Contents

[Normal distribution 2](#_Toc521529573)

[Binomial distribution 2](#_Toc521529574)

[Exponential 3](#_Toc521529575)

[Poisson distribution 4](#_Toc521529576)

[Uniform distribution 5](#_Toc521529577)

[T distribution 5](#_Toc521529578)

[F distribution 7](#_Toc521529579)

[Numericals 7](#_Toc521529580)

[T distribution 7](#_Toc521529581)

[Normal Distribution 8](#_Toc521529582)

[Binomial distribution 15](#_Toc521529583)

[Exponential 19](#_Toc521529584)

[Poisson Distribution 20](#_Toc521529585)

[Interview Questions 24](#_Toc521529586)

[Reasons for the Non Normal Distribution 24](#_Toc521529587)

[Dealing with Non Normal Distributions 25](#_Toc521529588)

[What is law of large numbers 25](#_Toc521529589)

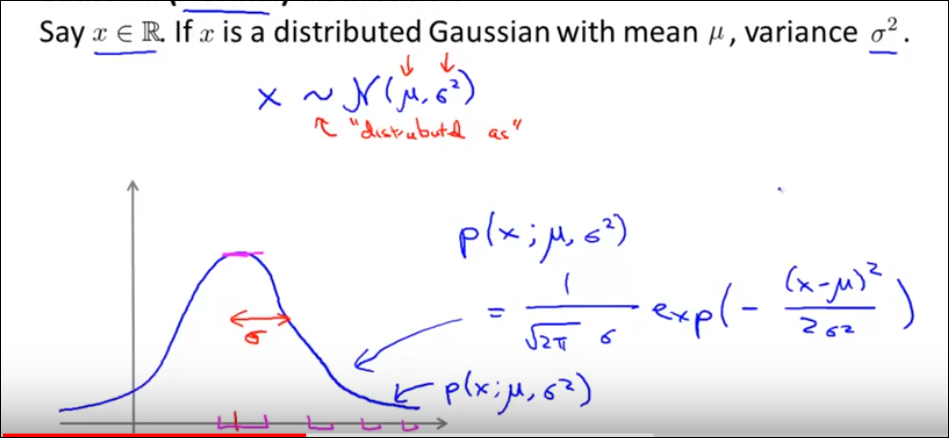
|  |  |  |
| --- | --- | --- |
|  | Expected value | Variance |
| Binomial | Np | Np(1-p) |
| Normal | Mu | Sigma2 |
| Exponential | 1/ λ | 1/ λ2 |
| Uniform | (a+b)/2 | (b-a)^2/12 |
| Poisson | μ | μ |
|  |  |  |
|  |  |  |
|  |  |  |

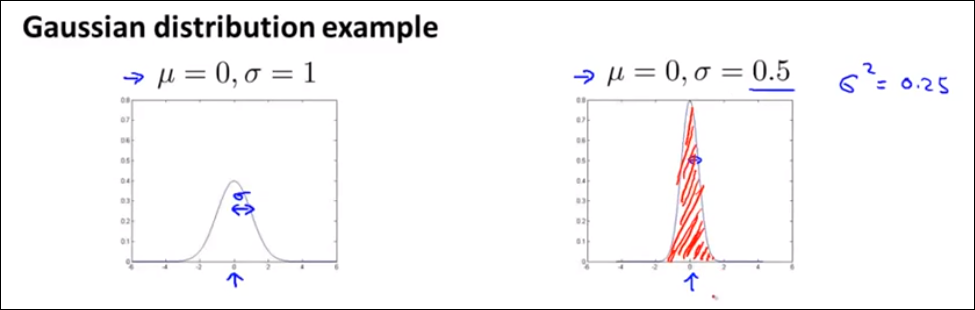
|  |
| --- |
| **Problem 1**: (4 pts each, *Circle the correct answer*) Classify the random variable as finite discrete (FD), infinite discrete (ID), or continuous (C).  The number of rolls of three dice until at least two of them produce 6,  **ID**  The product of the outcomes of rolling three dice, **FD**  m  The length of time from when the dice are released until the first one hits the ground, C  Time between arrival of 2 customers C  Tomorrow’s temperature at high noon C    The largest value obtained when rolling three dice. **FD** |

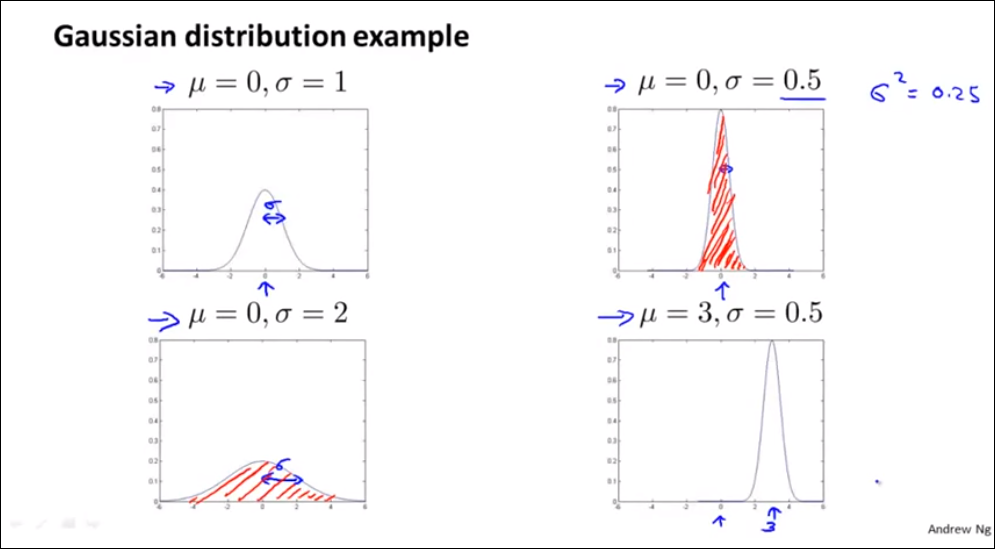
# Normal distribution

Normal distribution are always continuous

Width of normal distribution is sigma

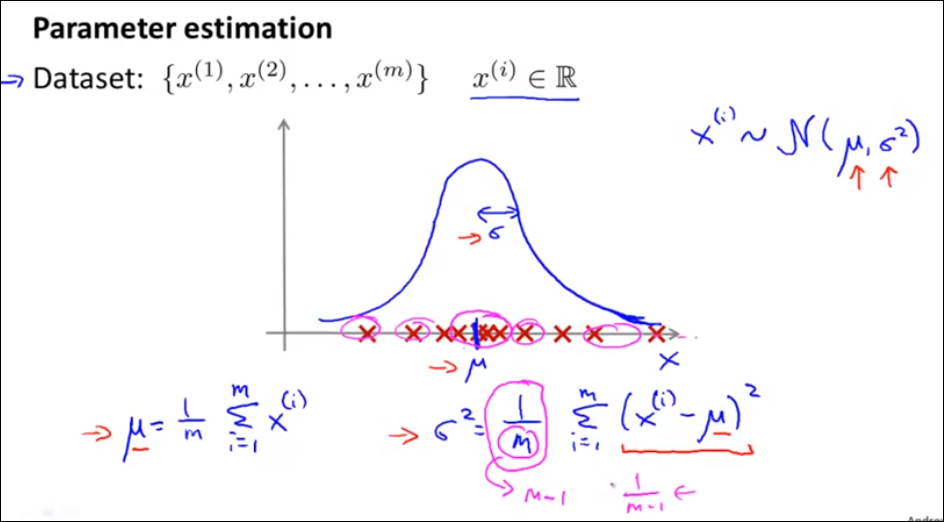






Area under the curve will remain 1

**Parameter estimation**



**To check if data is normal – mean should be equal to median , skewness should be equal to 0 or less than 1 ,kurtosis should be almost equal to 3 , and sd should be 1**

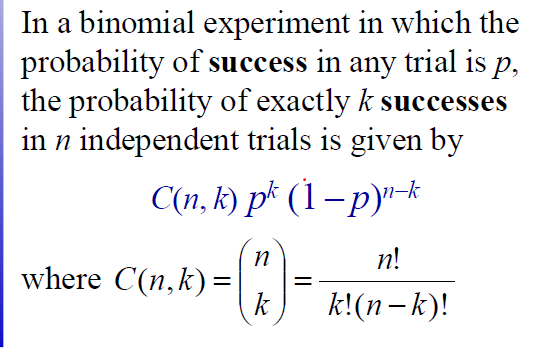
# Binomial distribution

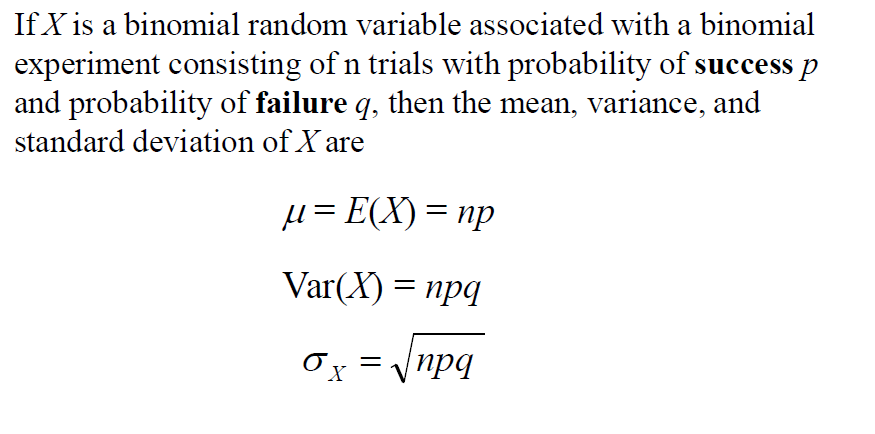
1. Number of trials in the experiment is fixed,

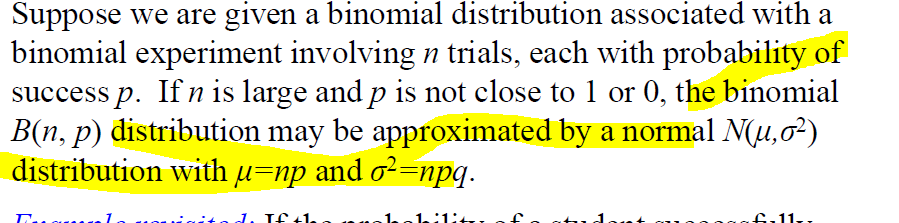
2. The only outcomes are success and failure,

3. In each trial the success probability is the same, and

4. The trials are independent of each other.

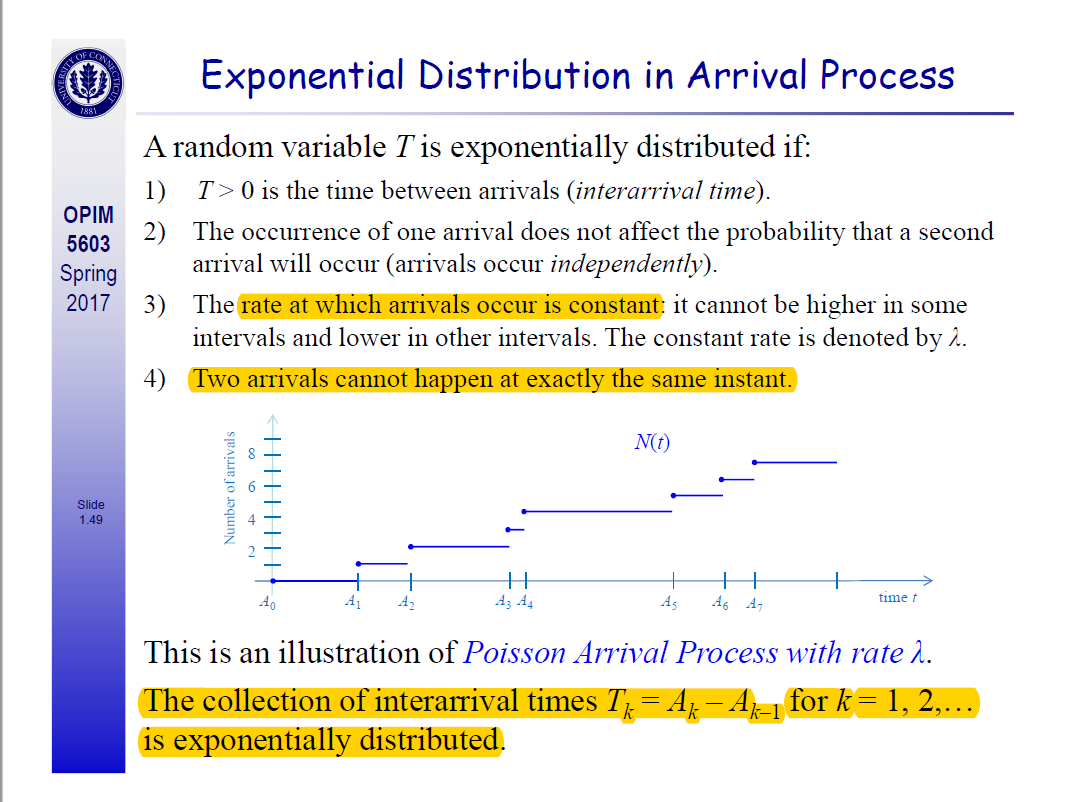






# Exponential

|  |  |  |
| --- | --- | --- |
| Exponential | Expected value =1/ λ | Variance=1/ λ2 |

* Interarrival time forms the exponential distribution
* It defines the time between 2 events(2 success)
* Graph will be x axis , y will be probability
* It is memoryless
* 

# Poisson distribution

* When we have to find how many have arrived between a particular time then it becomes poisson
* E(x) =λ(end time -start time ) ( expected number of arrivals is lam\*(et-st))

The Poisson random variable satisfies the following conditions:

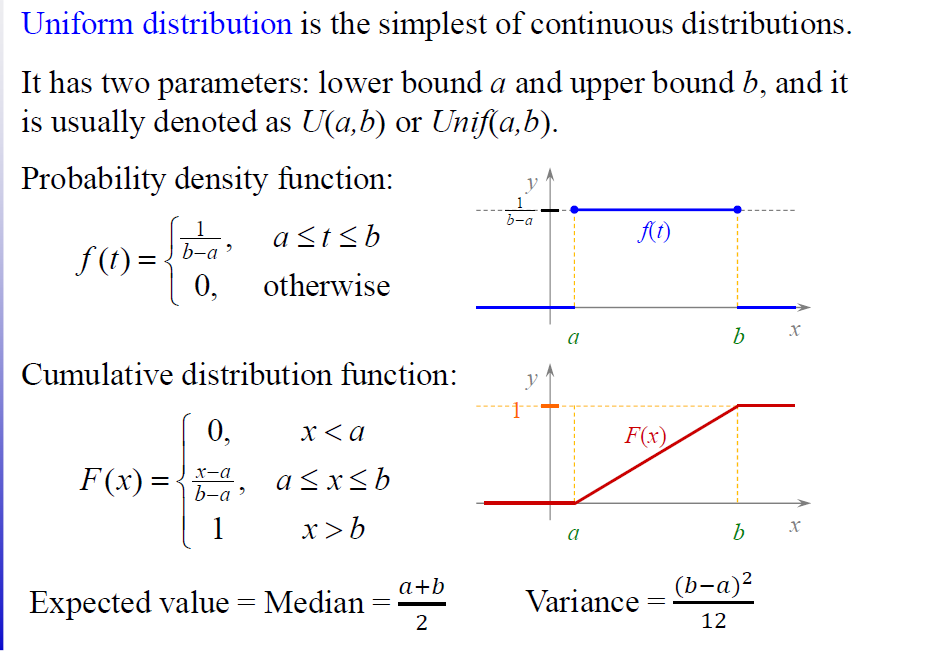
1. The number of successes in two disjoint time intervals is independent.
2. The probability of a success during a small time interval is proportional to the entire length of the time interval.

Apart from disjoint time intervals, the Poisson random variable also applies to disjoint regions of space.

Applications

* the number of deaths by horse kicking in the Prussian army (first application)
* birth defects and genetic mutations
* rare diseases (like Leukemia, but not AIDS because it is infectious and so not independent) - especially in legal cases
* car accidents
* traffic flow and ideal gap distance
* number of typing errors on a page
* hairs found in McDonald's hamburgers
* spread of an endangered animal in Africa
* failure of a machine in one month

# Uniform distribution



Consider the toss of a single die. The outcome of this toss is a random variable that can take on any of six possible values: 1, 2, 3, 4, 5, or 6. Each of these outcomes is equally likely to occur. The probability that any particular outcome will occur is equal to 1/6. Therefore, the outcome from the toss of a single die has a uniform distribution

. Graphically, a uniform distribution has no clear peaks.

# T distribution

According to the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem), the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of a statistic (like a sample mean) will follow a [normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution), 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](http://stattrek.com/Help/Glossary.aspx?Target=Z-score), and use the normal distribution to evaluate probabilities with the sample mean.

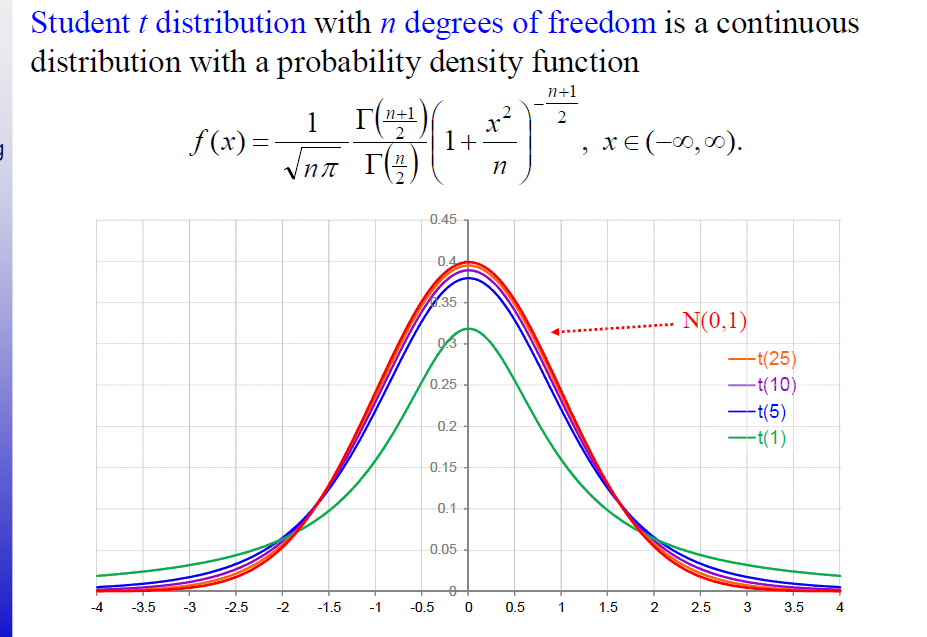
But sample sizes are sometimes small, and often we **do** **not** **know** the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic**(also known as the **t score**), whose values are given by:

t = [ x - μ ] / [ s / sqrt( n ) ]

**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.

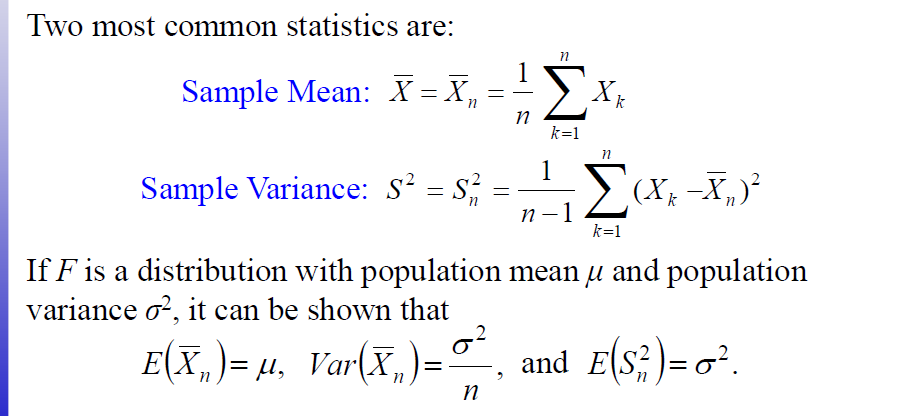
When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the *t* statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly,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

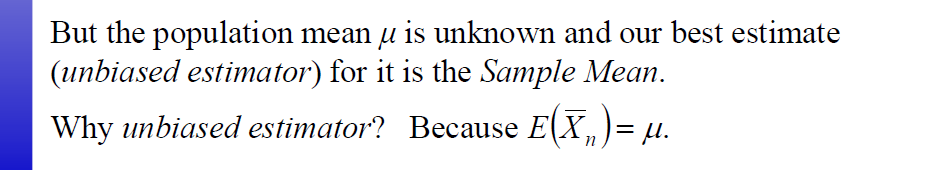


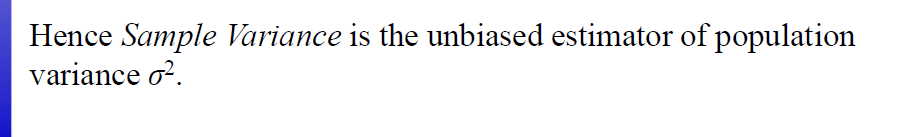
# F distribution

The F distribution is actually a collection of distribution curves. The F distribution is related to [chi-square](http://www.statisticshowto.com/chi-square/), because the f distribution is the ratio of two [chi-square distributions](http://www.statisticshowto.com/chi-square-distribution/) with [degrees of freedom](http://www.statisticshowto.com/degrees-of-freedom/) ν1 and ν2 (note: each chi-square is first been divided by its [degrees of freedom](http://www.statisticshowto.com/degrees-of-freedom/)). Each curve depends on the [degrees of freedom](http://www.statisticshowto.com/degrees-of-freedom/) in the numerator (dfn) and the denominator (dfd). These depend upon your sample characteristics.  
For example, in a simple one-way ANOVA between-groups,  
Dfn = a-1  
dfd = N-a  
where:  
a = the number of groups  
n = the total number of subjects in the experiment

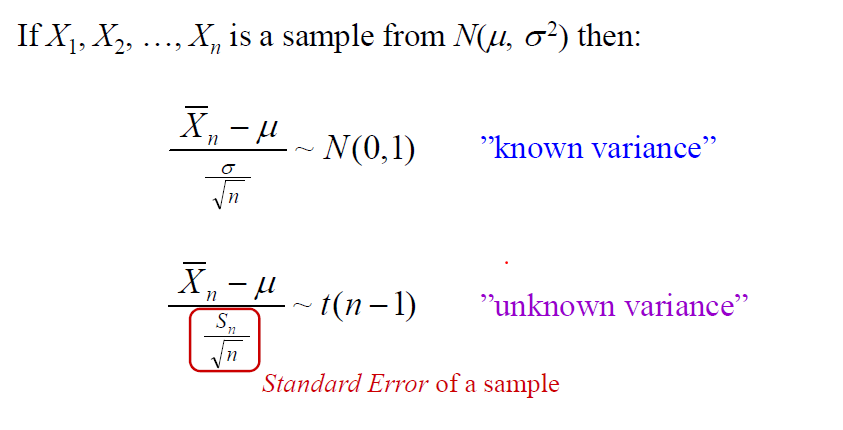
# Sample Distribution



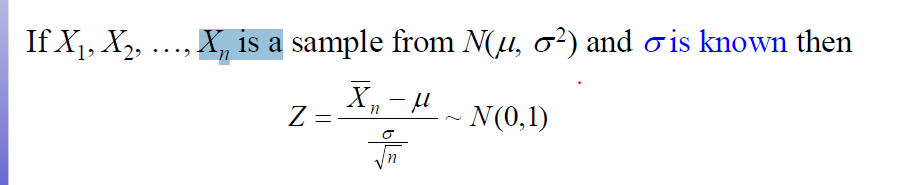


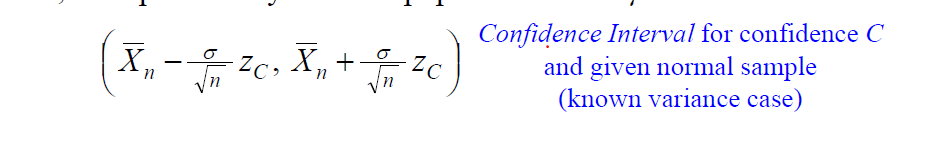


|  |
| --- |
| Problem 5: (5 pts) If X1, X2, …, X7 is a sample from N(17, 25), what is the distribution of its Sample Mean?  Since the variation is known the distribution of the sample would be **normal**  And mean and variance would be    So, distribution of the sample mean would be -:  N(17,25/n) = N(17,25/7) |

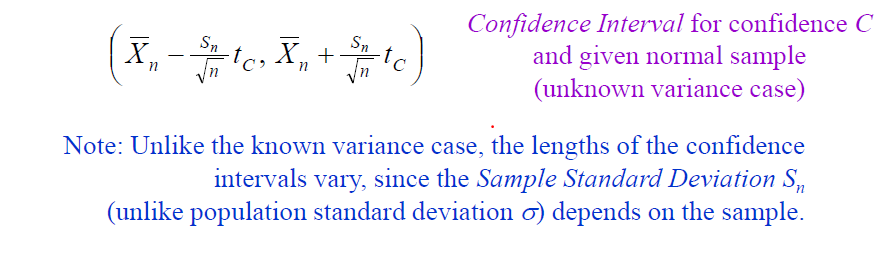


**For known variance**





**For unknown variance**



Unlike the known variance case, the lengths of the confidence

intervals vary, since the Sample Standard Deviation Sn

(unlike population standard deviation σ) depends on the sample.

Population mean μ will be inside approximately 950 of those

intervals, and outside of approximately 50 of them.

|  |
| --- |
| 1. Suppose a simple random sample of size *n* = 34 is obtained from a population with *μ* = 30 and *σ* = 4. Describe the sampling distribution of . 2. Approximately normal with  and 3. Approximately normal with  and . 4. Approximately normal with  and 5. Approximately normal with  and .   Sample mean = population mean , sample variance = population variance /n |

|  |
| --- |
| The high temperature in Chicago for the month of August is approximately normally distributed with mean *μ* = 80° F and standard deviation *σ* = 8° F. Use this information to answer the next two questions.  1. What is the probability that on any given day in August, the high temperature is above 90°F?    (a) 0.0125  (b) 0.8944  (c) 0.1056  (d) 0.0001    1-pnorm(90,80,8)  2. What is the probability that a random sample of 16 days in August have a sample mean,  high temperature of at least 82° F?    (a) 0.5987  (b) 0.2500  (c) 0.1587  (d) Cannot be determined. |

|  |
| --- |
| 1. A Census Bureau report on the income of Americans says that with 95% confidence the mean income of all U.S. households in 2006 was $57,010 with a margin of error of ±$350. This means that 2. 95% of all households had incomes in the range $57,010 ± $350. 3. We can be sure that the mean income for all households in the country lies in the range $57,010 ± $350. 4. We are 95% confident that the mean household income of all Americans will be in the range $57,010 ± $350 5. 95% of the households in the sample interviewed by the Census Bureau had incomes in the range $57,010 ± $350.   C |

|  |
| --- |
|  |

1. When estimating a population mean by a sample mean, which of the following does the margin of error not depend on?
2. The sample size
3. The sample mean
4. The confidence level
5. The population standard deviation

Ans sample mean

1. Keeping everything else the same, which statement would result in a larger margin of error?
2. A smaller confidence level.
3. A smaller sample size.
4. A larger sample mean.
5. A smaller standard deviation.

|  |
| --- |
| A savings and loan association needs information concerning the checking account balances of its local customers. A random sample of 14 accounts was checked and yielded a mean balance of $664.14 and a standard deviation of $297.29. Find a 90% confidence interval for the true mean checking account balance for local customers.  A) $492.52 to $835.76 B) $493.71 to $834.57  C) $523.43 to $804.85 D) $455.65 to $872.63  Answer: C  Objective: (8.4) Find Confidence Interval for Mean (Sigma Unknown)  0.90 =P(-tc <T<tc)  = 0.90  n=14  conf =0.90  sn=297  SM=664    tC = qt(conf+(1-conf)/2,n-1) # tC value for the given confidence    width = tC\*sn/sqrt(n) # half-width of the confidence interval  endpts = c(SM-width,SM+width) # confidence interval end points |

# Numericals

## T distribution

**Problem 2**: (6 pts each) Let *fn* denote the pdf of the Student *t*(*n*) distribution (with *n* degrees of freedom).

1. Write an *R* function *vt*(*x*,*n*) that returns the sequence (vector) *f*1(*x*), *f*2(*x*), …, *fn*(*x*).

vt= function(x,n)

{

dt(x,seq(1,n))

}

1. Write a function *dnt*(*x*,*n*) that returns the sequence *f*(*x*) – *f*1(*x*), *f*(*x*) – *f*2(*x*), …, *f*(*x*) – *fn*(*x*), where *f* is the pdf of the standard normal distribution. You can use the function from part (a) if you find it useful.

dnt = function(x,n){

dnorm(x) - dt(x,seq(1,n))

}

1. Calculate *dnt*(*x*,100) for three values of *x* of your choice? What do you observe?

dnt(-5,100) = value is negative

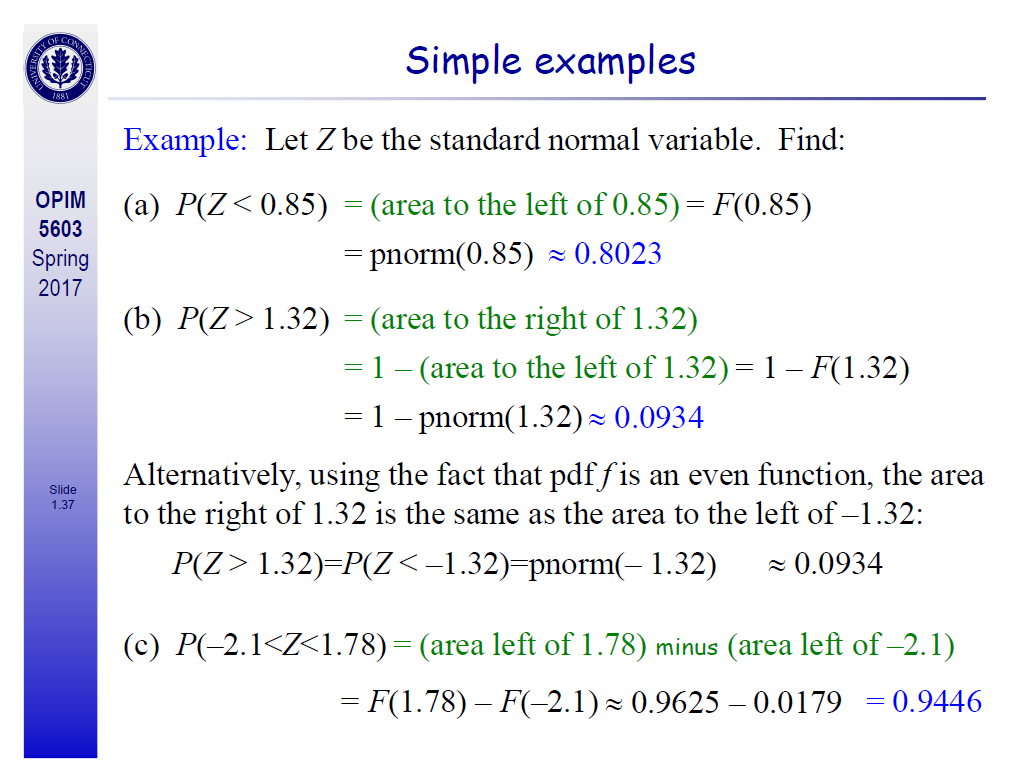
dnt(0,100) = value is positive

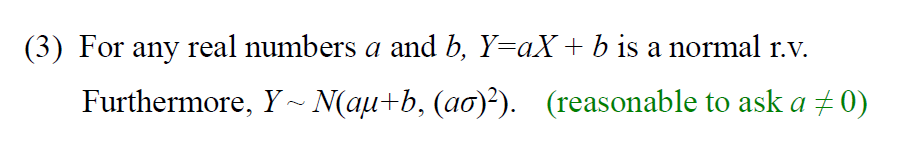
dnt(5,100) = value is negative

**observation**

* Normal pdf has shorter tails than t distribution . T distribution will have broader curves
* But normal distribution has higher peak
* As degree of freedom → ∞ , T distribution will converge to N(0,1)

## Normal Distribution





|  |
| --- |
| ***Questio 7***  **Problem 3**: (10 pts) Let *X*~ *N*(3,4). What is the distribution of the random variable 2*X* – 1?  That function will be normal distribute  X ~ N(μ, σ2)  Y(2X-1) ~ N(aμ+b, (aσ)2)  Y(2X-1) ~ N(2(3) +(-1) , (2\*2)2  N(μ, σ2) = N(5,16)  E(2X – 1) = 2EX – 1 = 2・ 3 – 1 = 5  Var(X) = 16 |

|  |
| --- |
| ***Questio 8***  A particular rash has shown up at an elementary school. It has  been determined that the length of time that the rash will last is  normally distributed with μ = 6 days and σ = 1.5 days. Find the  probability that for a student selected at random, the rash will  last for less than 3 days.  P(X < 3)=P(N(6,1.52) < 3)  Pnorm(2,6,1.5) |

|  |
| --- |
| 54. The expected (mean) life of a particular type of light bulb is 1,000 hours with a standard deviation of 50 hours. The life of this bulb is normally distributed. What is the probability that a randomly selected bulb would last longer than 1150 hours?  a) 0.4987  b) 0.9987  c) 0.0013  d) 0.5013  e) 0.5513     1. Pnorm(3)   1-pnorm(1150,1000,50) |

|  |
| --- |
| 43. Let z be a normal random variable with mean 0 and standard deviation 1. What is P(1.3 < z < 2.3)?  a) 0.4032  b) 0.9032  c) 0.4893  d) 0.0861  e) 0.0086  pnorm(2.3) - pnorm(1.3) =0.086 |

|  |
| --- |
| . Let z be a normal random variable with mean 0 and standard deviation 1. What is P(z > -1.1)?  z > -1.1 => z<=F(-1.1) => 1- pnorm(-1.1) |

|  |
| --- |
| 47. Let z be a normal random variable with mean 0 and standard deviation 1. What is  P(-2.25 < z < -1.1)?  pnorm(-1.1) - pnorm(-2.25) |

|  |
| --- |
| A standard normal distribution has the following characteristics:  a) the mean and the variance are both equal to 1  b) the mean and the variance are both equal to 0  c) the mean is equal to the variance  d) the mean is equal to 0 and the variance is equal to 1  e) the mean is equal to the standard deviation  Ans: d  Response: See section 6.2, Normal Distribution  Difficulty: Medium |

|  |
| --- |
| 58. Sure Stone Tire Company has established that the useful life of a particular brand of its automobile tires is normally distributed with a mean of 40,000 miles and a standard deviation of 5000 miles. What is the probability that a randomly selected tire of this brand has a life of at most 30,000 miles?  Atomost 30,000 => F(x<=30000)  pnorm(30000,40000,5000)  If it is discrete F(x<=30000-1) |

|  |
| --- |
| A manufacturer of a special type of one-size glove wants to design the glove to fit at least 99 percent of the population. Hand span is known to be normally distributed with a mean of 195 millimeters and a standard  deviation of 17 millimeters. What range of hand spans must the glove accommodate?  Here probability is given to us which is 99%.  So what is left is 1% together on both side  So 0.5% on each side  To find the value below which 0.5 percent of the population falls, use the command:  > qnorm(0.005, 195, 17)  [1] 151.2109  Similarly, to find the value above which 0.5 percent of the population falls, use the command:  > qnorm(0.005, 195, 17, lower.tail=F)  Or  qnorm(.995, 195, 17)  [1] 238.7891  The remaining 99 percent of the population falls between these two values. So, to accommodate 99 percent of  the population, the gloves must be designed to fit hands with a span between 151 and 239 millimeters. |

|  |
| --- |
| 1. P(X<70) => pnorm(70,79,sqrt(144)) 2. P(64<X<96) => Pnorm(96) -Pnorm(64) |

|  |
| --- |
| Since 20% of the members stay for more than 125 minutes, 80% stay for less than or equal to 125 minutes. For a standard normal distribution,  qnorm(.80) will give the z score so,  80% of the area under the curve lies to the left of z = 0.841621234. We know that  z = (x - µ)/σ   0.841621234 = (125 - 90)/σ   σ = (125 - 90)/0.841621234 = 35/0.841621234 = 41.58640325  b) Now that we know sigma, we can find the z score for 25 minutes,  z = (25 - 90)/41.58640325 = -65/41.58640325 = -1.563010862  The area to the left of that z score is 0.05902502, so that's the probability that a visit will last less then 25 minutes.  **pnorm(25,90,41.5)**  c) Since Tara arrives at 8 PM, she has only two hours (120 minutes) to use the facility. The z score for 120 minutes is  z = (120 - 90)/41.58640325 = 30/41.58640325 = 0.721389629  The area under the curve to the left of that z score is 0.764665087, so we're losing 0.235334913, the area to the right. In other words, our distribution has about 23.5% of its area chopped off, so it's not a suitable model. |

|  |
| --- |
| The random variable X is normally distributed with mean 79 and variance 144: X ~ N (79,144)    Total area is 1  So we 1-0.6463/3 =0.1179  Sd=12 given  qnorm(0.2358,79,12,lower.tail=F) =79+b  b =8.63 |

|  |
| --- |
|  |

## Binomial distribution

|  |
| --- |
| ***Questio 5***  Five cards are drawn, with replacement, from a standard 52-card  deck. If drawing a club is considered a success, find the mean,  variance, and standard deviation of X (where X is the number of  successes).  Number of trials are fixed  It is a binomial distribution as each card is either a club or not.  And since it is done with replacement every experiment is independent of other experiment  And probability in each experiment is same  So p =1/4 , q=1-1/4 , mean = np, var = npq |

|  |
| --- |
|  |

|  |
| --- |
| **Ten percent of computer parts produced by a certain supplier are defective. What is the probability that a sample of 10 parts contains more than 3 defective ones?**  n= 10,x=3 ,P = probability of success - .10  P(X > 3) = 1 − F(3)  1- pbinom(3,10,1/10)  If the question is atleast 3 parts  P(x>=3) 1-F(2) |

|  |
| --- |
| 1. dbinom(8,8,0.82) 2. dbinom(0,8,0.82) 3. atleast 6 pass =P(x>=6) = 1- Prob(x<=5)   1-pbinom(5,8,0.82) |

|  |
| --- |
| The Stanley Cup winner is determined in the final series between two teams. The  first team to win 4 games wins the Cup. Suppose that Dallas Stars advance to the final series, and they have a probability of 0.55 to win each game, and the game results are independent of each  other. Find the probability that  a) Dallas Stars wins the Stanley Cup  b) seven games are required to determine the Cup winner  (Hint: Without loss of generality, you can assume that the series continues until 7 games are played, even if the Cup winner is determined earlier. This ”change of Stanley Cup rules” will not change the answer to the problem!)  Suppose that the series continues until Dallas Stars win 4 games, even if the other rival wins the Cup earlier.   1. n = 7 , p = 0.55,x=3   P( Dallas wins ) = P(X >=4) = 1 - F(3) = 0.6083     1. pbinom(x,n,p) 2. First team to win 4 games will win the tournament so if 7 games are required to determine te game winner that means no team must have won 4 games   n = 6 ; p = 0.55 ;x=3  pbinom(3,n,p) - pbinom(2,n,p) |

|  |
| --- |
| **Suppose you toss a fair coin 12 times. What is the probability of getting exactly 7 Heads.**  X=7  N=12  P=1/2  dbinom(7,12,1/2)  atmost 7 heads  pbinom(7,12,1/2) |

|  |
| --- |
| ) An internet search engine looks for a certain keywordin a sequence of independent web sites. It is believed that 20% of the sites contain this keyword.  c) Out of the first 10 websites, let Y be the number of sites that contain the keyword. Find the distribution of Y .  d) Compute the expected value and the standard deviation of Y .  e) Compute the probability that at least 5 of the first 10 websites contain the keyword.  f) Compute the probability that the search engine had to visit at least 5 sites inorder to find the first occurrence of a keyword.  c) Y is Binomial(n = 10; p = 0:2).  d) E(Y ) = np = 2 and Std(Y ) =sqrt( np(1-p))  e) From the Binomial Table, P(X >=5) = 1- F(4) = 1 - 0:9672 = 0.0328 .  f) P(Y >= 5) = (1 – p(4) = 0.4096 : |

|  |
| --- |
| **Extra Credit**: (10 pts) Coin is tossed 11 times and 2 Heads and 9 Tails are recorded. The null hypothesis is that the coin is fair. Do you accept it at 5% significance level?  Hint: Take the number of Heads as your statistic of interest. Can you compute the *p value* for this statistic? Explain your reasoning.  H0 = coin is fair (p=0.5)  Ha = coin is not fair(p ≠ 0.5)  Binomial distribution  Pval = 2\*(pbinom(2,11,0.5)) = 0.06542968  Or we can just do  binom.test(2,11,0.5)  p-value = 0.06543  so we will accept the null hypothesis **, coin is fair** |

## Exponential

|  |
| --- |
| Problem 3: (5 pts) Let X~ Exp(0.3). Write a single line in R that calculates P(1 < X < 4) and execute it. Paste (or rewrite) the textual content of your RStudio console as the solution.    P(1<x<4) = P(X<=4) – P(X<=1)  pexp(4,rate=0.3)-pexp(1,rate=0.3)  [1] 0.439624 |

|  |
| --- |
| Malfunctions in a particular type of electronic device are known to follow an exponential distribution with a mean time of 24 months until the device malfunctions. What is the probability that a randomly selected device will malfunction within the first 6 months?  Expected value =1/lambda  Lambda =1/expected value  You can answer the question using the pexp function  > pexp(6, 1/24) |

|  |
| --- |
| Malfunctions in a particular type of electronic device are known to follow an exponential distribution with a mean  time of 24 months until the device malfunctions.  The probability of malfunction within six months is 0.22 (22%).  What is the probability that a randomly selected device will last more than 5 years (60 months) without malfunction?  > pexp(60, 1/24, lower.tail=F)  Or 1-pexp(60,1/24)  [1] 0.082085  The probability that it will last more than 5 years is approximately 0.08, or 8 percent. |

|  |
| --- |
| Malfunctions in a particular type of electronic device are known to follow an exponential distribution with a mean  time of 24 months until the device malfunctions. After how many months will 40 percent of the devices already  have malfunctioned?  E(x) = 1/lambda  To find the length of time within which 40 percent of devices will have malfunctioned, use the command:  > qexp(0.4, 1/24)  [1] 12.25981  So 40 percent of devices will malfunction within 12.3 months. |

## Poisson Distribution

|  |
| --- |
| **Problem 4**: Arrival of cars at New Jersey Turnpike toll booth on Saturdays between 3AM and 5AM is modeled as a *Poisson Arrival* process with a rate 2.1 per minute. Let *X* be a random variable that counts the arrivals between 4:00AM and 4:10AM.  (a)(5 pts) Is this variable finite discrete, infinite discrete, or continuous?  Infinite discrete  (b)(5pts) What is the distribution of X?  It is a **Poisson** distribution since all the cars are independent events with Pois(λ ∆t) = Pois(21)  Probability distribution = P(N=k) = e -μ μk/k! = e -2121 k/k !  (c)(5 pts) What is the expected value of X?  E(x) =λ(end time -start time ) = 2.1\*10 =21 |

|  |
| --- |
| 2= mu = λ ∆t  ∆t =5  Lambda =0.4  Atmost 5 errors = Pois(5, mu)  So deltaT =15  Mu = 15 \*0.4 =6  Ppois(5,6) |

|  |
| --- |
| The switchboard in a Denver law office gets an average of 2.5 incoming phonecalls during the noon hour on Thursdays. Experience shows that the existing lunch hour staff can handle up to 5 calls in an hour. They seem to be well covered. But just to test the edges, what is the actual chance that 6 calls will be received during the lunch period, on some particular Thursday?  dpois(6,2.5\*1) |

|  |
| --- |
| Suppose that the number of inquiries arriving at a certain interactive system follows  a Poisson distribution with arrival rate of 12 inquiries per minute.  Find the probability of 10 inquiries arriving  a) in a 1-minute interval;  b) in a 3-minute interval.  c. What is the expectation and the variance of the number of arrivals during each of these  intervals?  λ = 12/min  mu = λ ∆t = 12\*1 =12  x = 10  dpois(10,12)      λ = 12/min  mu = λ ∆t = 12\*3 =36  x = 10  dpois(10,36)   1. . For Poisson distribution with parameter mu, we have E(X) = V ar(X) = mu. |

|  |
| --- |
| To find 3 more flaws  Probability=*P*(*X*≥3) = 1-ppois(2,2.3) |

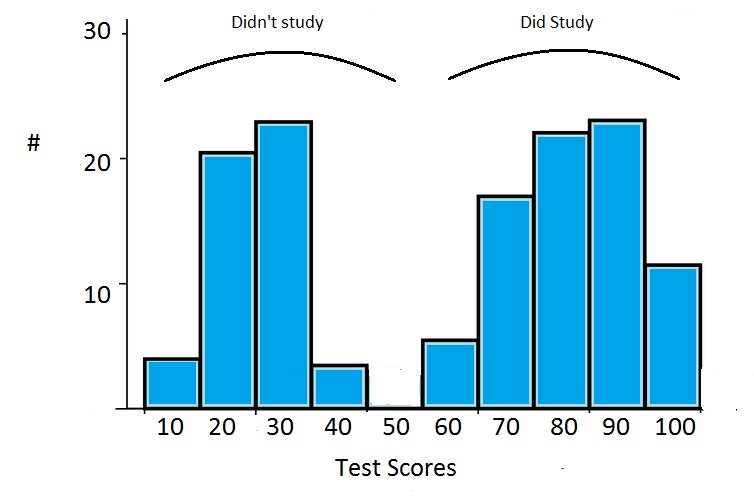
|  |
| --- |
|  |

|  |
| --- |
| To calculate the rate mu = lambda delta t  3 = lambda (20)  Lambda =3/20  Ppois(1,3/20) |
|  |

|  |
| --- |
| **Problem 4**: Arrival of cars at New Jersey Turnpike toll booth on Saturdays between 3AM and 5AM is modeled as a *Poisson Arrival* process with a rate 2.1 per minute. Let *X* be a random variable that counts the arrivals between 4:00AM and 4:10AM.  (a)(5 pts) Is this variable finite discrete, infinite discrete, or continuous?  Infinite discrete  (b)(5pts) What is the distribution of X?  It is a **Poisson** distribution since all the cars are independent events with Pois(λ ∆t) = Pois(21)  Probability distribution = P(N=k) = e -μ μk/k! = e -2121 k/k !  (c)(5 pts) What is the expected value of X?  E(x) =λ(end time -start time ) = 2.1\*10 =21 |

# Interview Questions

## Reasons for the Non Normal Distribution

* [**Outliers**](http://www.statisticshowto.com/find-outliers/)can cause your data the become skewed. The [mean](http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)is especially sensitive to outliers. Try removing any extreme high or low values and testing your data again.
* **Multiple distributions may be combined in your data**, giving the appearance of a [bimodal](http://www.statisticshowto.com/what-is-a-bimodal-distribution/)or [multimodal distribution](http://www.statisticshowto.com/multimodal-distribution/). For example, two sets of normally distributed test results are combined in the following image to give the appearance of bimodal data.  
  [](http://www.statisticshowto.com/wp-content/uploads/2013/07/bimodal-distribution-2.jpg)

* **Insufficient Data**can cause a normal distribution to look completely scattered. For example, classroom test results are usually normally distributed. An extreme example: if you choose three random students and plot the results on a graph, you won’t get a normal distribution. You might get a [uniform distribution](http://www.statisticshowto.com/uniform-distribution/) (i.e. 62 62 63) or you might get a [skewed distribution](http://www.statisticshowto.com/probability-and-statistics/skewed-distribution/) (80 92 99). If you are in doubt about whether you have a sufficient [sample size](http://www.statisticshowto.com/probability-and-statistics/find-sample-size/), collect more data.
* **Data may be inappropriately graphed**. For example, if you were to graph people’s weights on a scale of 0 to 1000 lbs, you would have a skewed cluster to the left of the graph. Make sure you’re graphing your data on appropriately labeled axes.

## Dealing with Non Normal Distributions

You have several options for handling your non normal data. Several tests, including the[one sample Z test](http://www.statisticshowto.com/one-sample-z-test/), [T test](http://www.statisticshowto.com/probability-and-statistics/t-test/)and [ANOVA](http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/)assume normality. You may still be able to run these tests if your [sample size](http://www.statisticshowto.com/probability-and-statistics/find-sample-size/) is large enough (usually over 20 items).

You can also choose to transform the data with a function, forcing it to fit a normal model. However, if you have a very small [sample](http://www.statisticshowto.com/sample/), a sample that is skewed or one that naturally fits another distribution type, you may want to run a [non parametric test](http://www.statisticshowto.com/parametric-and-non-parametric-data/).

A non parametric test is one that doesn’t assume the data fits a specific distribution type. Non parametric tests include the [Wilcoxon signed rank test](http://www.statisticshowto.com/wilcoxon-signed-rank-test/), the [Mann-Whitney U Test](http://www.statisticshowto.com/mann-whitney-u-test/) and the Kruskal-Wallis test.

See also:

* [How to Use Chebyshev’s Thoerem (Video)](http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/chebyshevs-theorem-inequality/)
* [How to Use Slovin’s Formula (video)](http://www.statisticshowto.com/how-to-use-slovins-formula/)
* [The Mean of the Sampling Distribution for a Proportion.](http://www.statisticshowto.com/sampling-distribution/#MeanSDP)

"Data" can never be normal; the normality assumption does \*not\* refer to the observed data. Rather, the assumption is that the \*process\* that produces the data is a normally distributed process.  And that process, likewise, can never be precisely normal, because of asymmetries, discreteness, and boundedness of the observable data. So the question is never "normal or not". It is always "how different from normal is the distribution that produces the data?" When that distribution is close to normal, the methods should be fine. And, depending on the method and sample size, the methods may be ok even when the distribution is far from normal because of the Central Limit Theorem. There is no "one size fits all" answer, rather the answer is always, "it depends."  The best way to understand the issue for a particular method, sample size, and data type is to perform a simulation study