Confusion matrix

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

Ex:

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

A confusion matrix has two possible predicted classes i.e. ‘yes’ and ‘no’

Basic terms of confusion matrix:

1. True Positives(TP): These are the case that the model has predicted ‘yes’ which is true with the actual result.
2. True Negatives(TN):These are the cases that the model predicted no which is true with the actual result.

3. False Positives(FP): These are the cases that the model predicted ‘yes’ where the actual results are ‘no’.
4. False Negatives(FN): These are the cases that the model predicted ‘no’ where the actual results are ‘yes’.

The confusion matrix allows to compute the following:

Accuracy: Overall, how often is the model correct?

(TP+TN)/total

Misclassification Rate: Overall, how often is it wrong?

(FP+FN)/total

True Positive Rate: When it's actually yes, how often does it predict yes?

TP/actual yes

False Positive Rate: When it's actually no, how often does it predict yes?

FP/actual no

Specificity: When it's actually no, how often does it predict no?

TN/actual no ; equivalent to 1 minus False Positive Rate

Precision: When it predicts yes, how often is it correct?

TP/predicted yes

Prevalence: How often does the yes condition actually occur in our sample?

actual yes/total

Basics of Probability Distribution

Random Variables

A random variable denotes the possible outcome of an event.

It can be discrete (i.e. finite with many possible outcomes of a random event) or continuous.

***Examples of discrete random variable*:**

1. Outcome of a coin toss (Head/Tail) denoted by 0/1.
2. Outcome of a dice roll (1,2,3,4,5,6)

Examples of continuous random variable:

1. Heights of students in a classroom.

2. Test scores of a student in an examination.

For discrete random variable, probability is measured using the 'Probability Mass Function(PMF)'

For continuous random variable, probability is measured through probability density function (i.e. probability within an interval).

Some Important Probability Distributions

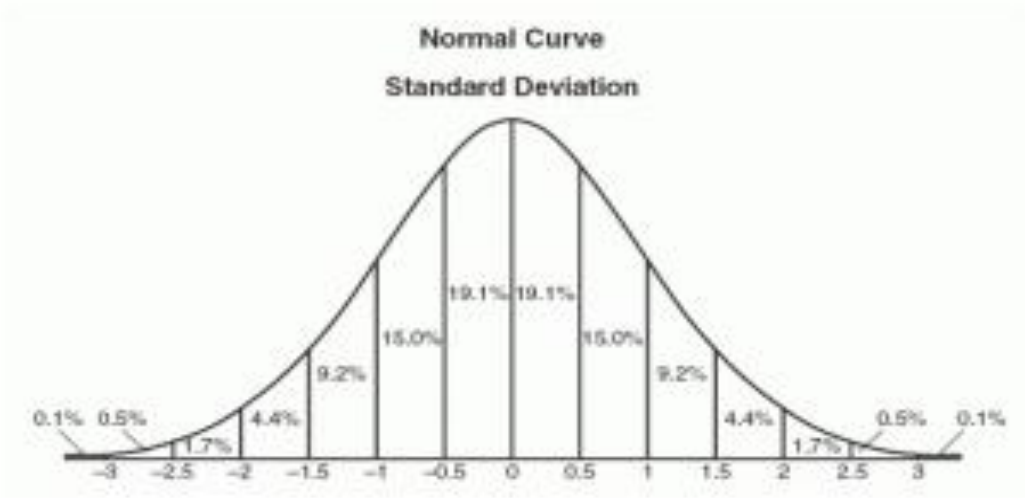
1. The Normal Distribution

It is the most important and widely used distribution in statistics.

It is sometimes called the “bell curve” and also sometimes as “Gaussian curve”

It is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Ex: Height is one simple example of something that follows a normal distribution .



Seven important features of normal distribution

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.

   (i.e. the peak is always in the middle and the curve is always symmetrical.)
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

Exercise

A survey of daily travel time(in minutes) of a passenger had the following results:

26	**33**	**65**	**28**	**34**	**55**	**25**	**44**	**50**	**36**	**26**	**37**	**43**	**62**	**35**	**38**	**45**	**32**	**28**	**34**

Prepare a normal distribution graph

Mean = $\frac{26+33+65+28+34+55+25+44+50+36+26+37+43+62+35+38+45+32+28+34}{20}$

=776/20 = 38.8 minutes ; Standard Deviation=11.4 minutes

Convert the values to z-scores and prepare the Normal Distribution Graph.

Formula for z-score is $z = \frac{x-\mu}{\sigma}$ where z= the z-score(standard score)

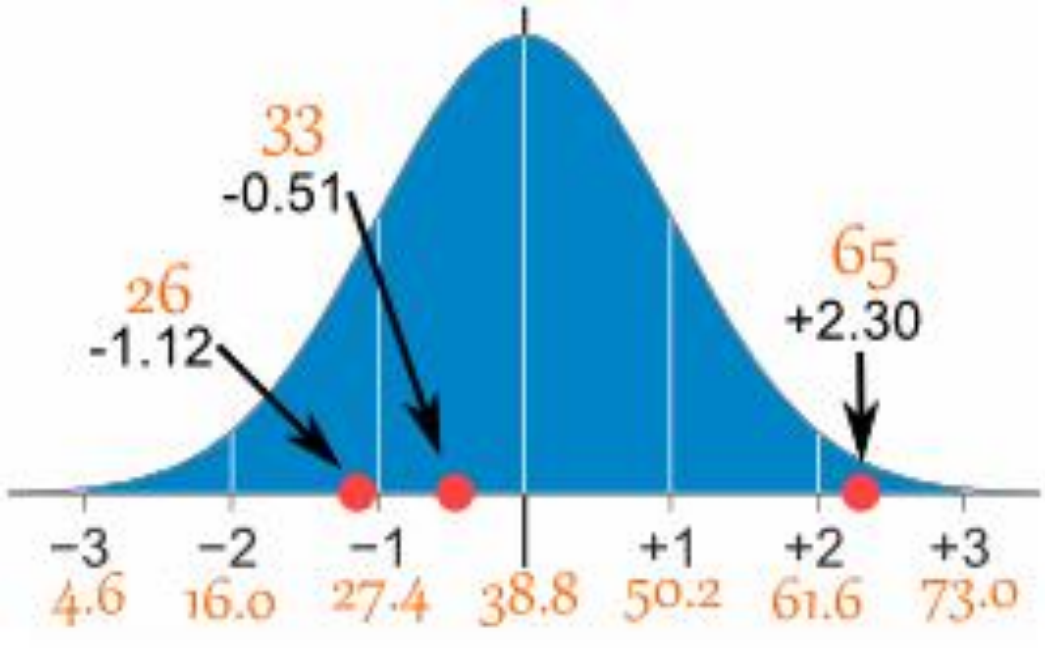
x= the value to be standardized

μ= mean; σ = the standard deviation

The conversion

Original Value	Calculation	Standard Score(z-score)
26	(26-38.3)/11.4	-1.12
33	(33-33.8)/11.4	-0.51
65	(65-33.8)/11.4	-2.30
28	(28-33.8)/11.4	-0.50
34	(34-33.8)/11.4	0.01
55	(55-33.8)/11.4	1.85
25	(25-33.8)/11.4	-0.77
44	(44-33.8)/11.4	0.89
50	(50-33.8)/11.4	1.42
36	(36-33.8)/11.4	0.19
26	(16-33.8)/11.4	-1.56
37	(37-33.8)/11.4	0.28
43	(43-33.8)/11.4	0.80
62	(62-33.8)/11.4	2.47
35	(35-33.8)/11.4	0.10

Original Value	Calculation	Standard Score(z-score)
38	(38-33.8)/11.4	0.36
45	(45-33.8)/11.4	0.98
32	(32-33.8)/11.4	-0.15
28	(28-33.8)/11.4	-0.50
34	(34-33.8)/11.4	0.01



Discrete Probability Distribution

Introduction

Probability distribution can be continuous or discrete depending on whether it defines probabilities for continuous or discrete variables.

A discrete distribution describes the probability of occurrence of each value of a discrete random variable.

With a discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability.

A discrete probability distribution is often presented in tabular form.

1. Geometric Distribution

It is the distribution of number of trials of needed to get the first success in repeated independent Bernouli trials.

Independent trials can result in one of two possible outcomes labeled *success* and *failure*

P(Success) =p and P(Failure) = 1-p

Random variable ‘x’ represents the number of successes in n trials.

If ‘x’ denotes the no. of trials, then to get first success in this x trials:

1. The first x-1 trials must be failures.(i.e. $(1-p)^{x-1}$)
2. The x^{th} trail must be success. (i.e. p)

Hence $P(X=x)=(1-p)^{x-1}\ p$; for x=1,2,3,…..

$$\mu = 1/p;\ \sigma^2 = (1-p)/p^2$$

Example:

In a population of employees at a company, 30% have received job skills training. If employees from this population are randomly selected, what is the probability that the 6^{th} person sampled is the first person that has received the training.

$P(X=x)=(1-p)^{x-1}\ p$; Here x= 6 , p= 30% i.e. 0.3

$P(X=6)=(1-0.3)^{6-1}\ 0.3=0.0504$

Build a distribution pattern for x=1,2,3,…….15

Original value	Calculation	Probability
P(X=1)	$(1-0.3)^{1-1}$ 0.3	0.3
P(X=2)	$(1-0.3)^{2-1}$ 0.3	0.21
P(X=3)	$(1-0.3)^{3-1}$ 0.3	0.147
P(X=4)	$(1-0.3)^{4-1}$ 0.3	0.1029
P(X=5)	$(1-0.3)^{5-1}$ 0.3	0.072
P(X=6)	$(1-0.3)^{6-1}$ 0.3	0.0504
P(X=7)	$(1-0.3)^{7-1}$ 0.3	
P(X=8)	$(1-0.3)^{8-1}$ 0.3	
P(X=9)	$(1-0.3)^{9-1}$ 0.3	
P(X=10)	$(1-0.3)^{10-1}$ 0.3	
P(X=11)	$(1-0.3)^{11-1}$ 0.3	
P(X=12)	$(1-0.3)^{12-1}$ 0.3	
P(X=13)	$(1-0.3)^{13-1}$ 0.3	
P(X=14)	$(1-0.3)^{14-1}$ 0.3	
P(X=15)	$(1-0.3)^{15-1}$ 0.3	

Observations

Geometric distribution is always right skewed.

Mode of GD(the most likely value) is always1

Mean $\mu = 1/p = 1/0.3 = 3.33$

Variance $\sigma^2 = (1-p)/p^2 = (1-0.3)/(0.3)^2$

$= 7.7$

Standard deviation $\sigma = \sqrt{7.7} = 2.79$

What is the probability that the first person trained in the training occurs on or before the 3rd person sampled?

i.e. $P(X<=3) = P(X=1)+P(X=2)+P(X=3)$

$=0.3+0.21+0.147$

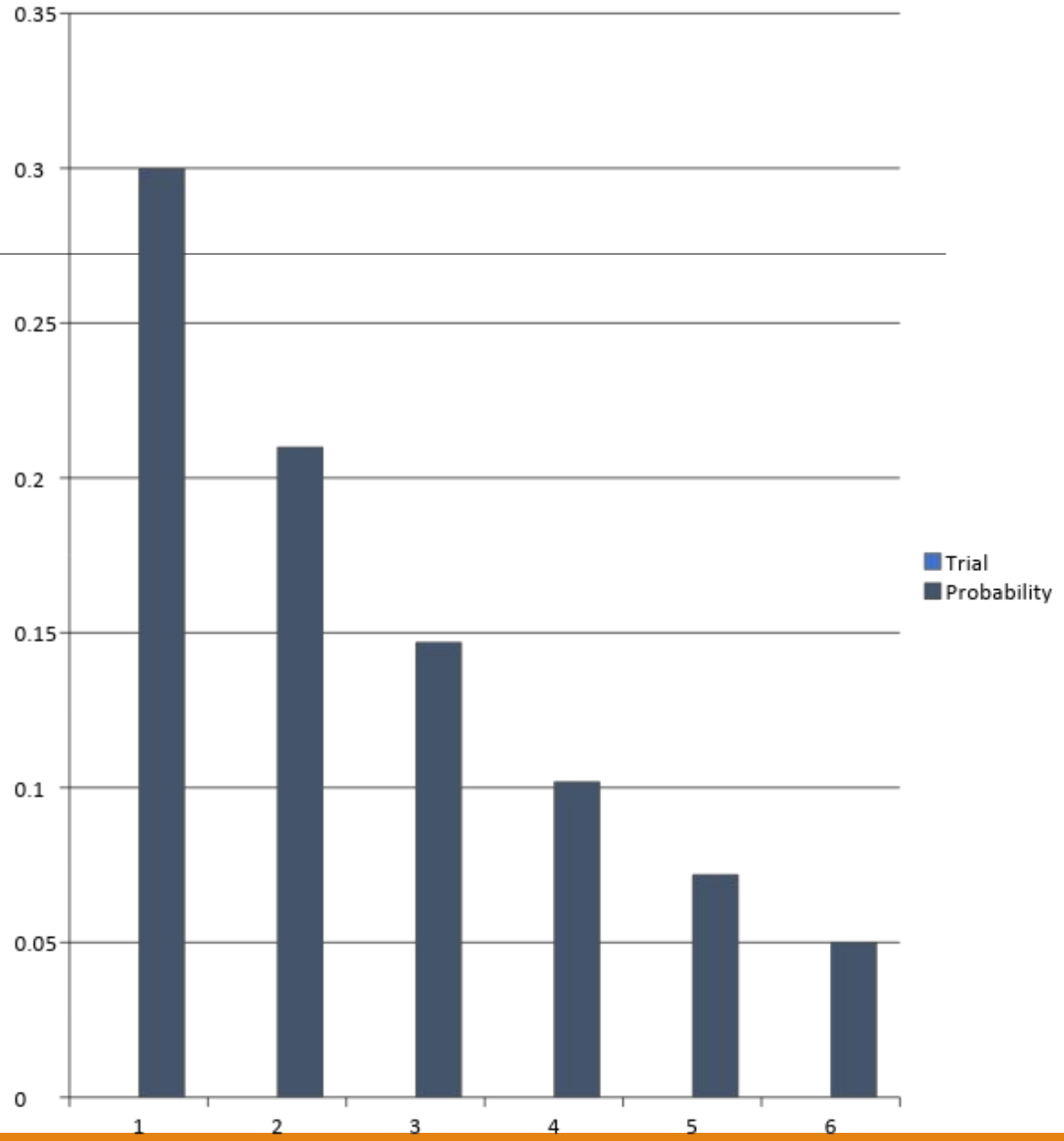
$=0.657$

The cumulative distribution function is

$F(x)= P(X <=x) = 1-(1-p)^x$; x=1,2,3,…

$P(X>3)=(1-0.3)^3 = 0.7^3$

$P(X<=3)= 1-0.7^3 = 0.657$



Bernoulli Distribution

It is another important discrete probability distribution.

If X denotes a random variable where X=1 if a success occurs and X=0 if a failure occurs, then X has a Bernoulli distribution as:

$P(X=x) = p^x (1-p)^{1-x}$; For x=0,1

Mean of a Bernoulli random variable(μ) = p

Variance($\sigma^{2)}$ = p(1-p)

Example:

Approximately 1 in 200 of American adults are lawyers. If one American lawyer is selected at random, what is the distribution of the number of lawyers?

Solution:

If 'X' denotes random variable then X=1 if a success occurs and X=0 if a failure occurs

Then X has Bernoulli distribution:

$P(X=x) = p^x (1-p)^{1-x}$; For x=0,1

p=1/200

$P(X=x) = (1/200)^x (1-1/200)^{1-x}$; for x=0,1

P(X=1) = 1/200 ; P(X=0) =199/200

Binomial Distribution

It is the most important discrete probability distribution.

It is based on the principle that the number of successes in *n* independent Bernoulli trials has a binomial distribution.

Independent trials can result in one of two possible outcomes labeled *success* and *failure*

P(Success) =p and P(Failure) = 1-p

Random variable 'x' represents the number of successes in n trials.

Then X has a binomial distribution as

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \quad ; \text{for } x = 0,1,2,\ldots,n.$$

Mean of binomial distribution $\mu = np$; Variance $\sigma^2 = np(1-p)$

Ex:

If a fair die is rolled three times, what is the probability of getting 5 exactly twice?

X represents no. of fives in 3 rolls (i.e. X=2)

N =3 and p=1/6 (i.e. probability of getting a 5 out of 6)

$$P(X=2) = \binom{3}{2} (1/6)^2 (1-1/6)^{3-2}$$

Example:

According to the statistics of Canada life tables, the probability a randomly selected 90-year old Canadian male survives for at least another year is approximately 0.82.

If twenty 90-year old Canadian males are randomly selected, what is the probability exactly 18 survive at least another year?

Success: The man survives for at least one year.

Failure: The man dies within one year.

In this case, n=20, p=0.82

To find the probability that 18 males survive,

$$P(X=18) = \binom{20}{18} 0.82^{18} (1-0.82)^{20-18} = 0.173$$

Mean (μ) = np = 20 x 0.82 = 16.4 ; i.e. 16.4 out 20 randomly selected Canadian males will survive at least for an year.

Variance(σ^2) =np(1-p)=20 x 0.82 x (1-0.82) = 2.952

Calculate probabilities for X=0,2,4,6,8,10,12,14,16,18,20

Poisson Distribution

Events in a space can occur independently.

The probability that an event occurs in a given length of time does not change through time.

If events occur randomly and independently then a random variable X which represents the number of events in a fixed unit of time has a Poisson distribution.

$$P(X=x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

For Poisson distribution mean (μ) = λ

Variance (σ^{2}) = λ

Example:

Plutonium-239 is an isotope of plutonium that is used in nuclear weapons and reactors. One nanogram of Plutonium-239 will have an average of 2.3 radioactive decays per second, and the number of decays will follow a Poisson distribution.

What is the probability that in a 2 second period there are exactly 3 radioactive decays?

Solution:

Let 'x' represent the number of decays in a 2 second period.

$\lambda = 2.3 \times 2 = 4.6$

$$P(X=3) = \frac{4.6^{3} e^{-4.6}}{3!} = 0.163$$; Similarly calculate for different values of X i.e. X=0,1,2,…..,15

Poisson distribution has some right skewness depending on the value of λ.

When distribution is large λ would be close to symmetric and when λ is close to zero right skewness can be pretty strong.

Continuous Probability Distribution

Introduction

Continuous random variables can take on an infinite number of possible values, corresponding to every value in an interval.

Ex: Height of adults

Continuous random variables can be modeled with a separate method called a *'probability density function'.*

Ex: We can model the distribution of time to failure of a particular type of light bulb as continuous probability distribution curve.

For continuous random variables, probabilities are areas under the curve. The area under the entire curve is equal to one.

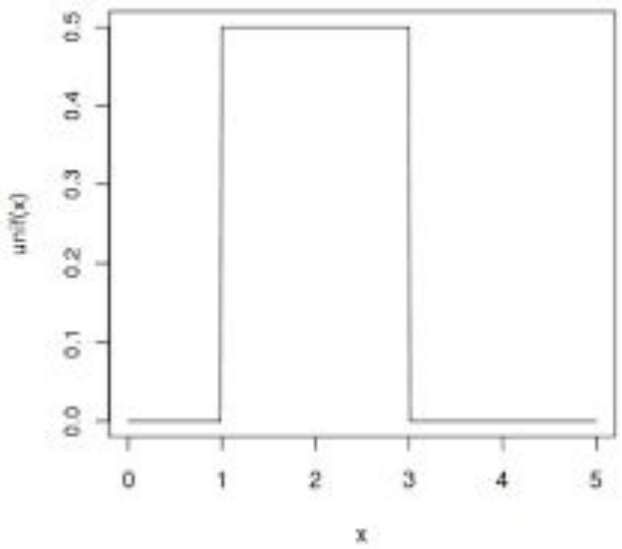
An important notion of a continuous random variable 'x' equal to any specific value would be zero

i.e. $P(X=a) = 0$ for any a or $P(X=b)=0$ for any b

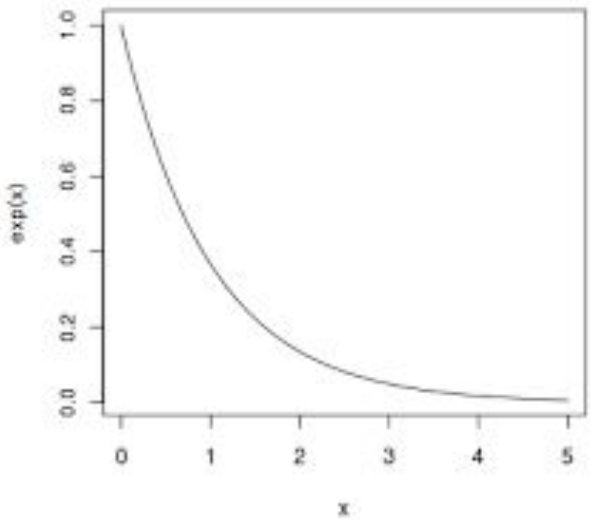
So the random variable x only falls in the interval between a and b. Mathematically:

$P(a \leq X \leq b)$

Types



Continuous Uniform Distribution



Continuous Exponential Distribution

Exponential Distribution

It is one of the important continuous probability distributions.

Time is the base factor for such distribution i.e. exponential distribution will provide a description of the length of time between occurrences.

Generally the exponential distribution describes waiting time between Poisson occurrences

Ex:

1. How much time will elapse before an earthquake occurs in a given region?
2. How long a shopkeeper need to wait before a customer enters a shop?
3. How long does a machinery work without breaking down?

The probability density function is defined by:

$f(t) = \lambda e^{-\lambda t}$, $t > 0 = 0$ otherwise

Example: If jobs arrive every 15 seconds on average, $\lambda = 4$ per minute, what is the probability of waiting less than or equal to 30 seconds, i.e .5 min?

Let T = time that elapses after a Poisson event.

P(T > t) = probability that no event occurred in the time interval of length t.

The probability that no Poisson event occurred in the time interval [0, t]: $P(0, t) = e^{-\lambda t}$; where λ is the average Poisson occurrence rate in a unit time interval.

So: $P(T > t) = e^{-\lambda t}$

The PDF:

$f(t) = e^{-\lambda t}$, $t > 0$

= 0 otherwise

$$P(T \leq .5) = \int_0^{.5} 4e^{-4t}dt$$

$$= \left[-e^{-4t}\right]_{t=0}^{.5}$$

$$= 1 - e^{-2}$$

$$= 0.86$$

t-Distribution

The t-distribution is member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

According to the central limit theorem, the sampling distribution of a statistic will follow a normal distribution, as long as the sample size is sufficiently large.

The t-distribution with degrees of freedom "*n – 1*" is given below.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

where $\bar{x}$ is sample mean and μ is population mean

s

Example: The CEO of light bulbs manufacturing company claims that an average light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

Solution:

Compute the t static $t = \frac{290 - 300}{\frac{50}{\sqrt{15}}} = -10/12.909945 = 0.7745966$

The degrees of freedom are equal to 15 – 1 = 14.

Assuming the CEO's claim is true, the population mean equals 300.

The sample mean equals 290.

The standard deviation of the sample is 50.

The cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days

Inferential Statistics

Introduction

Inferential statistics is one of the two main branches of statistics.

It uses a random sample of data taken from a population to describe and make inferences about the population.

Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible.

Ex:

Measuring the diameter of each nail that is manufactured in a mill is impractical.

But a representative random sample of nails can be taken whose diameter can be measured.

Information from the sample to make generalizations about the diameters of all of the nails.

Two main areas of inferential statistics:

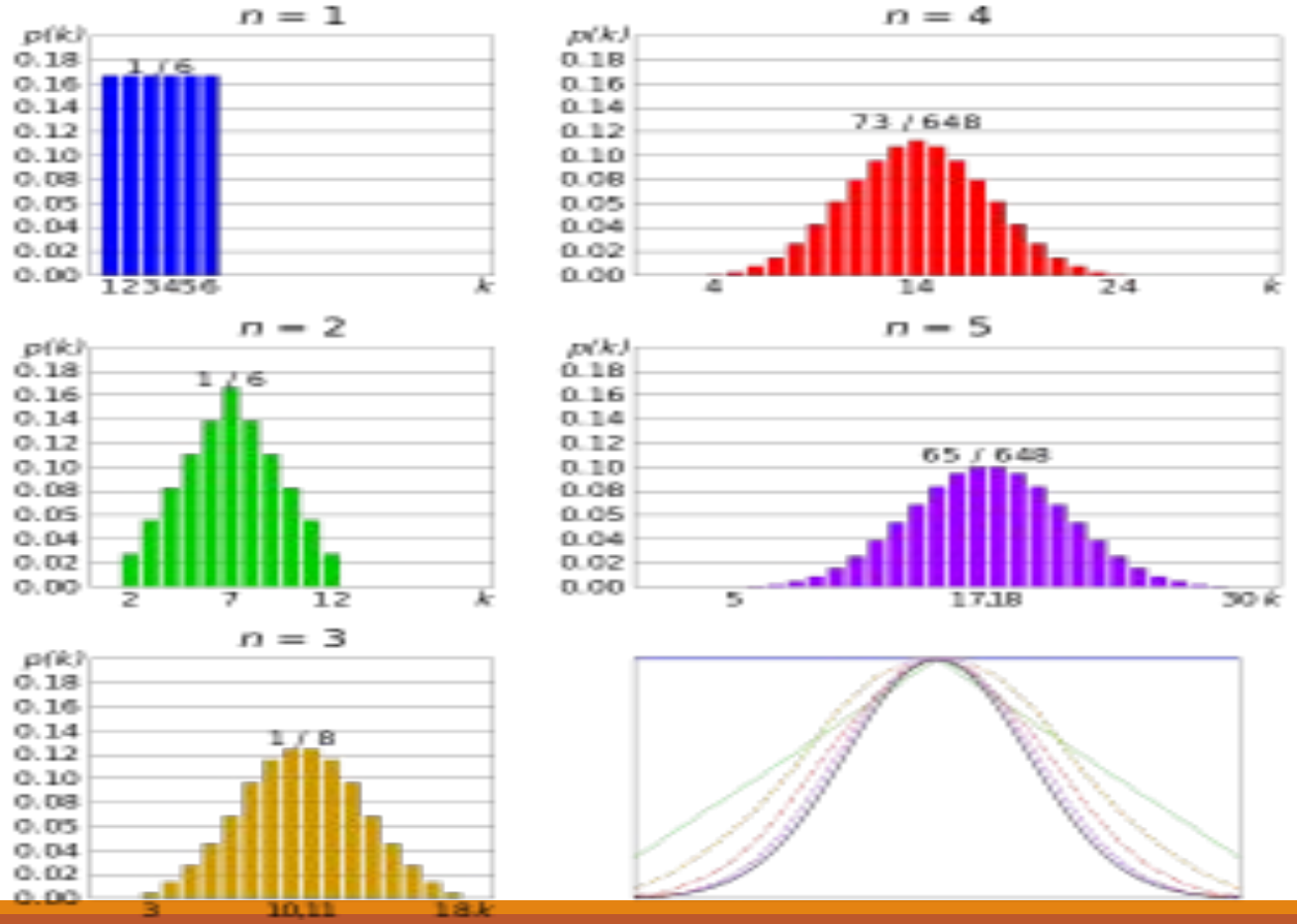
1. *Estimating parameters:* This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
2. *Hypothesis tests:* This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger — *no matter what the shape of the population distribution*.

It holds especially true for sample sizes over 30.

For large samples graph of the sample means will look more like a normal distribution.



Sampling Distribution

A sampling distribution is a graph of a statistic for the sample data.

A sampling distribution is where you a population (N), is taken and a statistic from that population is found.

Common statistics include:

1. Mean
2. Mean absolute value of the deviation from the mean
3. Range
4. Standard deviation of the sample
5. Unbiased estimate of variance
6. Variance of the sample

Understand Sampling Distribution

Sample proportion:

A sample proportion is where a random sample of objects *n* is taken from a population P; if x objects have a certain characteristic then the sample proportion "p" is: p = x/n.

Ex:

100 people are asked if they are democrat. If 40 people respond "yes" then the sample proportion is

p = 40/100.

Sampling Distribution of a Population

The sampling distribution of a proportion is when you repeat your survey for all possible samples of the population.

Ex:

Instead of polling 100 people once to ask if they are democrat, you'll poll them multiple times to get a better estimate of your statistic.

Confidence Interval and Hypothesis Testing

What is Confidence Interval?

An experiment can take a random sample from a lot or population and compute a statistic such as mean from the data to help understand the mean of the population.

However the challenge /issue can be like how well the computed sample statistic(i.e. sample mean) estimates the underlying population.

'Confidence interval' is the solution to address the issue as it provides a range of values which is likely to contain the population parameter of interest.

How to construct confidence interval?

A confidence interval is constructed at a *confidence level* (usually 95%),selected by the user.

i.e. if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases.

Types of confidence intervals

Confidence intervals can be one or two-sided.

A two-sided confidence interval brackets the population parameter from above and below.

A one-sided confidence interval brackets the population parameter either from above or below and furnishes an upper or lower bound to its magnitude.

Understand Confidence Interval

What is Hypothesis?

A hypothesis is an **educated guess** about something in the world around us.

An experiment conducted statistically to know the true relationship between an independent variable and dependent variable.

Ex: A new medicine might work better in treating headache.

A good hypothesis statement should include an “if” and “then” statement addressing the independent and dependent variables.

It should be testable by experiment, survey or other scientifically sound technique.

Should be based on information in prior research and have a design criteria.

How to prove a hypothesis?

Hypothesis Testing

Aims at testing the results of a survey or experiment to see if the results are meaningful.

i.e. testing whether the results are valid by figuring out the results that have happened by chance.

If results may have happened by chance, then the experiment won’t be repeatable and has little use.

The Process

Hypothesis testing usually involves:

1. Figure out null hypothesis.
2. State null hypothesis.
3. Choose the kind of statistical test(s) to be performed.
4. Either support or reject the null hypothesis.

What is Null Hypothesis?

Null hypothesis is always the accepted fact.(*generally being accepted as true)*

Ex:

1. There are eight planets in the solar system(excluding pluto).
2. DNA is a double shaped helix.
3. Moon is a satellite of earth.

Null Hypotheses & Alternative Hypotheses

Used in the context of statistical analysis.

Null hypotheses is symbolized as H_0 and usually expresses the uniformity/equality among the items in a sample.

Ex: In a sample of two methods if a researcher has an assumption where both methods are equally good, then such assumption is termed as the null hypotheses.

In the sample of two methods if a researcher thinks that method A is superior or method B is inferior, then such assumption is termed as alternative hypotheses.

In a research if a sample does not support the null hypotheses, then it can be concluded that null hypotheses can be rejected and the researcher looks for a set of alternatives that form an alternative hypotheses.

If μ(mean) of a population is equal to the hypothesised mean(μ_0) then it is evident that null hypotheses can be

population mean is equal to hypothesised mean.

(or)

$\mu = \mu_0 = 100$.

How to State Null Hypothesis?

To state a null hypothesis, it would be sufficient if one can figure out the hypothesis from the world problems of real life situations which can be a little trickier than just figuring out what the accepted fact is.

Ex:

A researcher might believe that knee surgery patients need a physical therapy twice a week instead of 3 times which otherwise results in longer recovery period (i.e. average recovery time is 8.2 weeks).Mathematically,

$H_1 : \mu > 8.2$

Now to state null hypothesis, mathematically :

$H_0 : \mu \leq 8.2$

Type 1 and Type II Errors

Two types of errors are possible in the context of testing of hypotheses

Null hypotheses H_0 can be rejected if it is true (Type I error) and H_0 may be accepted when it is not true(Type II error).

Type I error- Rejection of hypotheses which should have been accepted.

Type II error-

Possible Hypotheses Test Outcomes		
	Actual Situation	
Decision	*H_0 True*	*H_0 False*
Accept H_0	No Error	Type II Error
	Probability= 1- alpha	Probability= beta
Reject H_0	Type I Error	No Error
	Probability = alpha	Probability= 1- beta

Testing The Hypotheses

Given a hypotheses H_0 and alternative hypotheses H_1, a researcher may accept H_0 (i.e. reject H_1) or reject H_0 (i.e. accept H_1)

Ex: Suppose the mean age of the people in a city is 40 years. To conduct hypotheses testing,

i. Mean age of population in a city is 40 years

H_0 : μ= 40 Against $H_{1:}$ μ= 40

Ii. Mean age of the people in a city is 40 years or higher

H_0 : μ= 40 Against $H_{1:}$ μ> 40

iii. Mean age of the people in a city is 40 years or lower

H_0 : μ= 40 Against $H_{1:}$ μ< 40

Suppose sample mean be 20 years. This is lower than population mean(40 years). If H_0 is true then probability of getting such a different sample mean would be very small.

If sample mean is close to the assumed population mean, H_0 is accepted.

If sample mean is far-off from the assumed population mean, H_0 is rejected.

Note: In hypotheses testing, it is assumed that the null hypotheses is true and we proceed to reject null hypotheses using the sample.

Statistical Hypothesis Testing

t-test

It is also called ‘Student’s t-test’.

It compares two means(averages) and tells if they are different from each other.

It also tells how significant the differences are(i.e . whether the differences could have happened by chance)

Ex 1:

A person suffering from cold tried a naturopathic remedy and the cold lasts for a couple of days. The next time when the person has cold, bought a pharmaceutical tablet and the cold lasts a week. After surveying the friends, it was figured out that their colds were of a shorter duration(say 3 days on average)when they took homeopathic remedy.

Are these results repeatable?

Solution:

A t-test can tell by comparing the means of the two groups and letting you know the probability of those results happening by chance.

Ex2:

a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there’s always a control group he control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student’s t-test to find out if the results are repeatable for an entire population.

Types of t-test

There are **three main types of t-test:**

An Independent Samples t-test compares the means for two groups.

A Paired sample t-test compares means from the same group at different times (say, one year apart).

A One sample t-test tests the mean of a single group against a known mean.

A paired t test (also called a **correlated pairs t-test**, a **paired samples t test** or **dependent samples t test**) is where you run a t test on dependent samples. Dependent samples are essentially connected

Ex:

1. Two tests on the same person before and after training
2. Two blood pressure measurements on the same person using different equipment.

Choose the paired t-test if you have two measurements on the same item and should also choose this test if you have two items that are being measured with a unique condition.

Ex:

Measuring car safety performance in Vehicle Research and Testing and subject the cars to a series of crash tests. Although the manufacturers are different, you might be subjecting them to the same conditions.

A “regular” two sample t test, compares the means for two different samples.

Ex:

1. Testing two different groups of customer service associates on a business-related test.
2. Testing students from two universities on their English skills.

If random sample is taken from each group separately and they have different conditions, then the samples are independent and we should run an independent samples t test (also called between-samples and unpaired-samples).

The null hypothesis for the for the independent samples t-test is $\mu_1 = \mu_2$. (i.e. means are equal).

With the paired t test, the null hypothesis is that the *pairwise difference* between the two tests is equal(i.e. (H_0: $\mu_d = 0$).

Paired Samples t-test by hand

z-test

A Z-test is a type of hypothesis test.

A Z-test, is used when your data is approximately normally distributed.

When to use a z-test?

✓ Sample size should greater than 30. Otherwise use t-test

✓Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.

✓Data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.

✓Data should be randomly selected from a population, where each item has an equal chance of being selected.

✓Sample sizes should be equal if at all possible.

How to run a z-test?

Running a Z test on your data requires five steps:

State the null hypothesis and alternate hypothesis.

Choose an alpha level.

Find the critical value of z in a z table.

Calculate the z test statistic (see below).

Compare the test statistic to the critical z value and decide if you should support or reject the null hypothesis.

Two Proportion z-test

A company is testing two flu drugs A and B. Drug A works on 41 people out of a sample of 195. Drug B works on 351 people in a sample of 605. Are the two drugs comparable? Use a 5% alpha level.

Two Proportion z-test for Drug Efficiency