heights, and population 2 the population of female heights. Suppose samples of 50 each are selected from both populations.

## 6.5 Example 3

Ten replicate analyses of the concentration of mercury in a sample of commercial gas condensate gave the following results (in ng/ml) :

| 23.3 | 22.5 | 21.9 | 21.5 | 19.9 | 21.3 | 21.7 | 23.8 | 22.6 | 24.7 |

Compute 99% confidence limits for the mean.

# 7 Hypothesis Tests for Two Means

If the population standard deviations $\sigma_1$ and $\sigma_2$ are known, the test statistic is of the form:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2}$$

The critical value and p-value are looked up in the normal tables.

# 8 Using the p-value

# 9 Significance

# 10 ANOVA

# 11 Modelling

The purpose of modelling can be twofold. The first is prediction. Given a set of analytical data, we want to be able to predict properties of the samples that cannot be measured easily. An example is the assessment of whether a specific treatment will be useful for a patient with particular characteristics.

Such an application is known as classification - one is interested in modelling class membership (will or will not respond).

The other major field is regression, where the aim is to model continuous real variables (blood pressure, protein content, ...). Such predictive models can mean a big improvement in quality of life, and save large amounts of money.

The prediction error is usually taken as a quality measure: a model that is able to predict with high accuracy must have captured some real information about the system under study. Unfortunately, in most cases no analytical expressions can be derived for prediction accuracy, and other ways of estimating prediction accuracy are required in a process called validation. A popular example is **_cross-validation_**.

# 12 Testing Normality

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality - graphically and numerically.

## 12.1 Independent one-sample $t$-test

In testing the null hypothesis that the population mean is equal to a specified value $\mu_0$, one uses the statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{3}$$

where $s$ is the sample standard deviation and $n$ is the sample size. The degrees of freedom used in this test is $n - 1$.

## 12.2 Confidence intervals

Most studies will only sample part of a population and then the result is used to interpret the null hypothesis in the context of the whole population. Any estimates obtained from the sample only approximate the population value. Confidence intervals allow statisticians to express how closely the sample estimate matches the true value in the whole population. Often they are expressed as 95% confidence intervals. Formally, a 95% confidence interval of a procedure is any range such that the interval covers the true population value 95% of the time given repeated sampling under the same conditions.

If these intervals span a value (such as zero) where the null hypothesis would be confirmed then this can indicate that any observed value has been seen by chance. For example a drug that gives a mean increase in heart rate of 2 beats per minute but has 95% confidence intervals of -5 to 9 for its increase may well have no effect whatsoever.

The 95% confidence interval is often misinterpreted as the probability that the true value lies between the upper and lower limits given the observed sample. However this quantity is more a credible interval available only from Bayesian statistics.

### 12.2.1  Confidence intervals - example

A researcher was investigating computer usage among students at a particular university. Three hundred undergraduates and one hundred postgraduates were chosen at random and asked if they owned a laptop. It was found that 150 of the undergraduates and 80 of the postgraduates owned a laptop.

Find a 95% confidence interval for the difference in the proportion of undergraduates and postgraduates who own a laptop. On the basis of this interval, do you believe that postgraduates and undergraduates are equally likely to own a laptop?

## 12.3  difference of two proportions - example

Two time-sharing systems are compared according to their response time to an editing command. The mean response time of 100 requests submitted to system 1 was measured to be 600 milliseconds with a known standard deviation of 20 milliseconds. The mean response time of 100 requests on system 2 was 592 milliseconds with a known standard deviation of 23 milliseconds. Using a significance level of 5%,test the hypothesis that system 2 provides a faster response time than system 1. Clearly state your null and alternative hypotheses and your conclusion.

## 12.4  Difference between two population proportions

When we wish to test the hypothesis that the proportions in two populations are not different, the two sample proportions are pooled as a basis for determining the standard error of the difference between proportions. Note that this differs from the procedure used in Section 9.5 on statistical estimation, in which the assumption of no difference was not made.

Further, the present procedure is conceptually similar to that presented in Section 11.1, in which the two sample variances are pooled as the basis for computing the standard error of the difference between means. The pooled estimate of the population proportion, based on the proportions obtained in two independent samples

## 12.5  F-test of equality of variances

The test statistic is

$$F = \frac{S_X^2}{S_Y^2} \tag{4}$$

has an F-distribution with $n-1$ and $m-1$ degrees of freedom if the null hypothesis of equality of variances is true.

Further Inference

# 13  Fractional factorial design

(d) Define the following terms used in fractional factorial design; Defining relation, Generator, Confounding, Resolution. Which design resolution is considered optimal?

# 14  Chi Square

$p_i$ = Expected proportion for digit $i$.

For this test we used the chi-squared test statistic which is given by:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

The Chi Square test tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable.

- $X^2$ = the test statistic that asymptotically approaches a $\chi^2$ distribution.

- $O_i$ = an observed frequency;

- $E_i$ = an expected (theoretical) frequency, asserted by the null hypothesis;

- $n$ = the number of possible outcomes of each event.

The chi-square statistic can then be used to calculate ap-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom is equal to the number of cells "n", minus the reduction in degrees of freedom, "p".

## 14.1   Chi Square example

In reading a burette to 0.01ml the final figure has to be estimated. The following frequency table gives the final figures of 40 such readings.

| Digit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------|---|---|---|---|---|----|---|---|
| Frequency | 1 | 6 | 4 | 5 | 3 | 11 | 2 | 8 |

The null hypothesis is that each digit has equal chances of occurring. Since we have ten digits this implies that each digit should have a 12.5% chance of occurring.

Mathematically we can express the null hypothesis as:

$H_0 : p_0 = p_1 = p_1 = \ldots = p_7 = 0.125$

$$X^2 = \frac{(1-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(11-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(8-5)^2}{5} \quad (6)$$

## 14.2   Goodness of fit example

Pressure readings are taken regularly from a meter. It transpires that, in a random sample of 100 such readings, 45 are less than 1, 35 are between 1 and 2, and 20 are between 2 and 3.

Perform a $\chi^2$ goodness of fit test of the model that states that the readings are independent observations of a random variable that is uniformly distributed on (0, 3).

## 14.3   Chi Square contingency tables

In a survey the samples from five factories were examined for the number of skilled or unskilled workers.

| Factory | skilled workers | unskilled workers |
|---------|-----------------|-------------------|
| A | 80 | 184 |
| B | 58 | 147 |
| C | 114 | 276 |
| D | 55 | 196 |
| E | 83 | 229 |

Does the population proportion of skilled and unskilled workers vary with the factory?

# 15 Inference

## 15.1 Confidence interval of a mean (small sample)

If the data have a normal probability distribution and the sample standard deviation $s$ is used to estimate the population standard deviation $\sigma$, the interval estimate is given by:

$$\bar{X} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} \tag{7}$$

where $t_{1-\alpha/2,n-1}$ is the value providing an area of $\alpha/2$ in the upper tail of a Students t distribution with n - 1 degrees of freedom.

### 15.1.1 Example using R

Finding confidence intervals for the mean for the nitrate ion concentrations in Example 2.7.1.

```
#Typing data in
x=c(102,97,99,98,101,106)
mean(x)
sd(x)
n=length(x)
#setting the confidence level
CL=0.95
#computing confidence interval
pm=sd(x)*c(qt(0.025,n-1),qt(0.975,n-1))/sqrt(n)
CI=mean(x)+pm
```