Subject Name: PROBABILTY&STATASTICS

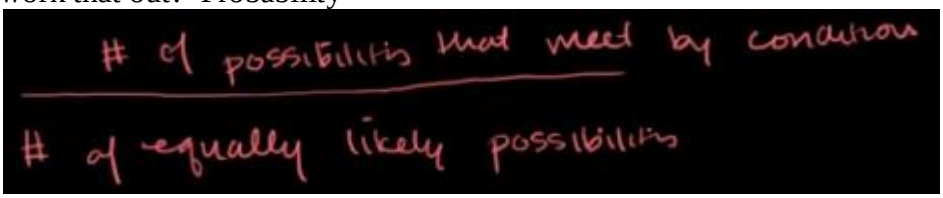
Prepared by (Faculty (s) Name): D. RUPALAKSHMI

Year and Sem, Department: II & I H&S

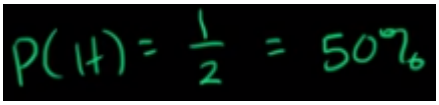
Unit-I: (Title)
BASIC PROBABITY

Important points / Definitions: (Minimum 15 to 20 points covering complete topics in that unit)

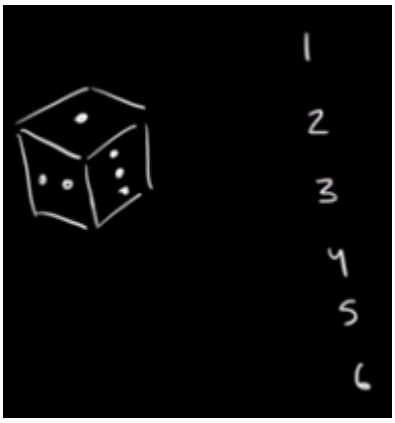
Probability is simply how likely something is to happen.
Whenever we're unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are. The analysis of events governed by probability is called statistics.
View all of Khan Academy's lessons and practice exercises on probability and statistics.
The best example for understanding probability is flipping a coin:
There are two possible outcomes—heads or tails.
What's the probability of the coin landing on Heads? We can find out using the equation P(H)=? P(H) = ? P(H)=?P, left parenthesis, H, right parenthesis, equals, question mark.You might intuitively know that the likelihood is half/half, or 50%. But how do we work that out? Probability =



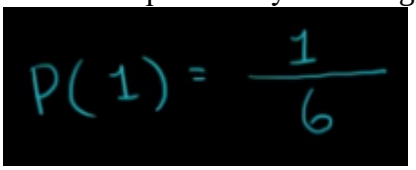
Formula for calculating the probability of certain outcomes for an event
In this case:



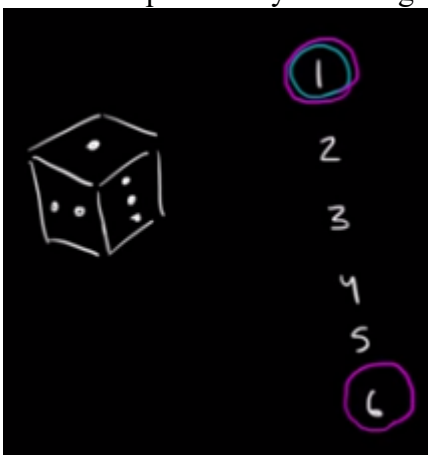
Probability of a coin landing on heads
Probability of an event = (# of ways it can happen) / (total number of outcomes)
P(A) = (# of ways A can happen) / (Total number of outcomes)
Example 1
There are six different outcomes.



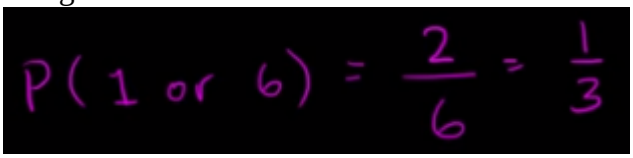
Different outcomes rolling a die
What's the probability of rolling a one?



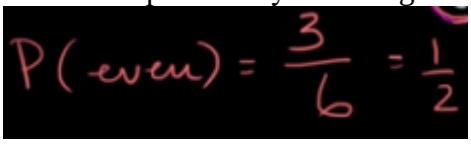
Probability formula for rolling a '1' on a die
What's the probability of rolling a one or a six?



Probability of a 1 or a 6 outcome when rolling a die
Using the formula from above:



Probability formula applied
What's the probability of rolling an even number (i.e., rolling a two, four or a six)?



Probability of rolling an even number? The formula and solution

Tips

- The probability of an event can only be between 0 and 1 and can also be written as a percentage.

- The probability of event A A AA is often written as P(A) P(A) P(A)P, left parenthesis, A, right parenthesis.
- If P(A)>P(B) P(A) > P(B) P(A)>P(B)P, left parenthesis, A, right parenthesis, is greater than, P, left parenthesis, B, right parenthesis, then event A A AA has a higher chance of occurring than event B B BB.
- If P(A)=P(B) P(A) = P(B) P(A)=P(B)P, left parenthesis, A, right parenthesis, equals, P, left parenthesis, B, right parenthesis, then events A A AA and B B BB are equally likely to occur.

Probability distributions are used in many fields but rarely do we explain what they are. Often it is assumed that the reader already knows (I assume this more than I should). So I'm going to try to explain what they are in this post.

What is a probability distribution?

Recall that a **random variable** is a variable whose value is the outcome of a random event (see the first introductory post for a refresher if this doesn't make any sense to you). For example, a random variable could be the outcome of the roll of a die or the flip of a coin.

*A **probability distribution** is a list of all of the possible outcomes of a random variable along with their corresponding probability values.*

To give a concrete example, here is the probability distribution of a fair 6-sided die.

Outcome of die roll	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

The probability distribution for a fair six-sided die

To be explicit, this is an example of a **discrete univariate probability distribution with finite support**. That's a bit of a mouthful, so let's try to break that statement down and understand it.

Discrete = This means that if I pick any two consecutive outcomes. I can't get an outcome that's in between. For example, if we consider 1 and 2 as outcomes of rolling a six-sided die, then I can't have an outcome in between that (e.g. I can have an outcome of 1.5). In mathematics, we would say that the list of outcomes is countable (but let's not go down the path of defining and understanding countable and uncountable sets. It gets weird). *You can probably guess when we get to continuous probability distributions this is no longer the case.*

Univariate = means that we only have one (random) variable. In this case, we only have the outcome of the die roll. In contrast, if we have more than one variable then we say that we have a *multivariate distribution*. In the specific case where we have 2 variables, we often say that it's a *bivariate distribution*.

finite support = This means that there is a limited number of outcomes. The *support* is essentially the outcomes for which the probability distribution is defined. So the support in our example is. 1, 2, 3, 4, 5 and 6. And since this is not an infinite number of values, it means that the support is finite.

Introduction to functions

Why are we talking about functions?

In the above example of rolling a six-sided die, there were only six possible outcomes so we could write down the entire probability distribution in a table. In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

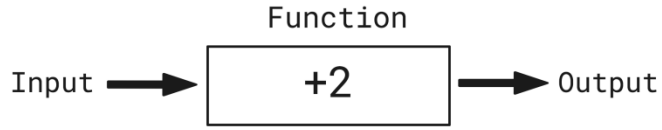
To get around the problem of writing a table for every distribution, we can define a function instead. *The function allows us to define a probability distribution succinctly*.

So let's first define what a function is generally and then we'll move onto functions used for probability distributions.

What is a function?

On a very abstract level, a function is a box that takes an input and returns an output. For the vast majority of cases, the function actually has to do something with the input for the output to be useful.

Let's define our own function. Let's say that this function takes in a number as input, adds 2 to the input number, and returns the new number as output. Graphically our function (as a box) looks like this:



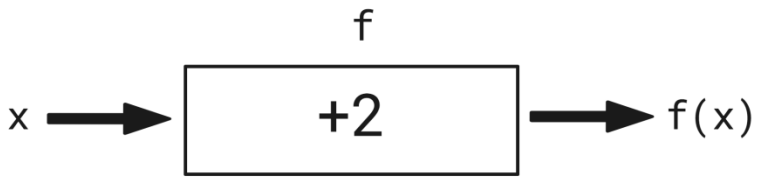
The abstract depiction of a function as a box that takes input and returns an output. In this case, the function adds 2 to the input.

So if our input was 5, our function would add 2 to it and return the output 5+2 = 7.

Notation for functions

Now it would be tedious to draw the digram above for every function that we want to create. Instead, we use symbols/letters to represent the diagram to make it more concise. Rather than

write the word “input” we use “x”, rather than write the word “function” we write “f” and rather than write the word “output” we write “f(x)”. So the above diagram can now be written as



Our function is written with symbols instead of words to make this more concise

This is better, however, we still have the problem that we have to draw a diagram to understand what the function is doing. We’re mathematicians though, we don’t want to waste precious energy by drawing a box, so we’ve come up with a better way of writing functions that means we don’t have to draw anything. We can define our function mathematically as

$$f(x) = x + 2$$

Our function is now written without having to draw arrows and boxes.

This is equivalent to the diagram above because we can see explicitly the input to the function f is x, we’ve called our function f and we know that the function adds 2 to the input and returns x + 2 as the output.

It’s important to note that the choice of letter for the function and the input was arbitrary. I could say that “a” is the input and I can call the function “add_two” so my function would be

$$\text{add\_two}(a) = a + 2$$

Another way of writing the same function

and this is completely equivalent to the function above.

One of the main takeaways from this is that with a function we can see how we would transform any input. With the function f(x) = x+ 2, We’d know what to do if the input was x=10 or if the input was x=10000. So we don’t need to write down a table like we did earlier in the post.

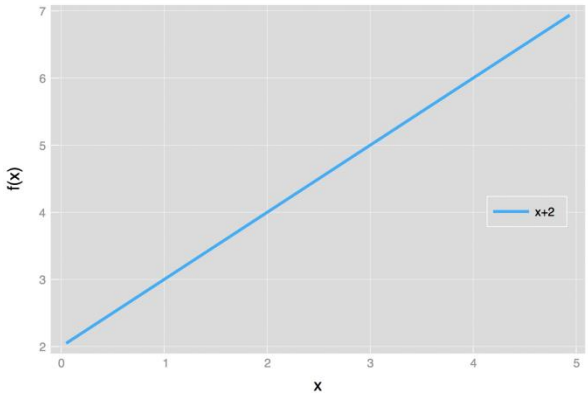
The final point I want to make here is that the functions that we’re going to use are solely going to work with numbers as both input and output. However, functions can take anything you like as an input and output anything you like (even output nothing). For example, we could write a function in a programming language that takes a string of text as input and outputs the first letter of that string. Here is an example of this function in the Python programming language

```
In [1]: def get_first_letter(my_string):
   ...:     return my_string[0]
   ...:
   ...:

In [2]: get_first_letter('Hello World')
Out[2]: 'H'
```

Representing functions graphically

Given that one of the main benefits of functions is to allow us to know how to transform any input, we can also use this knowledge to visualise the function explicitly. Let's stick with our example f(x) = x+2. Graphically it looks like this:



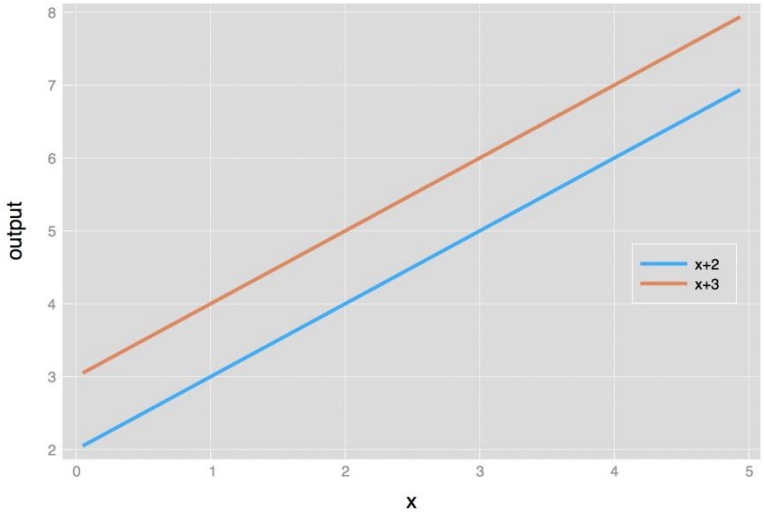
Graphical representation of our function f(x) = x + 2

We can read along the horizontal axis on the bottom as our input numbers and the corresponding numbers on the vertical axis on the left are the output values f(x) = x+2. For example, we can see that the blue line that represents the function crosses the point where the vertical (white) line at x=1 intersects the horizontal (white) line f(x) = 3. This graphically shows that f(1) = 1 + 2 = 3.

Parameters of functions

One of the most important features of functions are parameters. Parameters are the numbers that you find inside the functions that you don't necessarily feed in as input. In our example, f(x) = x + 2, the number "2" is a parameter because we need it to define the function but we don't include it as input to the function.

The reason that parameters are important is that they play a direct role in determining the output. For example, let's define another function h(x) = x+3. The only difference between the function f(x) = x+2 and our new function h(x) = x+3 is the value of the parameter (we now have a "3" instead of a "2"). This difference means that the outputs we get are completely different for the same input. Let's look at this graphically.



The difference between our functions f(x) = x + 2 and h(x) = x + 3

Parameters are arguably the most important feature of a probability (distribution) function because they define the output of the function which tells us the likelihood of certain outcomes in a random process. It's often parameters that we're trying to estimate in problems that come up in data science and I've previously written about 2 methods that we can use to estimate them: maximum likelihood estimation and Bayesian inference.

Now we're ready to talk about probability distributions using the language of functions.

Probability mass functions: Discrete probability distributions

When we use a probability function to describe a discrete probability distribution we call it a **probability mass function** (commonly abbreviated as pmf).

Remember from the first introductory post on probability concepts that the probability of a random variable, which we denote with a capital letter, X, taking on a value, denoted with a lowercase letter, x, is written as P(X=x). So if we use the dice roll as our example random variable, we can write the probability of the die landing on the number 3 as P(X=3) = 1/6.

A probability mass function, which we'll call "f" returns the probability of an outcome. Therefore, a probability mass function is written as:

$$f(x) = P(X = x)$$

I know this is getting a little horrible and mathematical but bear with me. The equation that we see above says that the probability mass function "f" just returns the probability of the outcome x.

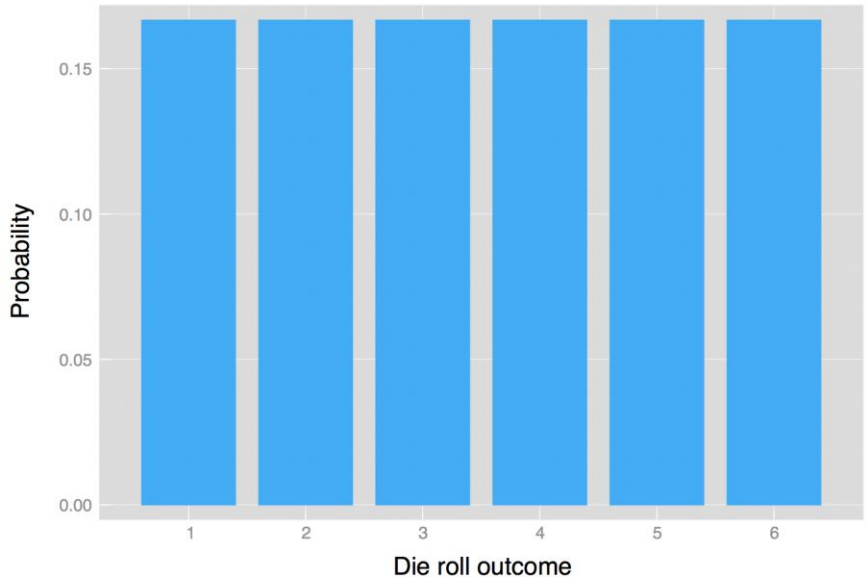
So let's go back the example of the fair 6-sided die (you're probably already sick of this example). The probability mass function, f, just returns the probability of the outcome. Therefore the probability of rolling a 3 is f(3) = 1/6. That's it.

Since a probability mass function returns probabilities it must obey the rules of probability (the axioms) that I described in my previous post. Namely, the probability mass function outputs values between 0 and 1 inclusive and the sum of the probability mass function (pmf) over all outcomes is equal to 1. Mathematically we can write these two conditions as

$$0 \leq f(x) \leq 1$$

$$\sum_i f(x_i) = f(x_1) + f(x_2) + \ldots = 1$$

So we've seen that we can write a discrete probability distribution as a table and as a function. We can also represent the die roll example graphically



Graphical representation of the probability distribution for outcomes of rolling a fair six-sided die

Example discrete probability distribution: The Bernoulli distribution

Some probability distributions crop up so often that they have been extensively studied and have names. One discrete distribution that crops up a lot is called the Bernoulli distribution. It describes the probability distribution of a process that has two possible outcomes. An example of this is a coin toss where the outcome is heads or tails.

The probability mass function of a Bernoulli distribution is

$$f(x) = p^x(1-p)^{1-x}$$

Here, x represents the outcome and takes the value 1 or 0. So we could say that heads = 1 and tails = 0. p is a parameter that represents the probability of the outcome being 1. So in the case of a fair coin where the probability of landed heads or tails is 0.5 we would set p = 0.5.

Often we want to be explicit about the parameters that are included in the probability mass function so we write

$$f(x;p) = p^x(1-p)^{1-x}$$

Notice that we use the semicolon to separate the input variables from the parameters.

Probability density functions: Continuous probability distributions

Sometimes we are concerned with the probabilities of random variables that have continuous outcomes. Examples include the height of an adult picked at random from a population or the amount of time that a taxi driver has to wait before their next job. For these examples, the random variable is better described by a continuous probability distribution.

When we use a probability function to describe a continuous probability distribution we call it a **probability density function** (commonly abbreviated as pdf).

Probability density functions are slightly more complicated conceptually than probability mass functions but don't worry, we'll get there. I think it'll be easiest to start with an example of a continuous probability distribution and then discuss the properties from there.

Example continuous probability distribution: The Normal distribution

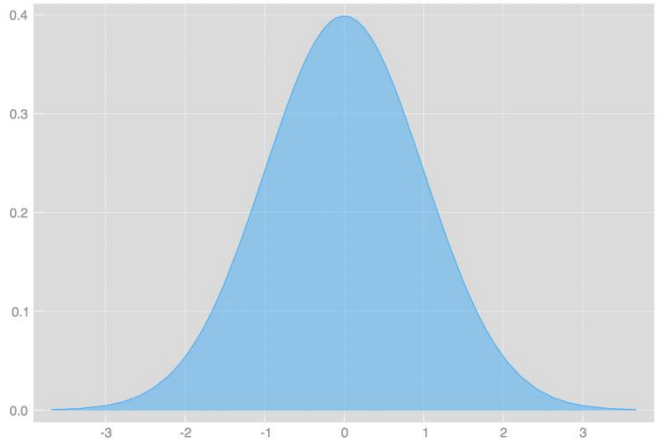
The normal distribution is probably the most common distribution in all of probability and statistics. One of the main reasons it crops up so much is due to the Central Limit Theorem. We're not going to go into it in this post but here is a nice article by Carson Forter called "The Only Theorem Data Scientists Need To Know" that explains what the theorem is and how it relates to the normal distribution.

The probability density function for the normal distribution is defined as

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Where the parameters (i.e. the symbols after the semicolon) represent the mean, μ, (the point where the centre of the distribution is) and the standard deviation, σ, (how spread out the distribution is) of the population.

If we set the mean to be equal to zero (μ=0) and the standard deviation equal to 1 (σ=1) then the distribution we get looks like this



Normal distribution with mean = 0 and standard deviation equal to 1

The normal distribution is an example of a *continuous univariate probability distribution with infinite support*. By infinite support, I mean that we can calculate values of the probability density function for all outcomes between minus infinity and positive infinity. In mathematics, we sometimes say that it's supported on the *whole real line.*

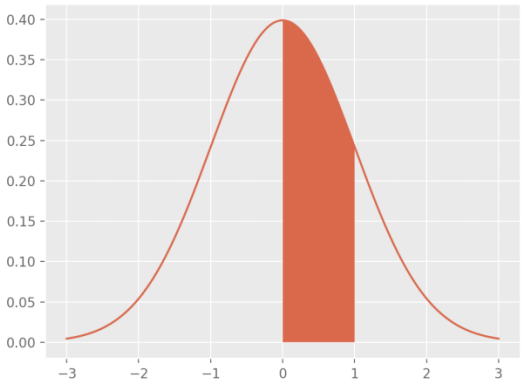
Properties of a continuous probability distribution

The first thing to notice is that the numbers on the vertical axis start at zero and go up. This is a rule that a probability density function has to obey. Any output value from a probability density function is greater than or equal to zero. In mathematical lingo we would say that the output is non-negative or write this mathematically as

$$f(x) \geq 0$$

However, unlike probability mass functions, the output of a probability density function is **not a probability value.** This is an incredibly important distinction, one of which I've been guilty of forgetting.

To get the probability from a probability density function we need to find the area under the curve. So from our example distribution with mean = 3 and standard deviation = 1, we can find the probability that the outcome is between 0 and 1 by finding the area shown in the image below



The area shaded is the probability of the outcome being between 0 and 1.

Mathematically we would write this as

$$\int_0^1 f(x;\mu,\sigma)\mathrm{d}x = P(0 < X < 1)$$

We can read this as "*the integral of the probability density function between 0 and 1 (on the left-hand side) is equal to the probability that the outcome of the random variable is between zero and 1 (on the right-hand side)*".

Forgive me as I haven't explicitly covered integrals and how they work (I have a brief conceptual introduction to integrals in my post on marginalisation but it doesn't teach you how to compute them). If you don't know about them then all you need to know for the moment is that it's a mathematical method for finding the area under a curve, which in this

case gives us the probabilities of outcomes. Perhaps I need to write a brief series covering introductory calculus.

We've now seen another property of probability density functions. Namely that the probability between two outcomes, let's say 'a' and 'b', is the integral of the probability density function between those two points (this is equivalent to finding the area under the curve produced by the probability density function between the points 'a' and 'b'). Mathematically this is

$$\int_a^b f(x)\mathrm{d}x = P(a < X < b)$$

Remember that we still have to follow the rules of probability distributions, namely the rule that says that the sum of all possible outcomes is equal to 1. We can cover all possible values if we set our range from 'minus infinity' all the way to 'positive infinity'. Therefore the following has to be true for the function to be a probability density function

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

This says that the area under the curve between minus infinity and positive infinity is equal to 1.

An important thing to know about continuous probability distributions (and something that may be really weird to come to terms with conceptually) is that the probability of the random variable being equal to a specific outcome is 0. For example, if we try to get the probability that the outcome is equal to the number 2 we would get

$$\int_2^2 f(x)\mathrm{d}x = P(2 < X < 2) = 0$$

This may seem weird conceptually but if you understand calculus then it should make a little more sense. I'm not going to cover calculus in this post. Instead, what I want you to take away from this fact is that we can only talk about probabilities occurring between two values. Or we can ask about the probability of an outcome being greater than or less than a specific value. We **can't** ask about the probability of an outcome being equal to a specific value.

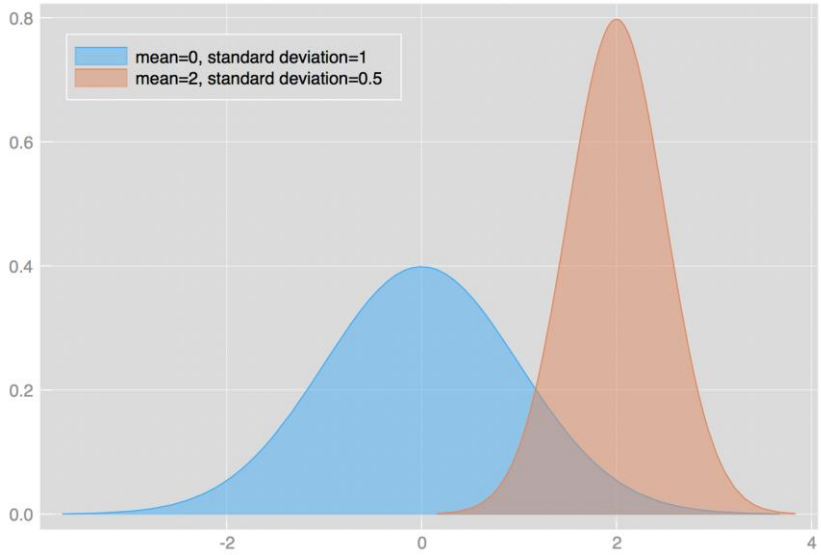
The eagle-eyed readers may have noticed that I've used "less than (<)" and "greater than (>)" symbols rather than "less than or equal to (≤)" and "greater than or equal to (≥)" symbols. For continuous probability distributions, it doesn't actually matter because they're the same. Explicitly I mean

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

So the probability of the random variable taking on a value between a and b exclusive is the same as the probability of it taking on a value between a and b inclusive.

The importance of parameters

We saw that parameter values can change the output values of a function and it's no different with probability distributions.



Two normal distributions with different parameters give completely different probability outcomes.

In the figure above we've plotted the probability density functions of two normal distributions. The blue distribution has parameter values μ=0 and σ=1 whereas the red distribution has parameter values μ=2 and σ=0.5.

It's clearer to see now why using the wrong parameter values can give results that differ wildly from what you might expect.

Summary

Wow! That was much longer than I intended. Let's summarise the main points:

- A **probability distribution** is a list of outcomes and their associated probabilities.
- We can write small distributions with tables but it's easier to summarise large distributions with functions.

- A function that represents a discrete probability distribution is called a **probability mass function**.
- A function that represents a continuous probability distribution is called a **probability density function.**
- Functions that represent probability distributions still have to obey the rules of probability
- The output of a *probability mass function is a probability* whereas *the area under the curve produced by a probability density function represents a probability*.
- Parameters of a probability function play a central role in defining the probabilities of the outcomes of a random variable.

Short Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1)Determine the probability for each of the following events .A non defective bolt will be found if out of 600 bolts already examined 12 were defective (dec2005,apr 2006)

2). write the axioms of probability(2009,2010,2014)

3)A box contains " n " tickets marked 1 through n. 2 tickets are drawn in succession without replacement . Determine the probability that the number on the tickets are consecutive integers(2004,2008)

4)Define sample space& simple event(2007)

5).A fair die is tossed twice . find the probability of getting a 4,5&6 on the first toss and 1,2,3&4 on the second toss(2009)

6).Define Bye's theorem or state and prove Bay's theorem(2007,2008,2009,2010)

7). If x is random variable & k is constant then E(X+K) =E(X)+K (2005,20067,2009)

8).If X is random variable& K is constant then E(X+Y)=E(X)+E(Y).provided E(X)&E(Y) exists (2009)

9).A random variable X has the following probability distribution(2008)

x	1	2	3	4	5	6	7	8
P(x)	K	2k	3k	4k	5k	6k	7k	8k

find the value of i).k ii).p(x≤2) iii).p(2≤X≤5)

10).X is a continuous random variable with p.d.f f(x)={ $x^{2,}$ 0≤x≤1
0, elsewhere if p (a≤x≤1)=19/81
.find the value 'a' (2004,2008)

Long Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Two marbles are drawn in succession from a box containing 10 red, 30white ,20 blue and 15 orange marbles, with replacement being made after each draw .find the probability that
i. Both are white
ii .First is red and second is white(2005,2006,2007,nov 2008,nov2010)
2. In a bolt factory machines A ,B,C manufacture 20%,30% and 50% of the total of the output,35 and 6%,3%and2% are defective .a bolt is drawn at a random and found to be defective . find the probabilities and that it is manufactured from
i. Machine A ii .Machine B iii. Machine C (2007,apr 2012)
3. Two dice are thrown . Let X assign to each point (a b) in S the maximum of its numbers i.e X(a b)= max(a b) .find the probability distribution. X is a random variable with X(s)={1,2,3,4,5,6}. also find the mean and variance of the distribution(2004,200,2008,2009,nov2010).
4. A random variable X has the following probability function

x	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	$K^{2,}$	$2k^2$	$7k^2+k$

i. Determine K ii. Evaluate p(X<6),p(X≥6),p(0<X<5)and p(0≤X≤6) iii. if p(X≤k)>1/2,find the minimum value of K and(iv). Determine the distribution function X (v).Mean (vi) Variance (nov 2011,dec2011,nov2012)
5. Find the mean& variance of the uniform probability distribution by f(x)= 1/n for x= 1,2,3,4.....n(2001,nov2009,nov2010,dec2011)
6. A continuous random variable has the probability density function
f(x)= { $kxe^{-\lambda x}$, x≥0,λ>0
{ 0 ,other wise
Determine(i).k (ii)mean(iii)variance (dec2009,nov2010,dec2011,may2011,nov2012, may2013)
7.The probability density f(x)of a continuous random variable is given by f(x)= $ce^{-IXI,}$ - ∞<x<∞ ,show that c=1/2 and find the mean and variance of distribution . Also find the probability that the variance lies between 0 and 4(jan 2007,nov2010,may 2013)
8.a)A probability density function of a random variable X is f(x)={1/2 sin x ,for 0≤x≤π,:0 elsewhere Find the mean ,mode and median of the distribution and also find the probability between 0 and π/2
b)If X is a continuous random variable and Y= ax+b ,prove that E(Y)= aE(x)+b and $V(Y)=a^2v(x)$ where V stands for variance & a,b are constants(2004,2008,nov2010)
9.calculate the first four moments of the following distribution about an arbitary origin

Class interval	60-62	63-65	66-68	69-71	72-74
Frequency	5	18	42	27	8

also find the moments about mean(nov2007,2009)
10.Compute the first 4 moments about the mean for the following distribution(nov2010)

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No of students	8	12	20	30	15	10	5

Also evaluate$\beta_1$andβ_2

Fill in the Blanks / Choose the Best: (Minimum 10 to 15 with Answers)
1. A...Random..... variable takes numerical value which is determined by the result of the random experiment
2.The binomial and poison distributions are the examples of a....Discrete..... probability distribution

3.the continuous probability distribution is a probability distribution of a .....continuous....random variable.
4. if x a continuous random variable and y=a+bx then the expected value of y=.aE(x)+b.....
5. The poison distribution for a discrete random variable x is=...$f(x,\lambda)=\lambda^2.e^{-\lambda}/x$.....
6.The mean of random variable can be ...Either positive or negative.....
7.Conditional probability of an event A given in event B is.....P(A∩B.)/P(B).....
8.If the random variable X has a standard normal distribution, then P(x) sample space f>=0)is...0.5.....
9.....Event.....is known as subset of sample space S
10.Probability space is known as...probability triple.....

Unit-II: (Title)

DISCRETE PROBABILITY DSTRIBUTION

Important points / Definitions: (Minimum 15 to 20 points covering complete topics in that unit)

Binomial Probability Distribution

To understand binomial distributions and binomial probability, it helps to understand binomial experiments and some associated notation; so we cover those topics first.

Binomial Experiment

A **binomial experiment** is a statistical experiment that has the following properties:

- The experiment consists of *n* repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by *P*, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant - 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation

The following notation is helpful, when we talk about binomial probability.

- *x*: The number of successes that result from the binomial experiment.
- *n*: The number of trials in the binomial experiment.
- *P*: The probability of success on an individual trial.
- *Q*: The probability of failure on an individual trial. (This is equal to 1 - *P*.)
- *n!*: The factorial of n (also known as n factorial).
- b(*x*; *n*, *P*): Binomial probability - the probability that an *n*-trial binomial experiment results in exactly *x* successes, when the probability of success on an individual trial is *P*.
- ${}_nC_r$: The number of combinations of *n* things, taken *r* at a time.

Binomial Distribution

A **binomial random variable** is the number of successes *x* in *n* repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a **binomial distribution**.

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The binomial distribution has the following properties:

- The mean of the distribution (μ_x) is equal to $n * P$.
- The variance (σ^2_x) is $n * P * (1 - P)$.
- The standard deviation (σ_x) is sqrt[$n * P * (1 - P)$].

Binomial Formula and Binomial Probability

The **binomial probability** refers to the probability that a binomial experiment results in exactly *x* successes. For example, in the above table, we see that the binomial probability of getting exactly one head in two coin flips is 0.50.

Given *x*, *n*, and *P*, we can compute the binomial probability based on the binomial formula:

Binomial Formula. Suppose a binomial experiment consists of *n* trials and results in *x* successes. If the probability of success on an individual trial is *P*, then the binomial probability is:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n - x}$$

or

$$b(x; n, P) = \{ n! / [x! (n - x)!] \} * P^x * (1 - P)^{n - x}$$

Example 1

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution: This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

$$b(2; 5, 0.167) = {}_5C_2 * (0.167)^2 * (0.833)^3$$
$$b(2; 5, 0.167) = 0.161$$

Cumulative Binomial Probability

A **cumulative binomial probability** refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

For example, we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

$$b(x \leq 45; 100, 0.5) =$$
$$b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + ... + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5)$$

Binomial Calculator

As you may have noticed, the binomial formula requires many time-consuming computations. The Binomial Calculator can do this work for you - quickly, easily, and error-free. Use the Binomial Calculator to compute binomial probabilities and cumulative binomial probabilities. The calculator is free. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

Binomial Calculator

Example 2

What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?

Solution: To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$$b(x \leq 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \ldots + b(x = 45; 100, 0.5)$$
$$b(x \leq 45; 100, 0.5) = 0.184$$

Example 3

The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

Solution: To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

$$b(x \leq 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3)$$
$$b(x \leq 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087$$
$$b(x \leq 2; 5, 0.3) = 0.8369$$

Example 4

What is the probability that the world series will last 4 games? 5 games? 6 games? 7 games? Assume that the teams are evenly matched.

Solution: This is a very tricky application of the binomial distribution. If you can follow the logic of this solution, you have a good understanding of the material covered in the tutorial, to this point.

In the world series, there are two baseball teams. The series ends when the winning team wins 4 games. Therefore, we define a success as a win by the team that ultimately becomes the world series champion.

For the purpose of this analysis, we assume that the teams are evenly matched. Therefore, the probability that a particular team wins a particular game is 0.5.

Let's look first at the simplest case. What is the probability that the series lasts only 4 games. This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

$$b(4; 4, 0.5) = {}_4C_4 * (0.5)^4 * (0.5)^0 = 0.0625$$

Similarly, when we compute the probability of the American League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be 0.0625 + 0.0625 = 0.125; since the series would end if either the American or National League team won 4 games in a row.

Now let's tackle the question of finding probability that the world series ends in 5 games. The trick in finding this solution is to recognize that the series can only end in 5 games, if one team has won 3 out of the first 4 games. So let's first find the probability that the American League team wins exactly 3 of the first 4 games.

$$b(3; 4, 0.5) = {}_4C_3 * (0.5)^3 * (0.5)^1 = 0.25$$

Okay, here comes some more tricky stuff, so listen up. Given that the American League team has won 3 of the first 4 games, the American League team has a 50/50 chance of winning the fifth game

to end the series. Therefore, the probability of the American League team winning the series in 5 games is 0.25 * 0.50 = 0.125. Since the National League team could also win the series in 5 games, the probability that the series ends in 5 games would be 0.125 + 0.125 = 0.25.

The rest of the problem would be solved in the same way. You should find that the probability of the series ending in 6 games is 0.3125; and the probability of the series ending in 7 games is also 0.3125.

Poisson Distribution

A Poisson distribution is the probability distribution that results from a Poisson experiment.

Attributes of a Poisson Experiment

A **Poisson experiment** is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- *e*: A constant equal to approximately 2.71828. (Actually, *e* is the base of the natural logarithm system.)
- μ: The mean number of successes that occur in a specified region.
- *x*: The actual number of successes that occur in a specified region.
- P(*x*; μ): The **Poisson probability** that exactly *x* successes occur in a Poisson experiment, when the mean number of successes is μ.

Poisson Distribution

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ. Then, the Poisson probability is:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

where *x* is the actual number of successes that result from the experiment, and *e* is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ .
- The variance is also equal to μ .

Poisson Distribution Example

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- μ = 2; since 2 homes are sold per day, on average.
- x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

$$P(3; 2) = (2.71828^{-2}) (2^{3}) / 3!$$

$$P(3; 2) = (0.13534) (8) / 6$$

$$P(3; 2) = 0.180$$

Thus, the probability of selling 3 homes tomorrow is 0.180 .

Poisson Calculator

Clearly, the Poisson formula requires many time-consuming computations. The Stat Trek Poisson Calculator can do this work for you - quickly, easily, and error-free. Use the Poisson Calculator to compute Poisson probabilities and cumulative Poisson probabilities. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

Poisson Calculator

Cumulative Poisson Probability

A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

Cumulative Poisson Example

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 5$; since 5 lions are seen per safari, on average.
- $x = 0, 1, 2,$ or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- $e = 2.71828$; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5). To compute this sum, we use the Poisson formula:

$$P(x \leq 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$$

$$P(x \leq 3, 5) = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$$

$$P(x \leq 3, 5) = [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6]$$

$$P(x \leq 3, 5) = [0.0067] + [0.03369] + [0.084224] + [0.140375]$$

$$P(x \leq 3, 5) = 0.2650$$

Thus, the probability of seeing at no more than 3 lions is 0.2650.

Short Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Define the binomial distribution(2013,2014)
2. State and prove mean of binomial distribution(2000,2013,2014)
3. The probability of a defective bolt 1/8 find (i) The mean (ii) The variance for the distribution of defective bolts of 640 (2004)
4. In 256 sets of 12 tosses of a coin ,in how many cases one can expect 8 heads and 4 tails(2003,2004)
5. The mean and variance of a binomial distribution are2& 8/5. Find n (Nov2015)

6. in eight throws of a die '5' (or) 6 is considered a success. Find the mean number of success and the S.D. (2004)
7. Traffic control engineer reports that 75% of the vehicles passing through a check post are from within state. What is the probability that fewer than '4' of the 9 are from out of the state. (2010 Nov)
8. State & Prove mean of the poison distribution (2008, 2009, 2012, 2013, 2014).
9. Derive the central moment of poison distribution. (May 2013)
10. Average number of accidents on any day on a national high way is 1.8. Determine the probability that the number of accidents (i) At least one (ii) At most one.

Long Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Derive the constants of Binomial distribution (i.e Mean of Binomial distribution& variance of B.D) (2000,2002,dec2013,mar2014,dec2014)
2. Ten coins are thrown simultaneously. Find the probability of getting at least
i) Seven heads ii)Six heads iii) One head (1999,2007,2008,2009,2013)
3). 20% of items produced from a factory are defective. find the probability that in a sample of 5 chosen at random i) none is defective ii) one is defective iii) $p(1<x<4)$
4.a)The mean of Binomial distribution is 3 and the variance is 9/4 .Find
i) the value of 'n" ii) $p(x\geq 7)$ iii) $p(1\leq x<6)$ (2007,2008,dec2009)
b) The probability that John hits a target is 1/2. He fires 6 times . Find the probability that he hits the target (i).exactly ii) more than 4 times iii). at least once (207,may 2011)
5.Fit a binomial distribution to the following data(2004,nov2011)

x	0	1	2	3	4	5
f	2	14	20	34	22	8

6.Four coins are tossed 160 times. The number of times X heads occur is given bellow(mar2004)

X	0	1	2	3	4
No of times	8	34	69	43	6

fit a Binomial distribution to this data on the hypothesis that coins are unbiased
7. Derive the mean & variance of poison distribution(may 2008 dec2009,2012,2013,204)
8.a)If a random variable has a poison distribution such that p(1)=p(2) find i) mean of distribution
ii) p(4) iii) $p(x\geq 1)$ iv) $p(1<x<4)$(dec2007)
b) If X is a poison variate such that $3p(x=4)=1/2\ p(x=2)+p(x=0)$ find i) The mean of x ii) $p(x\leq 2)$ (may2008,nov2010,may 2011)
9. a). given that $p(x=2)=9(x=4)+90\ p(x=6)$ for a poison variate X find
i). $p(x<2)$ ii)$p(x>4)$ iii) $p(x\geq 1)$(may2011)
b) If the mean of a poison variable is 1.8 than find
i) $p(x>1)$ ii) $p(x=5)$ iii)$p(0<x<5)$(dec2011)

10.a) Fit a poison distribution to the following data(2009,may2013)

X	0	1	2	3	4	5	Total
f	142	156	169	27	5	1	400

b).Fit a poison distribution for the following data and calculate the expected frequencies

x	0	1	2	3	4

F(x)	109	65	22	3	1

Fill in the Blanks / Choose the Best: (Minimum 10 to 15 with Answers)

1.he binomial distribution is..$n_{cr}.p^rq^{n-r}$.....
2.The binomial density function is given by....np.....
3. variance of binomial distribution is....npq.....
4. Mode of binomial distribution is ...λ.....
5. variance of binomial distribution is.....λ.....
6. Discrete distributions are of .....Two.....types.
7. ....Infinite..... number of occurrences of the event must be possible in the interval
8.In any extremely small portion of the interval , the probability of two or more occurrences of the event is.....Negligible.....
9. If the domain of the function is discrete for a discrete random varible , then the probability function is called as ..Discrete probability function.....
10.The occurrence of events in probability is...independent.....

Unit-III: (Title)
CONTINUOUS RANDOM VARIABLES & DISTRIBUTIONS

Important points / Definitions: (Minimum 15 to 20 points covering complete topics in that unit)

All probability distributions can be classified as discrete probability distributions or as continuous probability distributions, depending on whether they define probabilities associated with discrete variables or continuous variables.

Discrete vs. Continuous Variables

If a variable can take on any value between two specified values, it is called a **continuous variable**; otherwise, it is called a **discrete variable**.

Some examples will clarify the difference between discrete and continuous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Just like variables, probability distributions can be classified as discrete or continuous.

Discrete Probability Distributions

If a random variable is a discrete variable, its probability distribution is called a **discrete probability distribution**.

An example will make this clear. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents a *discrete* probability distribution because it relates each value of a discrete random variable with its probability of occurrence. On this website, we will cover the following discrete probability distributions.

- Binomial probability distribution
- Hyper geometric probability distribution
- Multinomial probability distribution
- Negative binomial distribution
- Poisson probability distribution

Note: With a discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution can always be presented in tabular form.

Continuous Probability Distributions

If a random variable is a continuous variable, its probability distribution is called a **continuous probability distribution**.

A continuous probability distribution differs from a discrete probability distribution in several ways.

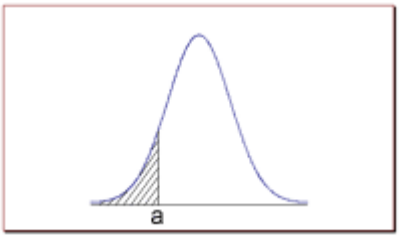
- The probability that a continuous random variable will assume a particular value is zero.
- As a result, a continuous probability distribution cannot be expressed in tabular form.
- Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a **probability density function**. Sometimes, it is referred to as a **density function**, a **PDF**, or a

pdf. For a continuous probability distribution, the density function has the following properties:

- Since the continuous random variable is defined over a continuous range of values (called the **domain** of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between *a* and *b* is equal to the area under the density function bounded by *a* and *b*.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable *X* was less than or equal to *a*. The probability that *X* is less than or equal to *a* is equal to the area under the curve bounded by *a* and minus infinity - as indicated by the shaded area.



Note: The shaded area in the graph represents the probability that the random variable *X* is less than or equal to *a*. This is a cumulative probability. However, the probability that *X* is *exactly* equal to *a* would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as *a*) is always zero.

On this website, we cover the following continuous probability distributions.

- Normal probability distribution
- Student's t distribution
- Chi-square distribution
- F distribution

The Normal Distribution

The **normal distribution** refers to a family of continuous probability distributions described by the normal equation.

The Normal Equation

The normal distribution is defined by the following equation:

The Normal Equation. The value of the random variable *Y* is:

$$Y = \{ 1/[\sigma * \text{sqrt}(2\pi)] \} * e^{-(x - \mu)2/2\sigma2}$$

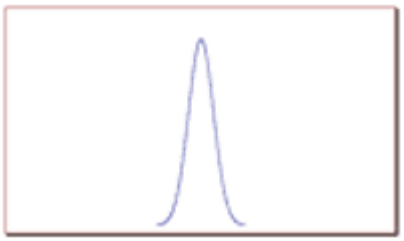
where X is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and e is approximately 2.71828.

The random variable X in the normal equation is called the **normal random variable**. The normal equation is the probability density function for the normal distribution.

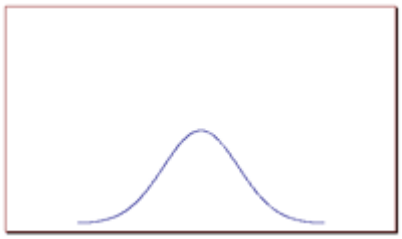
The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

Smaller standard deviation



Bigger standard deviation



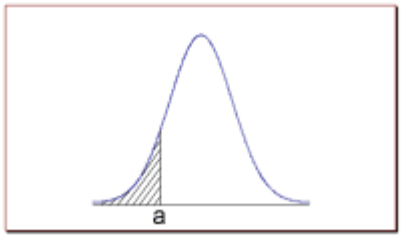
When the standard deviation is small, the curve is tall and narrow; and when the standard deviation is big, the curve is short and wide (see above)

Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the *non-shaded* area in the figure below).

- The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the *shaded* area in the figure below).



Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the **empirical rule** or the **68-95-99.7 rule**. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table. In the examples below, we illustrate the use of Stat Trek's Normal Distribution Calculator, a free tool available on this site. In the next lesson, we demonstrate the use of normal distribution tables.

Normal Distribution Calculator

The normal distribution calculator solves common statistical problems, based on the normal distribution. The calculator computes cumulative probabilities, based on three simple inputs. Simple instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

Normal Distribution Calculator

Test Your Understanding

Problem 1

An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

Solution: Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: P(X $\leq$ 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Problem 2

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution: Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

- To compute P(X < 110), we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that P(X < 110) is 0.84.
- To compute P(X < 90), we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that P(X < 90) is 0.16.

We use these findings to compute our final answer as follows:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$
$$P(90 < X < 110) = 0.84 - 0.16$$
$$P(90 < X < 110) = 0.68$$

Thus, about 68% of the test scores will fall between 90 and 110.

The **exponential distribution** is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

Example

Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes.

X is a **continuous random variable** since time is measured. It is given that $\mu = 4$ minutes. To do any calculations, you must know m, the decay parameter.

$m=1\mu$

. Therefore, $m=14=0.25$

The standard deviation, σ, is the same as the mean. $\mu = \sigma$

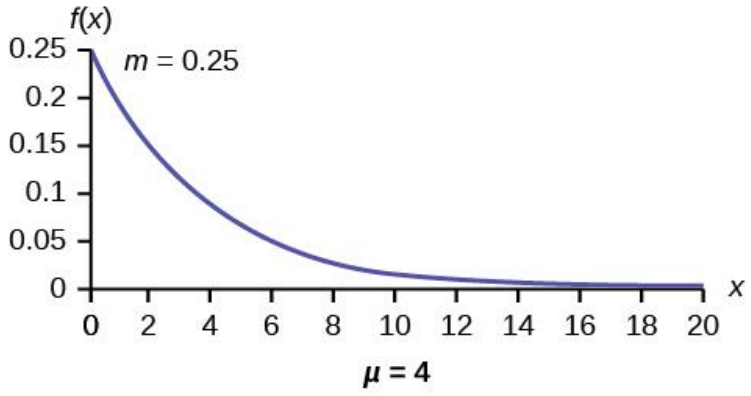
The distribution notation is $X \sim Exp(m)$. Therefore, $X \sim Exp(0.25)$.

The probability density function is $f(x) = me^{-mx}$. The number $e = 2.71828182846\ldots$ It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter one for x, the calculator will display the value e.

The curve is:

$f(x) = 0.25e^{-0.25x}$ where x is at least zero and $m = 0.25$.

For example, $f(5) = 0.25e^{-(0.25)(5)} = 0.072$. The postal clerk spends five minutes with the customers. The graph is as follows:



Notice the graph is a declining curve. When $x = 0$,

$f(x) = 0.25e^{(-0.25)(0)} = (0.25)(1) = 0.25 = m$. The maximum value on the y-axis is m.

Memorylessness of the Exponential Distribution

In example 1, recall that the amount of time between customers is exponentially distributed with a mean of two minutes ($X \sim Exp$ (0.5)). Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case–the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**. Specifically, the **memoryless property** says that

$P(X > r + t \mid X > r) = P(X > t)$ for all $r \geq 0$ and $t \geq 0$

For example, if five minutes has elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using $r = 5$ and $t = 1$ in the foregoing equation.

$P(X > 5 + 1 \mid X > 5) = P(X > 1) = e(–0.5)(1) \approx 0.6065$.

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

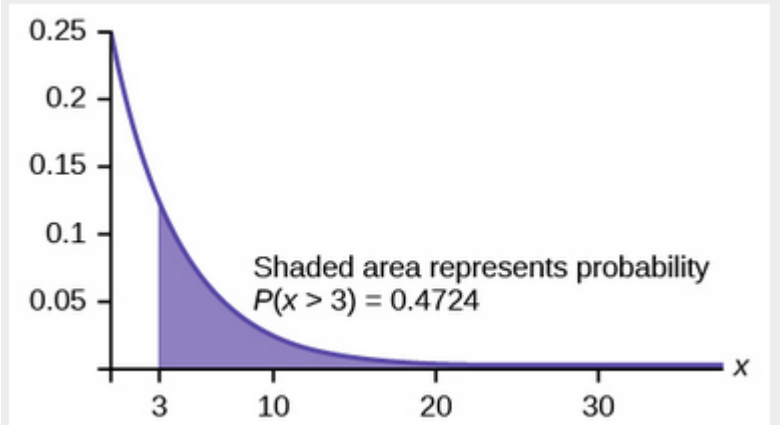
The exponential distribution is often used to model the longevity of an electrical or mechanical device. In example 1, the lifetime of a certain computer part has the exponential distribution with a mean of ten years ($X \sim Exp(0.1)$). The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is $P(X > 17|X > 10) = P(X > 7) = 0.4966$.

Example

Refer to example 1, where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

The decay parameter of X is m = 14 = 0.25, so X ~ Exp(0.25).
The cumulative distribution function is P(X < x) = 1 – e–0.25x. We want to find P(X > 7|X > 4). The memoryless property says that P(X > 7|X > 4) = P (X > 3), so we just need to find the

probability that a customer spends more than three minutes with a postal clerk.
This is $P(X > 3) = 1 – P(X < 3) = 1 – (1 – e–0.25·3) = e–0.75 ≈ 0.4724$.



Relationship between the Poisson and the Exponential Distribution

There is an interesting relationship between the exponential distribution and the Poisson distribution. Suppose that the time that elapses between two successive events follows the exponential distribution with a mean of μ units of time. Also assume that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a Poisson distribution with mean $\lambda = 1/\mu$. Recall that if X has the Poisson distribution with mean λ, then $P(X=k)=\lambda k e-\lambda k!$

. Conversely, if the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution. ($k!$ = k*(k-1*)(k–2)*(k-3)…3*2*1)

Example

At a police station in a large city, calls come in at an average rate of four calls per minute. Assume that the time that elapses from one call to the next has the exponential distribution. Take note that we are concerned only with the rate at which calls come in, and we are ignoring the time spent on the phone. We must also assume that the times spent between calls are independent. This means that a particularly long delay between two calls does not mean that there will be a shorter waiting period for the next call. We may then deduce that the total number of calls received during a time period has the Poisson distribution.

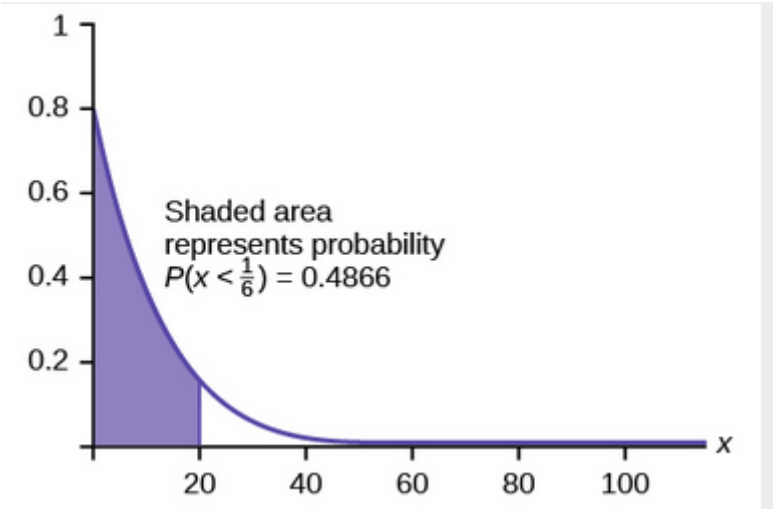
1. Find the average time between two successive calls.

2. Find the probability that after a call is received, the next call occurs in less than ten seconds.
3. Find the probability that exactly five calls occur within a minute.
4. Find the probability that less than five calls occur within a minute.
5. Find the probability that more than 40 calls occur in an eight-minute period.

Solutions:

1. On average there are four calls occur per minute, so 15 seconds, or 1560

- = 0.25 minutes occur between successive calls on average.
- Let T = time elapsed between calls. From part a, μ=0.25, so m = 10.25 = 4. Thus, T ~ Exp(4). The cumulative distribution function is $P(T < t) = 1 - e^{-4t}$. The probability that the next call occurs in less than ten seconds (ten seconds = 1/6 minute) is P(T < 16) = 1 – e–416≈0.4866



- Let X = the number of calls per minute. As previously stated, the number of calls per minute has a Poisson distribution, with a mean of four calls per minute. Therefore, $X \sim Poisson(4)$, and so $P(X = 5)$ = 45e–45!≈
- 0.1563. (5! = (5)(4)(3)(2)(1))
- Keep in mind that X must be a whole number, so $P(X < 5) = P(X \leq 4)$.

To compute this, we could take $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$. Using technology, we see that $P(X \leq 4) = 0.6288$.

- Let Y = the number of calls that occur during an eight minute period.

Since there is an average of four calls per minute, there is an average of (8)(4) = 32 calls during each eight minute period.
Hence, $Y \sim Poisson(32)$. Therefore, $P(Y > 40) = 1 - P(Y \leq 40) = 1 - 0.9294 = 0.0707$.

Exponential: $X \sim Exp(m)$ where m = the decay parameter

- pdf: $f(x)$ = me–mx

- where $x \geq 0$ and $m > 0$
- cdf: $P(X \leq x)$ = 1 –e–mx

- • mean μ=1m
- • standard deviation $\sigma = \mu$
- percentile, k: k = ln(AreaToTheLeftOfK)−m
- • Additionally
  - $P(X > x) = e^{(-mx)}$
  - $P(a < X < b) = e^{(-ma)} - e^{(-mb)}$
- Memoryless Property: $P(X > x + k|X > x) = P(X > k)$
- Poisson probability: P(X=k)=λke−λk! with mean λ
- • $k! = k*(k-1)*(k-2)*(k-3)…3*2*1$

Gamma Distribution

Definition: Gamma distribution is a distribution that arises naturally in processes for which the waiting times between events are relevant. It can be thought of as a waiting time between Poisson distributed events.

Probability density function: The waiting time until the hth Poisson event with a rate of change λ is

$$P(x) = \frac{\lambda(\lambda x)^{h-1}}{(h-1)!}e^{-\lambda x}$$

For $X \sim \text{Gamma}(k, \theta)$, where $k = h$ and θ = 1 / λ, the gamma probability density function is given by

$$\frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}$$

where

- e is the natural number (e = 2.71828…)
- k is the number of occurrences of an event
- if k is a positive integer, then $\Gamma(k) = (k-1)!$ is the gamma function
- θ = 1 / λ is the mean number of events per time unit, where λ is the mean time between events. For example, if the mean time between phone calls is 2 hours, then you would use a gamma distribution with θ=1/2=0.5. If we want to find the mean number of calls in 5 hours, it would be 5 ×1/2=2.5.
- x is a random variable

Cumulative density function: The gamma cumulative distribution function is given by

$$\frac{\gamma(k, x/\theta)}{\Gamma(k)}$$

where

- if k is a positive integer, then $\Gamma(k) = (k-1)!$ is the gamma function
- $\gamma(k, x/\theta) = \int_0^{x/\theta} t^{k-1} e^{-t} dt$

Moment generating function: The gamma moment-generating function is

$$M(t) = (1 - \theta t)^{-k}$$

Expectation: The expected value of a gamma distributed random variable x is

$$E(X) = k\theta$$

Variance: The gamma variance is

$$Var(X) = k\theta^2$$

Applications

The gamma distribution can be used a range of disciplines including queuing models, climatology, and financial services. Examples of events that may be modeled by gamma distribution include:

- The amount of rainfall accumulated in a reservoir
- The size of loan defaults or aggregate insurance claims
- The flow of items through manufacturing and distribution processes
- The load on web servers
- The many and varied forms of telecom exchange

The gamma distribution is also used to model errors in a multi-level Poisson regression model because the combination of a Poisson distribution and a gamma distribution is a negative binomial distribution.

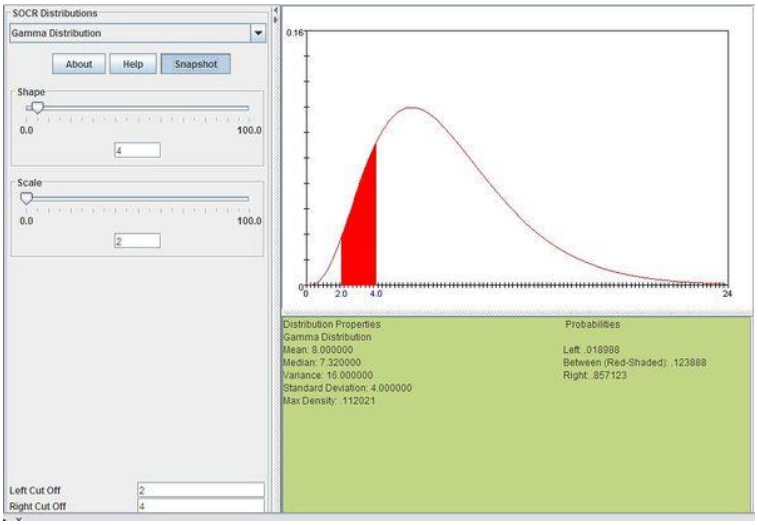
Example

Suppose you are fishing and you expect to get a fish once every 1/2 hour. Compute the probability that you will have to wait between 2 to 4 hours before you catch 4 fish.

One fish every 1/2 hour means we would expect to get θ = 1 / 0.5 = 2 fish every hour on average. Using θ = 2 and *k* = 4, we can compute this as follows:

$$P(2 \leq X \leq 4) = \sum_{x=2}^{4} \frac{x^{4-1} e^{-x/2}}{\Gamma(4) 2^4} = 0.12388$$

The figure below shows this result using SOCR distributions



Normal Approximation to Gamma distribution

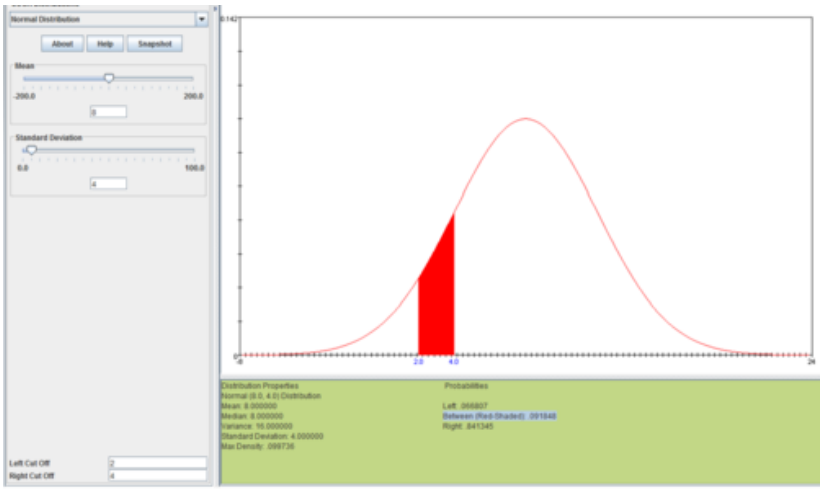
Note that if $\{X_1, X_2, X_3, \ldots\}$

is a sequence of independent Exponential(b) random variables then $Y_k = \sum_{i=1}^{k} X_i$ is a random variable with gamma distribution with the following shape parameter, **k** (positive integer indicating the number of exponential variable in the sum) and scale parameter **b** (which is the exponential parameter). By the central limit theorem, if k is large, then gamma distribution can be approximated by the normal distribution with mean $\mu = kb$ and variance $\sigma^2 = kb^2$. That is, the distribution of the variable $Z_k = \frac{Y_k - kb}{\sqrt{k}\, b}$

tends to the standard normal distribution as $k \longrightarrow \infty$.

For the example above, $\Gamma(k=4, \theta=2)$

, the SOCR Normal Distribution Calculator can be used to obtain an estimate of the area of interest as shown on the image below.



The probabilities of the real Gamma and approximate Normal distributions (on the range [2:4]) are not identical but are sufficiently close.

Short Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Derive the mean of Normal Distribution (2006, 2009 2014)
2. Derive the mode of Normal Distribution (2004, 2014)
3. Derive the division from the mean for Normal Distribution (2001, Dec 2009)
4. Derive the Chief Characteristics of the Normal Distribution (2004)
5. The importance and applications of the Normal Distribution (2009, 2005)
6. If X is normal variant with mean 30 and standard deviation '5' find P (26≤ X≤40), P (X≥45) (2000, 2009)
7. The mean height of students in a college is 155 cms and standard deviation is 15. What is the probability that the mean height of 36 students is less than 157 cms (2011, 2017)
8. If 'X" is normally distributed with mean '2' and variance 0.1, then find P ($|X-2| \geq 0.01$) ? (2010)
9. If X has the binomial distribution with mean 25 and probability of success 1/5, find P ($X<\mu-2\sigma$), where μ and σ^2 are the mean and variance of the distribution? (2009)
10. Find the probability of getting an even number on face 3 to 5 ties in throwing 10 dice together. (2012).

Long Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1.Derive the constants of Normal distribution (i.e., to prove mean= median= mode)(dec09,dec2014)
2. In a Normal distribution ,7% of the items are under 35 and 89% are under 63 . determine the mean and variance of the distribution(jan07,dec 2011,nov 2012)
3. The marks obtained in mathematics by 1000 students is normally distributed with mean 78% and standard deviation 11% determine(2006,2007,ov10,apr2012,nov2011,nov2012)
i) How many students got marks above 90%
ii) what was the highest mark obtained by the lowest 10% of the students
iii)within what limits did the middle of 90% of the students lie
4.If the masses of 300 students are normally distributed with mean 68 kgs and S.D 3kgs ,how many students have masses(2005,08,nov2010,dec2011,may2013)
i) Grater than 72 kgs
ii) less than or equal to 64 kgs

iii)Between 65& 71 kgs inclusive
5.If X is a normal vitiate , find the area A(nov2004,dec2005,may2007)
I) to the left of Z=-1.78
ii) to the right of Z=-1.45
iii) corresponding to $-0.8 \leq Z \leq 1.53$
iv) to the left of Z= -2.52 and to the right of Z=1.83
6.In a sample of 1000 cases, the mean of a certain test is 14 and S.D is2.5 Assuming the distribution to be normal ,find(nov2004,dec2007,nov2015)
i)How many students score between 12&15
ii) How many score above 18
iii) How many score below 18
7. Find the probability of getting an even number on face 3to 5 times in throwing 10 die together(n0v2012)
8.For an exponential distribution $\lambda=1.2$ find i) $p(x \geq 0.5)$ ii)$p(1 \leq x \leq 2)$.also find bits mean & variance (dec2014)
9.Derive the constants of the Gamma distribution (ie mean ,mode, variance)(dec2015)
10. Derive the coefficient of variance , skewers and kurtosis of gamma distribution (nov2010,may2011)

Fill in the Blanks / Choose the Best: (Minimum 10 to 15 with Answers)

1.If a random variable X takes any of the uncountable possible values at any point of time then it is called as ..continuous random variable.....
2.Mean deviation about mean of a uniform distribution is given by ....n=b-a/4.....
3.....Gamma..... distribution is used for calculating the amount of rainfall which is stored in a reservoir .
4. .....Beta.....distribution is used to calculate the number of defective items in a shipment.
5. When the value of location parameter (μ) is 0 and scale parameter (σ)is 1 then it is said to have ....Standard exponential distribution.....
6.If X and Y re two random variable then E(X+Y)=....E(X)+E(Y).....
7.The value of for which f(x)= reaches its maximum value is called...Mode of normal distribution.....
8. .....Covariance..... is measure that measures how much two random variables very together.
9.....Normal distribution..... is an example of continuous probability distribution.
10. Median of uniform distribution is given by.....Median(M)=b+a/2.....

Unit-IV: (Title))

APPLIED STATISTICS

Important points / Definitions: (Minimum 15 to 20 points covering complete topics in that unit)

Introduction to Curve Fitting

Introduction

Historians attribute the phrase *regression analysis* to Sir Francis Galton (1822-1911), a British anthropologist and
meteorologist, who used the term *regression* in an address that was published in *Nature* in 1885. Galton used the

term while talking of his discovery that offspring of seeds "did not tend to resemble their parent seeds in size, but
to be always more mediocre [i.e., more average] than they.... The experiments showed further that the mean filial
regression towards mediocrity was directly proportional to the parental deviation from it."
The content of Galton's paper would probably be called *correlation analysis* today, a term which he also coined.
However, the term *regression* soon was applied to situations other than Galton's and it has been used ever since.
Regression Analysis refers to the study of the relationship between a response (dependent) variable, Y, and one or
more independent variables, the X's. When this relationship is reasonably approximated by a straight line, it is
said to be *linear*, and we talk of linear regression. When the relationship follows a curve, we call it curvilinear
regression.
Usually, you assume that the independent variables are measured exactly (without random error) while the
dependent variable is measured with random error. Frequently, this assumption is not completely true, but when it
cannot be justified, a much more complicated fitting procedure is required. However, if the size of the measurement error in an independent variable is small relative to the range of values of that variable, least squares
regression analysis may be used with legitimacy.

Linear Regression Models

Perhaps the simplest example of a regression model is the familiar straight-line regression between two variables,
X and Y, expressed by the formula:

(1) $Y = B_0 + B_1X$

where B_0 and B_1 are called parameters, which are known constants linking Y and X. B_0 is the y-intercept, B_1 is the
slope.
The relationship in (1) is exact. If you know X, you can determine Y exactly. Exact relationships are hard to find
in applied science. Usually, you have to deal with empirical approximations determined from observed data.
These relationships are represented as follows:
(2) $Y_i = B_0 + B_1 X_i + e_i$
where Y_i and X_i are the i_{th} observed values of the dependent variable and the explanatory (regressor, predictor, or
independent) variable, respectively. B_0 and B_1 are unknown parameter constants which must be estimated. The
error term, e_i, represents the error at the i_{th} data point. It is customary to assume that $E(e_i)=0$ (unbiased) and
$V(e_i)=s^2$ (constant variance).
Actually, linear models include a broader range of models than those represented by equation (2). The main
requirement is that the model is linear in the parameters (the B-coefficients). Other linear models are:
(3) $\ln(Y_i) = B_0 + B_1 \ln(X_i) + e_i$
and
(4) B_i

1
X
$Y = e_{0} + B e_{i} + e_i$
At first, (4) appears nonlinear in the parameters. However, if you set 0
B
$C = e_{0}$, $C_1 = B_1$, and i
$Z = e^{X_i}$ you will
notice that it reduces to the form of (2). Models which may be reduced to linear models with suitable transformations are called intrinsically linear models. Model (5) is a second example of an intrinsically linear
model.
(5) i 0
B X
$Y = B [e_{1 i}] e_i$
Notice that applying a logarithmic transformation to both sides of (5) results in the following:
(6) $\ln(Y_i) = \ln(B_0) + B_1 X_i + \ln(e_i)$
This is now easily recognized as an intrinsically linear model.
You should note that if the errors are normally distributed in (5), their logarithms in model (6) will not be so
distributed. Likewise, if the errors, log(e_i), in (6) are normally distributed, the detransformed errors, e_i, in (5) will
not be. Hence, when you are applying transformations to simplify models, you should check to see that the
resulting error term has the desired properties. We will come back to this point later.

Nonlinear Regression Models

Nonlinear regression models are those which are not linear in the parameters to begin with nor can they be made
so by transformation. A general representation for the nonlinear regression model is:
(7) $Y_i = f(X_i, e_i; B_1, B_2, ..., B_p)$
where B_1, B_2, ..., B_p are the *p* parameters to be estimated from your data, and e_i is the error term.
Note that e_i is not necessarily additive as in (2), although this is a common form. An example of an additive model
is:
(8) i 0
B (X)
$Y = B e_{1 i} + e_i$
Linear models, such as those in (5), are preferred over nonlinear models, such as (8), for two main reasons. First,
the linear model is mathematically easier to work with. Parameters may be estimated with explicit expressions.
Nonlinear models must use iterative schemes, which may converge to several solutions. Second, often the
investigator does not know the actual form of the relationship and is looking for an approximation. The linear
model is an obvious place to start.

Least Squares Estimation of Nonlinear Models

The method of least squares minimizes the error sum of squares, Q, which is given by
(9) $Q = (Y Y)$
i=1
n
i i

$\sum - \cdot {}_2$

where $Y\cdot_i = f(X_i; B\cdot_1, B\cdot_2, \ldots)$ is the value predicted for a specific X_i using the parameters estimated by least squares. If the errors are normally distributed, the least squares estimates are also the maximum likelihood estimates. This is one of the reasons we strive for normally distributed errors.

The values of the B's that minimize *Q* in (9) may be found either of two ways. First, if *f()* is a simple function, such as in (2), you may find an analytic solution by differentiating *Q* with respect to B_1, B_2, ..., B_p, setting the resulting partial derivatives equal to zero, and solving the resulting p normal equations. Unfortunately, very few nonlinear models may be estimated this way.

The second method is to try different values for the parameters, calculating *Q* each time, and work towards the smallest *Q* possible. Three general procedures work toward a solution in this manner.

The *Gauss-Newton*, or *linearization*, method uses a Taylor series expansion to approximate the nonlinear model with linear terms. These may be used in a linear regression to come up with trial parameter estimates which may then be used to form new linear terms. The process iterates until a solution is reached.

The *steepest descent* method searches for the minimum *Q* value by iteratively determining the direction in which the parameter estimates should be changed. It is particularly useful when poor starting values are used.

The *Marquardt* algorithm uses the best features of both the Gauss-Newton and the steepest descent methods. This is the procedure that is implemented in this program. Note that *numerical derivatives* are used whenever derivatives are called for.

Starting Values

All iterative procedures require starting values for the parameters. This program finds the starting values for you. However, the values so found may fail to converge or you may be using a user-defined function which does not have preprogrammed starting values. Hence, you will have to supply your own starting values.

Unfortunately, there is no easy method for generating starting values for the B's in every case. However, we can provide you with some guidelines and a general method of attack that will work in many cases.

1. Try entering a 1 or 0 for each parameter and letting the program crank through a few iterations for you. You must be careful not to give impossible values (like taking the square root of a negative number), or the procedure will halt immediately. Even though the procedure may take longer to converge, the elapsed time will often be shorter than when using steps 2 and 3 below, since they require much more time and effort on your part.
2. Pick *p* observations that spread across the range of the independent variable and solve the model ignoring the error term. The resulting solution will often provide reasonable starting values. This includes transforming the model to a simpler form.

3. Consider the behavior of *f()* as X approaches zero or infinity and substitute in appropriate observations that
most closely approximate these conditions. This might be accomplished from a plot of your data or from an
examination of the data directly. Once some of the parameters have been estimated in this manner, others may be
found by applying step 2 above.

Inferences about Nonlinear Regression Parameters

The following results are from Seber (1989), chapter 5. They require the assumption that the errors are normally
distributed with equal variance.

Confidence Intervals for Parameters

Let

(10) $Y_i = f(X_i; B_1, B_2, \ldots) + e_i \ (i=1,2,\ldots,n)$

represent the nonlinear model that we are interested in fitting. Let B represent the parameters B_1, B_2, ..., B_p. The

asymptotic distribution of the estimates of B, which we call $B \cdot$, is given by

(11) $B \cdot \ N(B, C)$, $C = F.F.$, $F. = [(f$

Bj

$_p)]$

$\sim \sigma_{2-1}'$

∂

∂

For large n we have, approximately,

(12) $_{r\,n-p}$

$_{/2\,rr} B \cdot \pm t_\alpha s\, c \cdot$

which gives approximate, large-sample 100(1-a)% confidence limits for the individual parameters. Note the *s* is

an estimate of σ in (11), based on the residuals from the fit of (10).

These intervals are often referred to as the asymptotic-linearization confidence intervals because they are based on

a local linearization of the function (10). If the curvature of (10) is sharp near $B \cdot_i$, then the approximation will
have considerable error and (12) will be unreliable.

Confidence Intervals for a Predicted Value

Using (10) - (12) it is easy to give approximate, asymptotic 100(1-a)% confidence intervals (or prediction
intervals) for predicted values. These are:

(13) $_{0\,n-p}$

$_{/2}$

$_0$

$_{-1}$

$_0$

$_{1/2}$

$_0$

$_0$

$_1$

$_0$

$_2$

$Y\ t\ s[1 + f(F.F.)\, f]$, $f = (f(X)$

B
, *f(X)*
B
· ± ′ „...)
∂
∂
∂
′ ∂
α

Note that f_0 and *F.* must be estimated using the *B* · . Hence, if the fit of (10) is good and there is little curvature,
these confidence intervals will be accurate. If the fit is poor or there is sharp curvature near the region of interest,
these confidence limits may be unsatisfactory.

Parameterization

One of the first choices you must make is the way parameters are attached to the functional form of a model. For
example, consider the following two models:

(14) *i*
0 i
i 1
Y = *i B X*
X + *B*
+*e*

(15) *i*
i
0 i 1
Y = *i X*
C X +*C*
+*e*

These are actually the same basic model. Note that if we let $C_0=1/B_0$ and $C_1=B_1/B_0$, model (15) is simply a
rearrangement of (14). However, the statistical properties of these two models are very different. Equations (14)
and (15) are two parameterizations of the same basic model.

If there is no precedent for a particular model parameterization, then you should use that model with the best
statistical properties. If this case, trial-and-error methods will have to be used to find a model. Often this will
include comparing a plot of your data to a plot of the functional forms that are available, until a good match is
found. If there are several models possible, a careful study of the error terms (residuals) is necessary to help in
your selection.

A common misconception is the view that whether a parameter appears linearly or nonlinearly in the nonlinear
model relates directly to its estimation behavior. This is just not the case. (See Ratkowsky (1989) section 2.5.2.)

Another common misconception is that a complicated model is superior to a simple model. In general, the simpler

the model, the better the behavior of the estimation process. Adding an extra parameter has unpredictable results
on the estimation process. In some cases, it has little effect, while in others it has disastrous consequences.
Overparameterization (using too complicated a model) often leads to convergence problems. These models may
have multiple solutions. The estimates from these models are usually biased and nonnormally distributed. They

show high correlation among the parameter estimates. This problem may also occur when you use only a portion
of a complicated function to fit a set of data. It is always better to find a simpler function that exhibits the
functional behavior of your data. (See Ratkowsky (1989) section 2.5.4.)

The Stochastic Term e_i

A regression model such as (10) may be thought of as having a deterministic part *f(X;B1,B2,...)* and a stochastic
(random) part e_i. Often, assumptions about the e_i are necessary. The most common are:
1. Independently distributed
2. Identically distributed with constant variance
3. Normally distributed

Independence

Independence means that the error at one value of i (say i=4) is not related to the error at another value of i (say
i=5). Independence is often violated when data are taken over time and some carry-over effects are active.

Identicalness

Identicalness means that the distribution of the errors is the same for all values of i (for all data pairs X_i and Y_i).
In practice, this assumption is equated with constant variance in the errors. If the variance of the e_i increases or
decreases, then this assumption is violated.

Normality

The question of normality is very difficult to assess with small sample sizes (under 100). With large sample sizes,
normal probability plots (discussed later) do a pretty good job. Least-squares methods (those used by this
program) tend to create normality in the observed residuals even if the actual e_i's are not normal.
Some normality tests are available in the *Descriptive Statistics* module, so you can try them on your residuals.
However, most technicians agree that if your observed residuals have a bell-shaped distribution with no outliers,
the normality assumption is okay.

Summary

These assumptions are ideals that are only approximately met in practice. Least squares tends to be robust to
minor departures from these assumptions. Only when there are major departures such as outliers, a large shift in
the size of the variance, or a large serial correlation between successive residuals will estimates be significantly in
error.

Curve Fitting

Curve fitting is the process of introducing mathematical relationships between dependent and independent variables in the form of an equation for a given set of data.

Method of Least Squares

The method of least squares helps us to find the values of unknowns a

and b

in such a way that the following two conditions are satisfied:

- The sum of the residual (deviations) of observed values of Y

and corresponding expected (estimated) values of Y will be zero. $\sum(Y-\hat{Y})=0$

- .
- The sum of the squares of the residual (deviations) of observed values of Y and corresponding expected values ($\hat{Y}$) should be at least $\sum(Y-\hat{Y})^2$

Fitting of a Straight Line

A straight line can be fitted to the given data by the method of least squares. The equation of a straight line or least square line is $Y=a+bX$

, where a and b

are constants or unknowns.

To compute the values of these constants we need as many equations as the number of constants in the equation. These equations are called normal equations. In a straight line there are two constants a

and b

so we require two normal equations.

Normal Equation for '*a*' $\sum Y=na+b\sum X$

Normal Equation for '*b*' $\sum XY=a\sum X+b\sum X^2$

The direct formula of finding a and b is written as

$$b=\frac{\sum XY-(\sum X)(\sum Y)n}{\sum X2-(\sum X)2n},\ a=\overline{Y}-bX$$

Correlation and regression

The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

Correlation coefficient

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Figure 11.1 gives some graphical representations of correlation.

Figure 11.1 Correlation illustrated.

Looking at data: scatter diagrams

When an investigator has collected two series of observations and wishes to see whether there is a relationship between them, he or she should first construct a scatter diagram. The vertical scale represents one set of measurements and the horizontal scale the other. If one set of observations consists of experimental results and the other consists of a time scale or observed classification of some kind, it is usual to put the experimental results on the vertical axis. These represent what is called the "dependent variable". The "independent variable", such as time or height or some other observed classification, is measured along the horizontal axis, or baseline.

The words "independent" and "dependent" could puzzle the beginner because it is sometimes not clear what is dependent on what. This confusion is a triumph of common sense over

misleading terminology, because often each variable is dependent on some third variable, which may or may not be mentioned. It is reasonable, for instance, to think of the height of children as dependent on age rather than the converse but consider a positive correlation between mean tar yield and nicotine yield of certain brands of cigarette.' The nicotine liberated is unlikely to have its origin in the tar: both vary in parallel with some other factor or factors in the composition of the cigarettes. The yield of the one does not seem to be "dependent" on the other in the sense that, on average, the height of a child depends on his age. In such cases it often does not matter which scale is put on which axis of the scatter diagram. However, if the intention is to make inferences about one variable from the other, the observations from which the inferences are to be made are usually put on the baseline. As a further example, a plot of monthly deaths from heart disease against monthly sales of ice cream would show a negative association. However, it is hardly likely that eating ice cream protects from heart disease! It is simply that the mortality rate from heart disease is inversely related - and ice cream consumption positively related - to a third factor, namely environmental temperature.

Calculation of the correlation coefficient

A paediatric registrar has measured the pulmonary anatomical dead space (in ml) and height (in cm) of 15 children. The data are given in table 11.1 and the scatter diagram shown in figure 11.2 Each dot represents one child, and it is placed at the point corresponding to the measurement of the height (horizontal axis) and the dead space (vertical axis). The registrar now inspects the pattern to see whether it seems likely that the area covered by the dots centres on a straight line or whether a curved line is needed. In this case the paediatrician decides that a straight line can adequately describe the general trend of the dots. His next step will therefore be to calculate the correlation coefficient.

Table 11.1 Correlation between height and pulmonary anatomical dead space in 15 children

Child number	Height (cm)	Dead space (ml), y
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Total	2169	1004
Mean	144.6	66.933

When making the scatter diagram (figure 11.2) to show the heights and pulmonary anatomical dead spaces in the 15 children, the paediatrician set out figures as in columns (1), (2), and (3) of table 11.1 . It is helpful to arrange the observations in serial order of the independent variable when one of the two variables is clearly identifiable as independent. The corresponding figures for the dependent variable can then be examined in relation to the increasing series for the independent variable. In this way we get the same picture, but in numerical form, as appears in the scatter diagram.

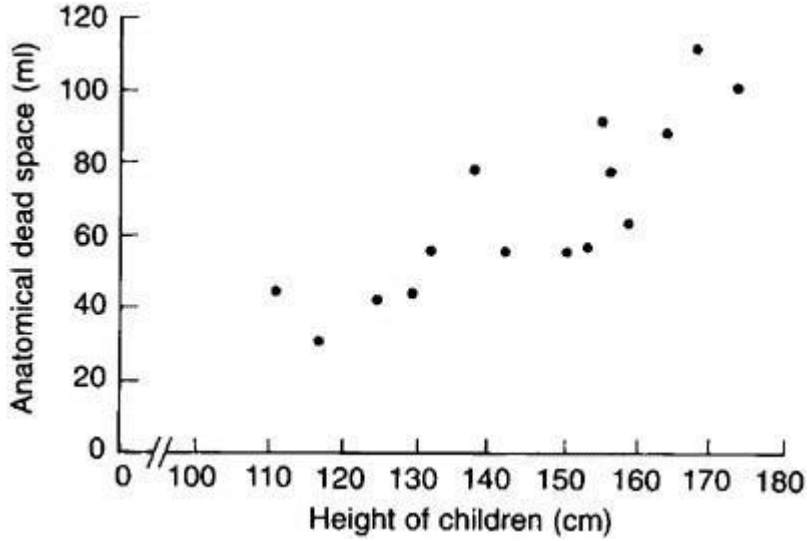


Figure 11.2 Scatter diagram of relation in 15 children between height and pulmonary anatomical dead space.

The calculation of the correlation coefficient is as follows, with x representing the values of the independent variable (in this case height) and y representing the values of the dependent variable (in this case anatomical dead space). The formula to be used is:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2(y-\bar{y})^2]}}$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2(y-\bar{y})^2]}}$$

which can be shown to be equal to:

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD\,(x)SD(y)}$$

Significance test

To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation:

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

The *t* is entered at n - 2 degrees of freedom.

For example, the correlation coefficient for these data was 0.846.

The number of pairs of observations was 15. Applying equation 11.1, we have:

$$t = 0{\cdot}846\sqrt{\frac{15-2}{1-0{\cdot}846^2}} = 5{\cdot}72.$$

Entering table B at 15 - 2 = 13 degrees of freedom we find that at t = 5.72, P<0.001 so the correlation coefficient may be regarded as highly significant. Thus (as could be seen immediately from the scatter plot) we have a very strong correlation between dead space and height which is most unlikely to have arisen by chance.

The assumptions governing this test are:

1. That both variables are plausibly Normally distributed.
2. That there is a linear relationship between them.
3. The null hypothesis is that there is no association between them.

The test should not be used for comparing two methods of measuring the same quantity, such as two methods of measuring peak expiratory flow rate. Its use in this way appears to be a common mistake, with a significant result being interpreted as meaning that one method is equivalent to the other. The reasons have been extensively discussed(2) but it is worth recalling that a significant result tells us little about the strength of a relationship. From the formula it should be clear that with even with a very weak relationship (say r = 0.1) we would get a significant result with a large enough sample (say n over 1000).

Spearman rank correlation

A plot of the data may reveal outlying points well away from the main body of the data, which could unduly influence the calculation of the correlation coefficient. Alternatively the variables may be quantitative discrete such as a mole count, or ordered categorical such as a pain score. A non-parametric procedure, due to Spearman, is to replace the observations by their ranks in the calculation of the correlation coefficient.

This results in a simple formula for Spearman's rank correlation, Rho.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where d is the difference in the ranks of the two variables for a given individual.

The regression equation

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. For instance, in the children described earlier greater height is associated, on average, with greater anatomical dead Space. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the

correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. The first of these is its distance above the baseline; the second is its

slope. They are expressed in the following *regression equation* :

With this equation we can find a series of values of y_{fit} the variable, that correspond to each of a series of values of x, the independent variable. The parameters α and β have to be estimated from the data. The parameter signifies the distance above the baseline at which the regression line cuts the vertical (y) axis; that is, when y = 0. The parameter β (the *regression coefficient*) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards.

The regression equation is often more useful than the correlation coefficient. It enables us to predict y from x and gives us a better summary of the relationship between the two variables. If, for a particular value of x, x i, the regression equation predicts a value of y fit , the prediction error is $y_1 - y_{fit}$. It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error $\Sigma(y_1 - y_{fit})$ of zero because the positive and negative terms exactly cancel. To remove the negative signs we square the differences and the regression equation chosen to minimise the sum of squares of the prediction errors, $S^2 = \Sigma(y_1 - y_{fit})^2$ We denote the sample estimates of Alpha and Beta by a and b. It can be shown that the one straight line that minimises S_2, the least squares estimate, is given by

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

it can be shown that

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)SD(x)^2}$$

which is of use because we have calculated all the components of equation (11.2) in the calculation of the correlation coefficient.

The calculation of the correlation coefficient on the data in table 11.2 gave the following:

$$\Sigma xy = 150605, SD(x) = 19.3679, \bar{y} = 66.93, \bar{x} = 144.6$$

Applying these figures to the formulae for the regression coefficients, we have:

$$b = \frac{150605 - 15 \times 66.93 \times 144.6}{14 \times 19.3679^2} = \frac{5426.6}{5251.6} = 1.033 \text{ml/cm}$$

$$a = 66.39 - (1.033 \times 144.6) = -82.4$$

Therefore, in this case, the equation for the regression of y on x becomes

$$y = -82.4 + 1.033x$$

This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.033 ml *over the range of measurements made*.

The line representing the equation is shown superimposed on the scatter diagram of the data in figure 11.2. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:

If x = 110, y = (1.033 x 110) - 82.4 = 31.2

If x = 140, y = (1.033 x 140) - 82.4 = 62.2

If x = 170, y = (1.033 x 170) - 82.4 = 93.2

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them.

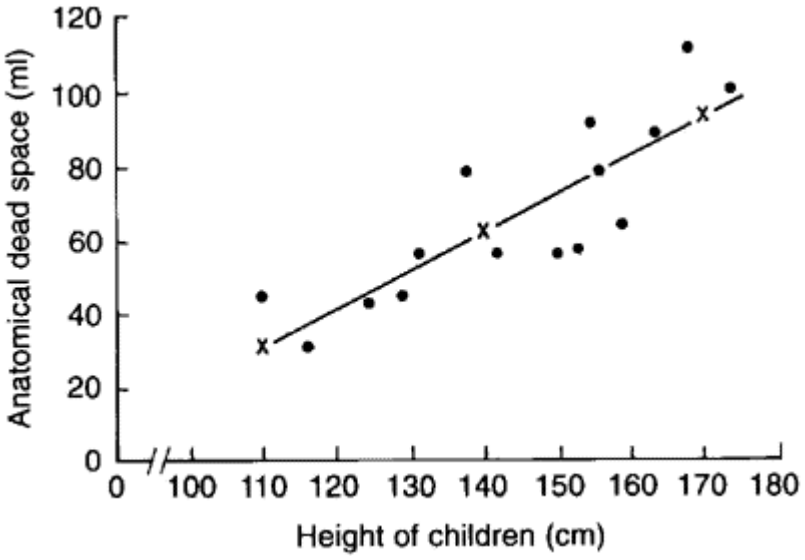


Figure 11.3 Regression line drawn on scatter diagram relating height and pulmonaiy anatomical dead space in 15 children

The standard error of the slope SE(b) is given by:

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x - \bar{x})^2}}$$

where S_{res} is the residual standard deviation, given by:

$$S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n - 2}}$$

This can be shown to be algebraically equal to

$$\sqrt{((SD((y)^2(1 - r^2)(n - 1) / (n - 2)}$$

We already have to hand all of the terms in this expression. Thus S_{res} is the square root of $23.6476^2(1 + -0.846^2)14 / 13 = \sqrt{171.2029} = 13.08445$. The denominator of (11.3) is 72.4680. Thus SE(b) = 13.08445/72.4680 = 0.18055.

We can test whether the slope is significantly different from zero by:

t = b/SE(b) = 1.033/0.18055 = 5.72.

Again, this has n - 2 = 15 - 2 = 13 degrees of freedom. The assumptions governing this test are:

1. That the prediction errors are approximately Normally distributed. Note this does not mean that the x or y variables have to be Normally distributed.
2. That the relationship between the two variables is linear.
3. That the scatter of points about the line is approximately constant - we would not wish the variability of the dependent variable to be growing as the independent variable increases. If this is the case try taking logarithms of both the x and y variables.

Note that the test of significance for the slope gives exactly the same value of P as the test of significance for the correlation coefficient. Although the two tests are derived differently, they are algebraically equivalent, which makes intuitive sense.

We can obtain a 95% confidence interval for b from

$$b - t_{0.05} \times SE(b) \text{ to } b + t_{0.05} \times SE(b)$$

where the tstatistic from has 13 degrees of freedom, and is equal to 2.160.

Thus the 95% confidence interval is

l.033 - 2.160 x 0.18055 to l.033 + 2.160 x 0.18055 = 0.643 to 1.422.

Short Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Derive the Normal Equations to fit the straight line y = a+bx. (2006, 2017)
2. By the method of least squares, find the straight line that best fits the following data (2010, 2011)

X	1	2	3	4	5
Y	14	27	40	55	68

3. Fit a polynomial of second degree to the data points given in the following table (2011, 2012)

X	0	1.0	2.0
Y	1.0	6.0	17.0

4. Write the normal equations to fit the parabola (2017)
5. Writ the relation between correlation and regression coefficients. Is it possible to have two variable x and y with regression coefficient as 2.8 an -0.5 ? explain (2006)
6. Write the properties of correlation coefficient (2010, 2012)
7. Fit a exponential curve of the form $y=ab^x$ for the data. (2015)

X	1	2	3	4
Y	7	11	17	27

8. Find a curve $y=ae^{bx}$ to the data. (2008, 2009)

X	0	2	4
Y	5.1	10	31.1

9. If 'Θ' is the angle between two regression lines and S.D. of y is twice the S.D. of x and r = 0.25, find tanΘ . (2010, 2012)
10. Given n=10, σx = 5.4, σy = 6.2 and sum of product of deviations from the mean of x & y is 66. Find the correlation coefficient. (2011)

Long Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1.Fit a straight line to the form Y=a+bx for following data(may 09,2010,dec2011)

x	0	5	10	15	20	25
y	12	15	17	22	24	30

2.Fit a second degree polynomial to the following data by the method of least squares(nov2008,nov2009)

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3y

3.Useing the method of least squares ,find the constants a & b such that $y= ae^{bx}$ fits the following data(jan2010,sep 2017)

x	0.0	0.5	1.0	1.5	2.0	2.5
y	0.10	0.45	2.15	9.15	40.35	180.75

4.Obtain a relation of the form $y= ab^x$ for the following data by the method of least squares(june2010,june2013)

x	2	3	4	5	6
y	8.3	15.4	33.1	65.2	127.4

5.a) Derive the normal equations to fit the straight line y= a+bx(may2006,june2010,june2011,may2012)
b) Derive the normal equations to fit the parabola $y= a+bx+cx^2$

6.what re the five methods available to study correlation?(may 2009)
7. prove that the correlation coefficient and discuss its properties.(nov2010)
8.prove that the correlation coefficient is independent of change of scale.may 2008,dec2009)
9.Find the coefficient of correlation between independent production and expor using the following data and comment on the result(may2005,dec2009)

Production (in corer tons)	55	56	58	59	60	60	62
Exports (in corer tons)	35	38	38	39	44	43	45

10. the correlation coefficient between two variable X and Y is r =0.6 if σx=1.50,σy=2.00,x=10and y= 20 find the regression lines of i) X on Y ii) Y on X(may 2008,nov20013)
11.Define rank correlation and derive rank correlation coefficient.(nov2005,09)

Fill in the Blanks / Choose the Best: (Minimum 10 to 15 with Answers)
1.Negative correlation is also known as ..Inverse Correlation.....
2. ....Karl Pearson's.....measure is known as Poisson coefficient of correlation.
3.Regression line of y on x is y=.a+ bx.....
4. The correlation coefficient lies between ...-1.....and...1.....
5. correlation is classified into ....Eight.....types.
6.....Correlation..... refers to the statistical tool for measuring the degree of relationship that exists between two or more variables.
7.The amount of correlation in a sample is measured by.....Sample correlation.....
8. Two independent variables are .... Uncorrelated.....
9. Regression analysis involves ...Dependent.....and .....Independent.....types of variables.
10.The formula used for concurrent deviation method is $\gamma=\frac{\sqrt{\pm(2c-n)}}{n}$

Unit-V: (Title) TEST OF HYPOTHESIS

Important points / Definitions: (Minimum 15 to 20 points covering complete topics in that unit)

Chi-Square Distribution

The distribution of the chi-square statistic is called the chi-square distribution. In this lesson, we learn to compute the chi-square statistic and find the probability associated with the statistic. And we'll work through some chi-square examples to illustrate key points.

The Chi-Square Statistic

Suppose we conduct the following statistical experiment. We select a random sample of size *n* from a normal population, having a standard deviation equal to σ. We find that the standard deviation in our sample is equal to *s*. Given these data, we can define a statistic, called **chi-square**, using the following equation:

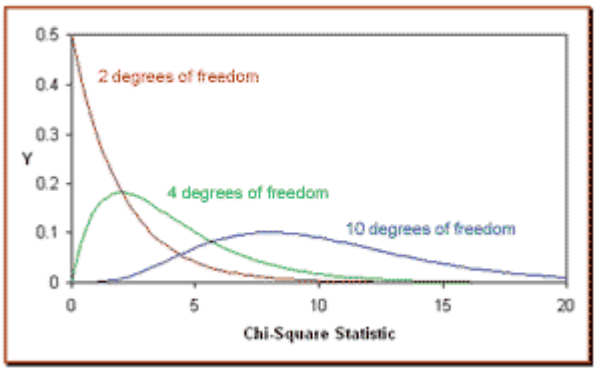
$$X^2 = [(n - 1) * s^2] / \sigma^2$$

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following probability density function:

$$Y = Y_0 * (X^2)^{(v/2 - 1)} * e^{-X2 / 2}$$

where Y_0 is a constant that depends on the number of degrees of freedom, X^2 is the chi-square statistic, $v = n - 1$ is the number of degrees of freedom, and *e* is a constant equal to the base of the natural logarithm system (approximately 2.71828). Y_0 is defined, so that the area under the chi-square curve is equal to one.

In the figure below, the red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is $n - 1 = 3 - 1 = 2$. Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4); and the blue curve, for samples of size 11 (degrees of freedom equal to 10).



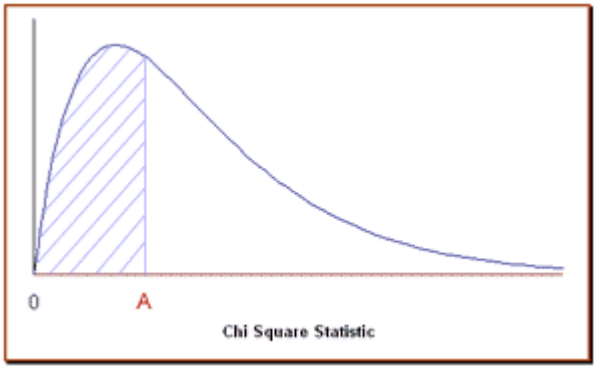
The chi-square distribution has the following properties:

- The mean of the distribution is equal to the number of degrees of freedom: $\mu = v$.
- The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2 * v$

- When the degrees of freedom are greater than or equal to 2, the maximum value for Y occurs when $X^2 = \nu - 2$.
- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

Cumulative Probability and the Chi-Square Distribution

The chi-square distribution is constructed so that the total area under the curve is equal to 1. The area under the curve between 0 and a particular chi-square value is a cumulative probability associated with that chi-square value. For example, in the figure below, the shaded area represents a cumulative probability associated with a chi-square statistic equal to *A*; that is, it is the probability that the value of a chi-square statistic will fall between 0 and *A*.



Fortunately, we don't have to compute the area under the curve to find the probability. The easiest way to find the cumulative probability associated with a particular chi-square statistic is to use the Chi-Square Calculator, a free tool provided by Stat Trek.

Chi-Square Calculator

The Chi-Square Calculator solves common statistics problems, based on the chi-square distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

Chi-Square Calculator

Test Your Understanding

Problem 1

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Solution

We know the following:

- The standard deviation of the population is 4 minutes.
- The standard deviation of the sample is 6 minutes.
- The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

$$X^2 = [(n - 1) * s^2] / \sigma^2$$
$$X^2 = [(7 - 1) * 6^2] / 4^2 = 13.5$$

where X^2 is the chi-square statistic, *n* is the sample size, *s* is the standard deviation of the sample, and σ is the standard deviation of the population.

Problem 2

Let's revisit the problem presented above. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

Solution

We know the following:

- The sample size *n* is equal to 7.
- The degrees of freedom are equal to *n* - 1 = 7 - 1 = 6.
- The chi-square statistic is equal to 13.5 (see Example 1 above).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic falls between 0 and 13.5, we enter the degrees of freedom (6) and the chi-square statistic (13.5) into the Chi-Square Distribution Calculator. The calculator displays the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means (by the subtraction rule) that the probability that the standard deviation would be *greater than* 6 minutes is 1 - 0.96 or 0.04.

Student's t Distribution

The **t distribution** (aka, **Student's t-distribution**) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

Why Use the t Distribution?

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic** (also known as the **t score**), whose values are given by:

$$t = [x - \mu] / [s / \sqrt{n}]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the *t* statistic is called the **t distribution** or the **Student t distribution**.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

Degrees of Freedom

There are actually many different t distributions. The particular form of the t distribution is determined by its **degrees of freedom**. The degrees of freedom refers to the number of independent observations in a set of data.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the *t* statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.

Properties of the t Distribution

The t distribution has the following properties:

- The mean of the distribution is equal to 0 .
- The variance is equal to $v / (v - 2)$, where v is the degrees of freedom (see last section) and $v \geq 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

When to Use the t Distribution

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The sampling distribution of a statistic should be bell-shaped if any of the following conditions apply.

- The population distribution is normal.
- The population distribution is symmetric, unimodal, without outliers, and the sample size is at least 30.
- The population distribution is moderately skewed, unimodal, without outliers, and the sample size is at least 40.
- The sample size is greater than 40, without outliers.

The t distribution should *not* be used with small samples from populations that are not approximately normal.

Probability and the Student t Distribution

When a sample of size n is drawn from a population having a normal (or nearly normal) distribution, the sample mean can be transformed into a t statistic, using the equation presented at the beginning of this lesson. We repeat that equation below:

$$t = [x - \mu] / [s / sqrt(n)]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, n is the sample size, and degrees of freedom are equal to n - 1.

The t statistic produced by this transformation can be associated with a unique cumulative probability. This cumulative probability represents the likelihood of finding a sample mean less than or equal to x, given a random sample of size n.

The easiest way to find the probability associated with a particular t statistic is to use the T Distribution Calculator, a free tool provided by Stat Trek.

T Distribution Calculator

The T Distribution Calculator solves common statistics problems, based on the t distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked

questions and sample problems provide straightforward explanations. The calculator is free. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

T Distribution Calculator

Notation and t Statistics

Statisticians use t_α to represent the t statistic that has a cumulative probability of $(1 - \alpha)$. For example, suppose we were interested in the t statistic having a cumulative probability of 0.95. In this example, α would be equal to (1 - 0.95) or 0.05. We would refer to the t statistic as $t_{0.05}$

Of course, the value of $t_{0.05}$ depends on the number of degrees of freedom. For example, with 2 degrees of freedom, $t_{0.05}$ is equal to 2.92; but with 20 degrees of freedom, $t_{0.05}$ is equal to 1.725.

Note: Because the t distribution is symmetric about a mean of zero, the following is true.

$$t_\alpha = -t_{1 - \text{alpha}} \qquad \text{And} \qquad t_{1 - \text{alpha}} = -t_\alpha$$

Thus, if $t_{0.05} = 2.92$, then $t_{0.95} = -2.92$.

Test Your Understanding

Problem 1

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme 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?

Note: There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented below. Solution A is the traditional approach. It requires you to compute the t statistic, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability. Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t statistic "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

Solution A

The first thing we need to do is compute the t statistic, based on the following equation:

$$t = [x - \mu] / [s / \text{sqrt}(n)]$$
$$t = (290 - 300) / [50 / \text{sqrt}(15)]$$
$$t = -10 / 12.909945 = - 0.7745966$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

Now, we are ready to use the T Distribution Calculator. Since we know the t statistic, we select "T score" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to 15 - 1 = 14.
- The t statistic is equal to - 0.7745966.

The calculator displays 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.

Solution B:

This time, we will work directly with the raw data from the problem. We will not compute the t statistic; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- 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 calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

Problem 2

Suppose scores on an IQ test are normally distributed, with a population mean of 100. Suppose 20 people are randomly selected and tested. The standard deviation in the sample group is 15. What is the probability that the average test score in the sample group will be at most 110?

Solution:

To solve this problem, we will work directly with the raw data from the problem. We will not compute the t statistic; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to 20 - 1 = 19.
- The population mean equals 100.
- The sample mean equals 110.
- The standard deviation of the sample is 15.

We enter these values into the T Distribution Calculator. The calculator displays the cumulative probability: 0.996. Hence, there is a 99.6% chance that the sample average will be no greater than 110.

F Distribution

The F distribution is the probability distribution associated with the f statistic. In this lesson, we show how to compute an f statistic and how to find probabilities associated with specific f statistic values.

The f Statistic

The ***f* statistic**, also known as an ***f* value**, is a random variable that has an F distribution. (We discuss the F distribution in the next section.)

Here are the steps required to compute an ***f* statistic**:

- Select a random sample of size n_1 from a normal population, having a standard deviation equal to σ_1.
- Select an independent random sample of size n_2 from a normal population, having a standard deviation equal to σ_2.
- The f statistic is the ratio of s_1^2/σ_1^2 and s_2^2/σ_2^2.

The following equivalent equations are commonly used to compute an f statistic:

$$f = [s_1^2/\sigma_1^2] / [s_2^2/\sigma_2^2]$$

$$f = [s_1^2 * \sigma_2^2] / [s_2^2 * \sigma_1^2]$$

$$f = [X^2_1 / v_1] / [X^2_2 / v_2]$$

$$f = [X^2_1 * v_2] / [X^2_2 * v_1]$$

where σ_1 is the standard deviation of population 1, s_1 is the standard deviation of the sample drawn from population 1, σ_2 is the standard deviation of population 2, s_2 is the standard deviation of the sample drawn from population 2, X^2_1 is the chi-square statistic for the sample drawn from population 1, v_1 is the degrees of freedom for X^2_1, X^2_2 is the chi-square statistic for the sample drawn from population 2, and v_2 is the degrees of freedom for X^2_2 . Note that degrees of freedom $v_1 = n_1 - 1$, and degrees of freedom $v_2 = n_2 - 1$.

The F Distribution

The distribution of all possible values of the f statistic is called an **F distribution**, with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

The curve of the F distribution depends on the degrees of freedom, v_1 and v_2. When describing an F distribution, the number of degrees of freedom associated with the standard deviation in the numerator of the *f* statistic is always stated first. Thus, *f*(5, 9) would refer to an F distribution with $v_1 = 5$ and $v_2 = 9$ degrees of freedom; whereas *f*(9, 5) would refer to an F distribution with $v_1 = 9$ and $v_2 = 5$ degrees of freedom. Note that the curve represented by *f*(5, 9) would differ from the curve represented by *f*(9, 5).

The F distribution has the following properties:

- The mean of the distribution is equal to $v_2 / (v_2 - 2)$ for $v_2 > 2$.
- The variance is equal to $[2 * v_2^2 * (v_1 + v_1 - 2)] / [v_1 * (v_2 - 2)^2 * (v_2 - 4)]$ for $v_2 > 4$.

Cumulative Probability and the F Distribution

Every *f* statistic can be associated with a unique cumulative probability. This cumulative probability represents the likelihood that the *f* statistic is less than or equal to a specified value.

Statisticians use f_α to represent the value of an *f* statistic having a cumulative probability of $(1 - \alpha)$. For example, suppose we were interested in the *f* statistic having a cumulative probability of 0.95. We would refer to that *f* statistic as $f_{0.05}$, since $(1 - 0.95) = 0.05$.

Of course, to find the value of f_α, we would need to know the degrees of freedom, v_1 and v_2. Notationally, the degrees of freedom appear in parentheses as follows: $f_\alpha(v_1,v_2)$. Thus, $f_{0.05}(5, 7)$ refers to value of the f statistic having a cumulative probability of 0.95, $v_1 = 5$ degrees of freedom, and $v_2 = 7$ degrees of freedom.

The easiest way to find the value of a particular *f* statistic is to use the F Distribution Calculator.

F Distribution Calculator

The F Distribution Calculator solves common statistics problems, based on the F distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can found in the Stat Trek main menu under the Stat Tools tab. Or you can tap the button below.

F Distribution Calculator

The use of the F Distribution Calculator is illustrated below in Problem 2.

Test Your Understanding

Problem 1

Suppose you randomly select 7 women from a population of women, and 12 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population standard deviation	Sample standard deviation
Women	30	35
Men	50	45

Compute the f statistic.

Solution A: The f statistic can be computed from the population and sample standard deviations, using the following equation:

$$f = [s_1^2/\sigma_1^2] / [s_2^2/\sigma_2^2]$$

where σ_1 is the standard deviation of population 1, s_1 is the standard deviation of the sample drawn from population 1, σ_2 is the standard deviation of population 2, and s_1 is the standard deviation of the sample drawn from population 2.

As you can see from the equation, there are actually two ways to compute an f statistic from these data. If the women's data appears in the numerator, we can calculate an f statistic as follows:

$$f = (35^2 / 30^2) / (45^2 / 50^2)$$

$$f = (1225 / 900) / (2025 / 2500)$$

$$f = 1.361 / 0.81 = 1.68$$

For this calculation, the numerator degrees of freedom v_1 are 7 - 1 or 6; and the denominator degrees of freedom v_2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate an f statistic as follows:

$$f = (45^2 / 50^2) / (35^2 / 30^2)$$

$$f = (2025 / 2500) / (1225 / 900)$$

$$f = 0.81 / 1.361 = 0.595$$

For this calculation, the numerator degrees of freedom v_1 are 12 - 1 or 11; and the denominator degrees of freedom v_2 are 7 - 1 or 6.

When you are trying to find the cumulative probability associated with an f statistic, you need to know v_1 and v_2. This point is illustrated in the next example.

Problem 2

Find the cumulative probability associated with each of the f statistics from Example 1, above.

Solution: To solve this problem, we need to find the degrees of freedom for each sample. Then, we will use the F Distribution Calculator to find the probabilities.

- The degrees of freedom for the sample of women is equal to n - 1 = 7 - 1 = 6.
- The degrees of freedom for the sample of men is equal to n - 1 = 12 - 1 = 11.

Therefore, when the women's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 6; and the denominator degrees of freedom v_2 is equal to 11. And, based on the computations shown in the previous example, the f statistic is equal to 1.68. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.78.

On the other hand, when the men's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 11; and the denominator degrees of freedom v_2 is equal to 6. And, based on the computations shown in the previous example, the f statistic is equal to 0.595. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.22.

Short Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1. Define statistical hypothesis (2010)
2. Writ the procedure for testing a hypothesis (2008)
3. Define the error of sample & explain type-I error & type-II errors (2009)
4. Derive the one-tailed & two-tailed tests & critical region. (2009, 2011, 2012, 2017)
5. Write the procedure for testing of hypothesis. (2009, 2011, 2013)
6. What is meant by level of significance one-tailed and two-tailed tests. (2011, 2009)
7. Define null hypothesis. (2015)
8. Explain briefly the students 't' - test . (2014)
9. Explain the t-test for the equality of '2' means in small samples. (2014)
10. Explain briefly variance of F-Test. (2013)

Long Questions (minimum 10 previous JNTUH Questions – Year to be mentioned)

1.What is statistical hypothesis? write about , null hypothesis and alternative hypothesis.(nov2009)
2. discuss in brief level of significance .(nov2007,dec2010)
3. write about (nov2014,may 2016)

i).critical region
ii) two -tailed test
iii) one -tailed test
4. write short notes on large sample test of single proportion(nov2005,june2011)
5. Derive the large sample test procedure for difference of means(n0v05,dec09)
6. Explain large sample test procedure for testing the significance of single mean, A sample of 900 members has mean 3.4 cms .is the sample from the population with mean 3.25 cms and S.D.2.61 cms Also find 95% confidence limits.(nov12,dec09)
7.A company producing computers states that the mean lifetime of computers is 1600 hours . Test this claim at o.o1 L.O.S against the AH: μ<1600hours if 100 computer produced by this company has lifetime of 1570 hours with s.d of 120 hmay2010,nov2011)
8.samplee of students were drawn from two universities and from their weights in kilograms and S.D are calculated .Make a large sample test to test the significance of difference between the means(nov08)

	Mean	S.D	Size of the sample
University A	55	10	400
University B	57	15	100

9.Discuss on student ' t- test for difference between two means(mar 09,nov2011)
10. explain about F- distribution , its properties (nov2011,may2014)

Fill in the Blanks / Choose the Best: (Minimum 10 to 15 with Answers)

1.Standand deviation of small sample is ....S= $\sqrt{\frac{1}{(n-1)}\sum(x_i-\bar{x})^2}$

2.....Statistical hypothesis... are statements about the probability distributions of the populations
3.The two different types of hypothesis are ....Null Hypothesis... and ...Alternative Hypothesis...
4. ...Level of Significance...refers to the probability of containing a statistic random value 't' in the critical region.

5.The test statistic corresponding to one sample mean X is given by ....z = $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ .....

6.If a sample is drawn from a finite population of size N, then the observed proportion of

...S.E(p)= $\sqrt{\left[\frac{N-n}{N-1}\right]\times\frac{PQ}{n}}$ ...

7.σ is the ..Standard Deviation....of the population.
8.Normal test can be applied to the fundamental property of normal distribution i.e.

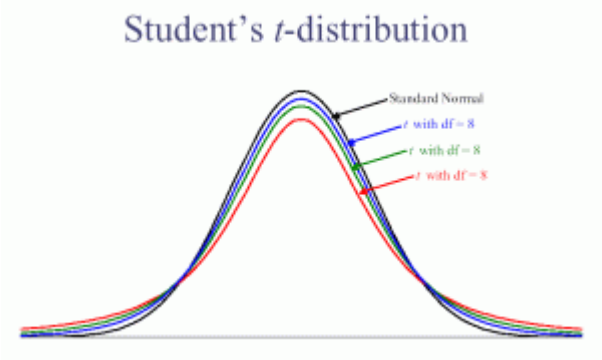
..z = $\frac{X-E(X)}{\sqrt{V(X)}}=\frac{X-nP}{\sqrt{nPQ}}$ ~N(0,1)....

9.The sampling distribution of small samples follow ..Student's t-distribution...
10. Null hypothesis is also known as....Simple Hypothesis....

T Test (Student's T-Test): Definition and Examples

Contents:

What is a T test?



The t-distribution, used for the t-test. Image: Carnegie Mellon.

The t test tells you how significant the differences between groups are; In other words it lets you know if those differences (measured in means/averages) could have happened by chance.

A very simple example: Let's say you have a cold and you try a naturopathic remedy. Your cold lasts a couple of days. The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week. You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you *really* want to know is, are these results repeatable? A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

Another example: Student's T-tests can be used in real life to compare means. For example, 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 (a group who are given a placebo, or "sugar pill"). The 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.

The T Score.

The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups. The smaller the t

score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different *from* each other as they are within each other. When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.

- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.

T-Values and P-values

How big is "big enough"? Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05. Low p-values are good; They indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

Calculating the Statistic / Test Types

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.

You probably don't want to calculate the test by hand (the math can get very messy, but if you insist you can find the steps for an independent samples t test here.

Use the following tools to calculate the t test:
How to do a T test in Excel.
T test in SPSS.
T distribution on the TI 89.
T distribution on the TI 83.

What is a Paired T Test (Paired Samples T Test / Dependent Samples T Test)?

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 — they are tests on the same person or thing. For example:

- Knee MRI costs at two different hospitals,
- Two tests on the same person before and after training,
- Two blood pressure measurements on the same person using different equipment.

When to Choose a Paired T Test / Paired Samples T Test / Dependent Samples T Test

Choose the paired t-test if you have two measurements on the same item, person or thing. You should also choose this test if you have two items that are being measured with a unique condition. For example, you might be 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.

With a "regular" two sample t test, you're comparing the means for two different samples. For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills. If you take a random sample each group separately and they have different conditions, your samples are independent and you 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 $\mu_1 = \mu_2$. In other words, it assumes the means are equal. With the paired t test, the null hypothesis is that the *pairwise difference* between the two tests is equal (H_0: $\mu_d = 0$). The difference between the two tests is very subtle; which one you choose is based on your data collection method.

Paired Samples T Test By hand

Sample question: Calculate a paired t test by hand for the following data:

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

Step 1: Subtract each Y score from each X score.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

Step 2: Add up all of the values from Step 1.
Set this number aside for a moment.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2
		SUM:	**-73**

Step 3: Square the differences from Step 1.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	

Step 4: Add up all of the squared differences from Step 3.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Step 5: Use the following formula to calculate the t-score:

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

ΣD: Sum of the differences (Sum of X-Y from Step 2)
ΣD^2: Sum of the squared differences (from Step 4)
$(\Sigma D)^2$: Sum of the differences (from Step 2), squared.

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -2.74$$

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this sample problem, with df=10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Note: You can ignore the minus sign when comparing the two t-values, as ± indicates the direction; the p-value remains the same for both directions.

F-Test

What is an F Test?

An "F Test" is a catch-all term for any test that uses the F-distribution. In most cases, when people talk about the F-Test, what they are actually talking about is The *F-Test to Compare Two Variances.* However, the f-statistic is used in a variety of tests including regression analysis, the Chow test and the Scheffe Test (a post-hoc ANOVA test).

General Steps for an F Test

If you're running an F Test, you should use Excel, SPSS, Minitab or some other kind of technology to run the test. Why? Calculating the F test by hand, including variances, is tedious and time-consuming. Therefore you'll probably make some errors along the way.

If you're running an F Test using technology (for example, an F Test two sample for variances in Excel), the only steps you really need to do are Step 1 and 4 (dealing with the null hypothesis). Technology will calculate Steps 2 and 3 for you.

1. State the null hypothesis and the alternate hypothesis.
2. Calculate the F value. The F Value is calculated using the formula F = (SSE_1 – SSE_2 / m) / SSE_2 / n-k, where SSE = residual sum of squares, m = number of restrictions and k = number of independent variables.
3. Find the F Statistic (the critical value for this test). The F statistic formula is:
   F Statistic = variance of the group means / mean of the within group variances.
   You can find the F Statistic in the F-Table.
4. Support or Reject the Null Hypothesis.

Back to Top

F Test to Compare Two Variances

A **Statistical F Test** uses an F Statistic to compare two variances, s_1 and s_2, by dividing them. The result is always a positive number (because variances are always positive). The equation for comparing two variances with the f-test is:
$F = s^2_1 / s^2_2$

If the variances are equal, the ratio of the variances will equal 1. For example, if you had two data sets with a sample 1 (variance of 10) and a sample 2 (variance of 10), the ratio would be 10/10 = 1.

You **always** test that the population variances are equal when running an F Test. In other words, you always assume that the variances are equal to 1. Therefore, your null hypothesis will always be that *the variances are equal*.

Assumptions

Several **assumptions** are made for the test. Your population **must be approximately normally distributed** (i.e. fit the shape of a bell curve) in order to use the test. Plus, the samples must be independent events. In addition, you'll want to bear in mind a few important points:

- The larger variance should always go in the numerator (the top number) to force the test into a right-tailed test. Right-tailed tests are easier to calculate.
- For two-tailed tests, divide alpha by 2 before finding the right critical value.
- If you are given standard deviations, they must be squared to get the variances.
- If your degrees of freedom aren't listed in the F Table, use the larger critical value. This helps to avoid the possibility of Type I errors.

Back to Top

F Test to compare two variances by hand: Steps

Warning: F tests can get really tedious to calculate by hand, especially if you have to calculate the variances. You're much better off using technology (like Excel — see below).

These are the general steps to follow. Scroll down for a specific example (watch the video underneath the steps).

Step 1: *If you are given standard deviations, go to Step 2. If you are given variances to compare, go to Step 3.*

Step 2: *Square both standard deviations to get the variances.* For example, if $\sigma_1 = 9.6$ and $\sigma_2 = 10.9$, then the variances (s_1 and s_2) would be $9.6^2 =$ **92.16** and $10.9^2 =$ **118.81**.

Step 3: *Take the largest variance, and divide it by the smallest variance to get the f-value.* For example, if your two variances were $s_1 = 2.5$ and $s_2 = 9.4$, divide 9.4 / 2.5 = **3.76**.
Why? Placing the largest variance on top will force the F-test into a right tailed test, which is much easier to calculate than a left-tailed test.

Step 4: Find your degrees of freedom. Degrees of freedom is your sample size minus 1. As you have two samples (variance 1 and variance 2), you'll have two degrees of freedom: one for the numerator and one for the denominator.

Step 5: *Look at the f-value you calculated in Step 3 in the f-table.* Note that there are several tables, so you'll need to locate the right table for your alpha level. Unsure how to read an f-table? Read What is an f-table?.

Step 6: Compare your calculated value (Step 3) with the table f-value in Step 5. If the f-table value is smaller than the calculated value, you can reject the null hypothesis.

That's it!
Back to Top

Two Tailed F-Test

The difference between running a one or two tailed F test is that the alpha level needs to be halved for two tailed F tests. For example, instead of working at $\alpha = 0.05$, you use $\alpha = 0.025$; Instead of working at $\alpha = 0.01$, you use $\alpha = 0.005$.

With a two tailed F test, you just want to know if the variances are not equal to each other. In notation:
$H_a = \sigma^2 1 \neq \sigma^2 2$

Sample problem: Conduct a two tailed F Test on the following samples:
Sample 1: Variance = 109.63, sample size = 41.
Sample 2: Variance = 65.99, sample size = 21.

Step 1: Write your hypothesis statements:
H_o: No difference in variances.
H_a: Difference in variances.

Step 2: Calculate your F critical value. Put the highest variance as the numerator and the lowest variance as the denominator:
F Statistic = variance 1/ variance 2 = 109.63 / 65.99 = 1.66

Step 3: Calculate the degrees of freedom:
The degrees of freedom in the table will be the sample size -1, so:
Sample 1 has 40 df (the numerator).
Sample 2 has 20 df (the denominator).

Step 4: Choose an alpha level. No alpha was stated in the question, so use 0.05 (the standard "go to" in statistics). This needs to be halved for the two-tailed test, so use 0.025.

Step 5: Find the critical F Value using the F Table. There are several tables, so make sure you look in the alpha = .025 table. Critical F (40,20) at alpha (0.025) = 2.287.

/	df_1=1	2	24	30	40	60	120	∞
df_2=1	647.7890	799.5000	97.2492	1001.414	1005.598	1009.800	1014.020	1018.258
2	38.5063	39.0000	39.4562	39.465	39.473	39.481	39.490	39.498
3	17.4434	16.0441	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.6491	8.5109	8.461	8.411	8.360	8.309	8.257
5	10.0070	8.4336	6.2780	6.227	6.175	6.123	6.069	6.015
16	6.1151	4.6867	2.6252	2.568	2.509	2.447	2.383	2.316
17	6.0420	4.6189	2.5598	2.502	2.442	2.380	2.315	2.247
18	5.9781	4.5597	2.5027	2.445	2.384	2.321	2.256	2.187
19	5.9216	4.5075	2.4523	2.394	2.333	2.270	2.203	2.133
20	5.8715	4.4613	2.4076	2.349	2.287	2.223	2.156	2.085

Step 6: Compare your calculated value (Step 2) to your table value (Step 5). If your calculated value is higher than the table value, you can reject the null hypothesis:
F calculated value: 1.66
F value from table: 2.287.
1.66 < 2 .287.
So we cannot reject the null hypothesis.
Back to Top

F-Test to Compare Two Variances in Excel

Watch the video or read the steps below:

F-test two sample for variances Excel 2013: Steps

Step 1: Click the “Data” tab and then click “Data Analysis.”
Step 2: Click “F test two sample for variances” and then click “OK.”
Step 3: Click the Variable 1 Range box and then type the location for your first set of data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box.
Step 4: Click the Variable 2 box and then type the location for your second set of data. For example, if you typed your data into cells B1 to B10, type “B1:B10” into that box.
Step 5: Click the “Labels” box if your data has column headers.
Step 6: Choose an alpha level. In most cases, an alpha level of 0.05 is usually fine.
Step 7: Select a location for your output. For example, click the “New Worksheet” radio button.
Step 8: Click “OK.”
Step 9: Read the results. If your f-value is higher than your F critical value, reject the null hypothesis as your two populations have unequal variances.

Warning: Excel has a small "quirk." Make sure that variance 1 is higher than variance 2. If it isn't switch your input data around (i.e. make input 1 "B" and input 2 "A"). Otherwise, Excel will calculate an incorrect f-value. This is because the variance is a ratio of variance 1/variance 2, and Excel can't work out which set of data is set 1 and set 2 without you explicitly telling it.