1. Statistical Distributions

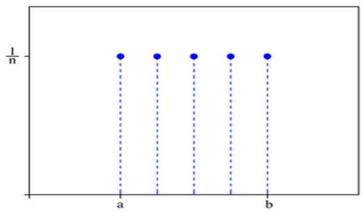
The distributions in the theory of statistics are classified mainly as discrete and continuous distributions.

Discrete Distributions		Continuous Distributions	
Uniform		Uniform	
Bernoulli		Normal & Standard Normal	
Binomial		t-Distribution	
Poisson		Chi-square	
Negative Binomial		F-Distribution	
Geometric		Exponential	
Hyper geometric		Gamma	
		Weibull	

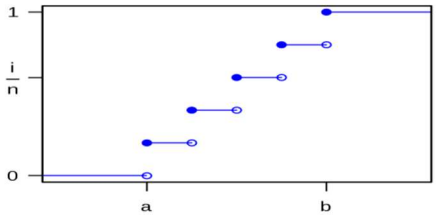
1.1 Discrete Distributions

Uniform Distribution

Probability Mass Function



Cummulative Distribution Function



Bernoulli and Binomial distributions

If an experiment has two possible outcomes, "success" and "failure", and their probabilities are, respectively, $\theta$and $(1-\theta)$, then the variable of number of successes(X), has a Bernoulli distribution with $pmf;\ f(x)=\theta^x(1-\theta)^{1-x}; x=0\ or\ 1.$

The experiment consists of n independent, repeated Bernoulli trials is said to be a binomial experiment. The variable of number of successes(X), then has a binomial distribution with $pmf;\ f(x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}; x=0,1,\dots,n$

It can be proved that the mean and the standard deviation for a large sample from a binomial distribution are given by

$$\bar{x} = np \quad \text{and} \quad \sigma = \sqrt{npq};$$

where X is approximately distributed as Normal (Normal approximation to Binomial)

Poisson Distribution

In a Poisson experiment, the random occurrence of number of events over an interval (usually a time interval) is observed. In the same experiment if the time between two events is observed, the variable will theoretically follow a continuous distribution which will be discussed later.

The probability mass function of the r.v. *X*

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}; x = 0,1,2,\dots$$

- λ – The mean number of counts in the interval (>0)
- E(X) = V(X) = λ

Negative Binomial distribution (Pascal distribution)

A **negative binomial experiment**

- The experiment consists of *x* repeated trials.
- Each trial results only in two outcomes, one 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.
- The experiment continues until *r* successes are observed, where *r* is specified in advance.

Eg: Consider the statistical experiment of flipping a coin repeatedly and count the number of times the coin lands on heads. Continue flipping the coin until it has Head 5 times on top. Then the number of trials needed to have Head turned on 5 times (X), follows a negative binomial distribution.

X ***:***	5	6	7	8	9	10	……
P(X) :	?	?	?	?	?	?	……

Notation and terminology

x : The number of trials required to produce *r* successes in a negative binomial experiment.

r :The number of successes in the negative binomial experiment.

P :The probability of success on an individual trial.

Q : The probability of failure on an individual trial, *1-P*.

$^{n}C_{r}$: The number of combinations of *n* things, taken *r* at a time.

Negative Binomial probability

b*(*x*; *r*, *P*) : - the probability that an *x*-trial negative binomial experiment results in the r^{th} success on the x^{th} trial, when the probability of success of an individual trial is P.

$$b^*(x;\, r,\, P) = {}^{x-1}C_{r-1} \,.P^{r} \,.\, (1 - P)^{x - r}$$

Mean of Negative Binomial distribution

$$\mu_x = \frac{r}{P}$$

Variance of Negative Binomial distribution

$$V_x = \frac{rQ}{P^2}$$

NB: When dealing with negative binomial distribution, check on how the negative binomial random variable is defined.

Alternative definitions can be:

- The negative binomial random variable is *R*, the number of successes before the binomial experiment results in *k* failures. The mean of *R* is $\mu_R = kP/Q$.
- The negative binomial random variable is *K*, the number of failures before the binomial experiment results in *r* successes. The mean of *K* is $\mu_K = rQ/P$.

Geometric Distribution (A special case of Negative Binomial)

This is a special case of the negative binomial distribution, where the variable of interest is the **number of trials required for a single success or the first success**. Thus, the geometric distribution is negative binomial distribution with the number of successes (*r*) is equal to 1.

An example of a geometric distribution would be asking for the probability that the first head occurs on the third flip. That probability is referred to as a **geometric probability** and is denoted by *g*(*x*; *p*). The formula for geometric probability is

$$g(x; p) = p.q^{x-1}$$

Mean of Geometric distribution

$$\mu_x = \frac{1}{P}$$

Variance of Geometric distribution

$$V_x = \frac{q}{P^2}$$

Hyper Geometric distribution

Hypergeometric Experiment

- A sample of size *n* is randomly selected without replacement from a population of *N* items.
- In the population, *k* items can be classified as successes, and *N* - *k* items can be classified as failures.

Eg: Consider the statistical experiment of randomly selecting 2 marbles without replacement from an urn of 10 marbles - 5 red and 5 green. The variable of interest is the number of red marbles selected. This is a hyper geometric experiment.

Note : As binomial experiment requires that the probability of success be constant on every trial, the above is not a binomial experiment. In the above experiment, the probability of a success changes on every trial. Further that if the marbles were selected with replacement, the probability of success would not change. It would be 5/10 on every trial. Then, this would be a binomial experiment.

Notations and terminology

N : The number of items in the population.

k : The number of items in the population that are classified as successes.

n : The number of items in the sample.

x : The number of items in the sample that are classified as successes.

${}^{k}C_{x}$: The number of combinations of *k* things, taken *x* at a time.

Hypergeometric probability

h(*x*; *N*, *n*, *k*): - the probability that an *n*-trial hypergeometric experiment results in exactly*x* successes, when the population consists of *N* items, *k* of which are classified as successes.

$$h(x; N, n, k) = [\ {}^{k}C_{x}\ ]\ [\ {}^{N-k}C_{n-x}\ ] / [\ {}^{N}C_{n}\ ]$$

Mean of the Hypergeometric distribution

$$\mu_x = nk / N$$

Variance of the Hypergeometric distribution

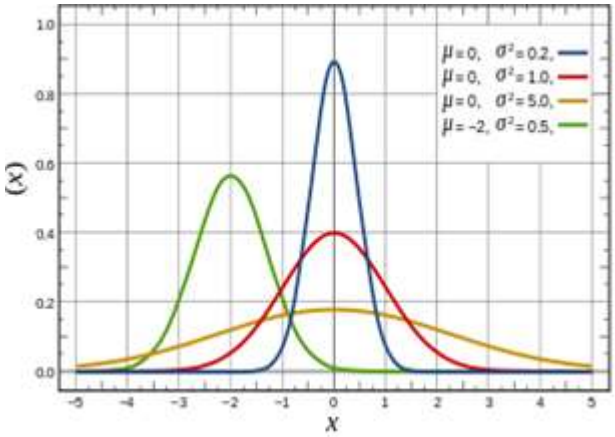
$$V_x = nk\,(N - k)(N - n) / [\ N^2\,(N - 1)\ ]$$

1.2 Continuous Distributions

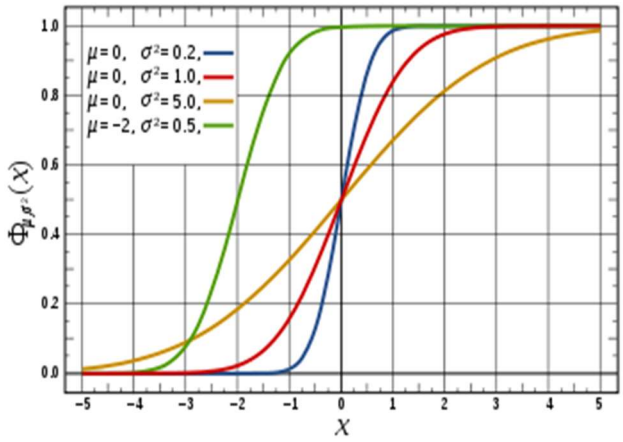
Normal Distribution and Standard Normal Distribution

$$(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$;Normal *pdf*

$$(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$;Standard Normal *pdf*



Cummulative density function



Exponential Distribution

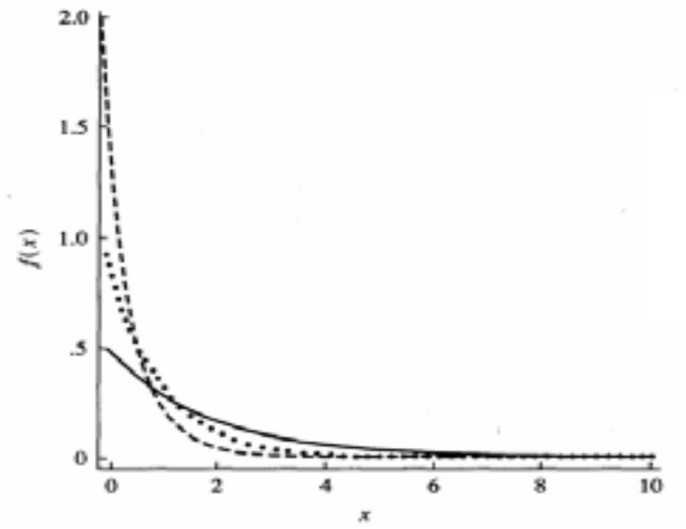
If a random variable X has the *pdf*

$$\boldsymbol{f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0 \; and \; \lambda > 0,}$$

then it is said to have the exponential distribution with parameter λ and written as $X \sim Exp(\lambda)$.

The exponential distribution is often used to model the length of time until an event occurs. The exponential distribution can be thought of as the continuous analogue of the geometric distribution.

This parameter $\boldsymbol{\lambda}$ represents the "**mean number of events per unit time**" e.g. the rate of arrivals or the rate of failures as same as in Poisson distribution.



$\lambda = 0.5(solid)$

$\lambda = 1(dotted)$

$\lambda = 2(dashed)$

Applications

- Model inter arrival times (time between arrivals) when arrivals are completely random; $\boldsymbol{\lambda}$ = arrivals / hour
- Model service times; $\boldsymbol{\lambda}$ = services / minute
- Model the lifetime of a component that fails catastrophically (i.e. light bulb); $\boldsymbol{\lambda} =$ failure rate

Properties of the random variable X which has exponential distribution

1. It is closely related to the Poisson distribution – if X describes the time between two failures then the number of failures per unit time has the Poisson distribution with parameter $\boldsymbol{\lambda}$, the same.

2. The *cdf* is $\boldsymbol{F_X(x) = \lambda \int_0^x e^{-\lambda y}\ dy\ = 1 - e^{-\lambda x}}$

3. The 100(1 −∝)% percentile is $\boldsymbol{x_\propto = -\frac{1}{\lambda} \ln \propto}$

4. Mean $\boldsymbol{\mu_x = 1/\lambda}$

5. Variance $\boldsymbol{V_x = 1/\lambda^2}$

6. Moment Generating Function (*mgf)* $M_X(t) = \boldsymbol{\lambda/(\lambda - t)}$

7. "Memoryless" property
   For all s>= 0 and t >=0

   P(X > s + t | X > s) = P(X > t)

Instance 1: If it is known that a component has survived s hours so far,the remaining amount of time that it survives follows the same distribution as the original distribution. It does not remember that it already has been used for s amount of time.

*Instance 2:*This means that the distribution of the waiting time to the next event remains the same regardless of how long we have already been waiting. This only happens when events occur (or not) totally at random, i.e., independent of past history

Exercise : Suppose the life of an industrial lamp is exponentially distributed with failure rate $\boldsymbol{\lambda}$=1/3 (one failure every 3000 hours on the avg.) Determine the probability that

a) the lamp will last no longer than its mean life time. (constant for any $\boldsymbol{\lambda}$**)**
b) the lamp will last longer than its mean life time
c) the industrial lamp will last between 2000 and 3000 hours.
d) the lamp will last for another 1000 hours given that it is operating after 2500 hours.

Answer:

a) $P(X \leq 3)$=
b) $P(X > 3)$=
c) $P(2 \leq X \leq 3)$=

d) $P(X > 3.5|X > 2.5)$=$P(X > 2.5 + 1|X > 2.5)$=$P(X > 1)$

Theorem :X has an exponential distribution **iff** X is a positive continuous r.v. and P(X>s+t | X>s) = P(X>t) for all s,t> 0.

Proof:*Omitted*

Gamma distribution

Gamma distribution is more suitable to describe some of the real world applications when they follow exponential patterns. The general command of a such probability density is given by

$$f(x) = \begin{cases} kx^{\alpha-1}e^{-x/\beta}; & for\ x > 0 \\ 0 \quad\quad\quad\quad\quad ; & elsewhere \end{cases}$$

where $\alpha > 0$ and $\beta > 0$, and k must be such that the total area under the curve is equal to 1.

In evaluating k, using calculus theory, the ***Gamma function*** which only depends on α is derived:

$$\tau(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy \qquad\qquad for\ \alpha > 0$$

The ***Gamma function*** follows the recursion formula

$$\tau(\alpha) = (\alpha - 1)\,\tau(\alpha - 1);$$

$$\tau(\alpha) = (\alpha - 1)!$$

where $\tau(1) = \int_0^1 y^0 e^{-y}dy = 1 \quad and \quad \tau(1/2) = \sqrt{\pi}$

Thus $\int_0^\infty kx^{\alpha-1}e^{-x/\beta}dx = k\beta^\alpha\tau(\alpha) = 1$

A random variable X has a ***Gamma distribution*** has the probability density function

$$g(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha\tau(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}, & for\ x > 0 \\ 0 \quad , & elsewhere \end{cases}$$

Where $\alpha > 0$ and $\beta > 0$.

- The mean $\mu = \alpha\beta$ and $V(X) = \alpha\beta^2$
- Observe the graphs of gamma functions for different pairs of values for $\alpha$and β

Exercise: In a certain city, the daily consumption of electric power in millions of kilowatt-hours can be treated as a random variable having a Gamma distribution with $\alpha = 3$ and $\beta = 2$.

(i) What is the average consumption of electric power per day by the city?

(ii) If the power plant of this city has a daily capacity of 12 million kilowatt-hours, what is the probability that this power supply will be inadequate on any given day?

Answer:

(i) Average = $\alpha\beta = 3 * 2 = 6$

(ii) $P(\text{daily consumption of electric power } \geq 12) = \int_{12}^{\infty} \frac{1}{2^3\tau(3)} x^{3-1} e^{-\frac{x}{2}} \ dx$

$$= 1 - \int_{0}^{12} \frac{1}{2^3\tau(3)} x^{3-1} e^{-\frac{x}{2}} \ dx$$

Sampling distributions

Let's draw all possible samples of size n from a given population of size N. Then consider computing a statistic; the mean or a proportion or the standard deviation for each sample.

The probability distribution of this statistic is called a **sampling distribution**.

Variability of a Sampling Distribution

The variability of a sampling distribution is measured by its variance (or by its std. deviation).

This variability will depend on;

- N : The number of observations in the population.
- n : The number of observations in the sample.
- The method used to select the samples at random.

Note: If N is much larger than n, then$n/N$is fairly small and the sampling distribution has roughly the same sampling error, irrespective of whether sampling is done with or without replacement.

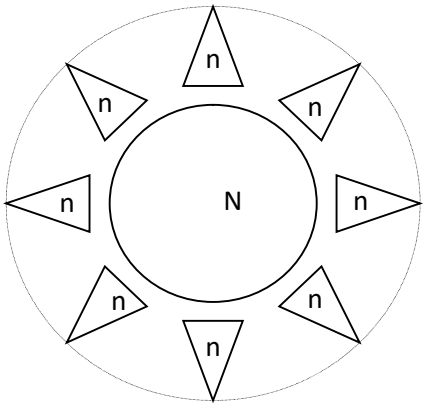
If sampling is done without replacement and the sample represents a significant fraction (say, 1/10) of the population size, the sampling error will be clearly smaller.

The Central Limit Theorem

The **Central Limit Theorem** (CLT) states that the probability distribution of any statistic (or the sampling distribution for any statistic) will be normal or nearly normal, if the sample size is "large enough". Thus the CLT permits approximate calculations for a variety of distributions.
Many statisticians say that a **sample size of 30 is "large enough"** as a rule of thumb.
These are some other instances in which the sample size can be considered as large enough.

- The population distribution is normal.
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

1. Sampling Distribution of the Mean



Let us take $\bar{x}$ as the mean of a sample of size n. Suppose there are$m$ number of such samples drawn from this large population.

If you take the average of the sample means by $\frac{\sum_{i=1}^{m} \bar{x}}{m}$; $\frac{\sum_{i=1}^{m} \bar{x}}{m} = \boldsymbol{\mu_{\bar{x}} = \mu}$ **(popl^n mean)**

And, the standard error of the sampling distribution

$\sqrt{\sigma^2_{\bar{x}}} = \sigma_{\bar{x}} = \sqrt{\sigma^2(1/n - 1/N)} = \sqrt{\sigma^2(1/n)}$ as N→∝

Thus, $\boldsymbol{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$

Therefore, we can specify the sampling distribution of the mean $\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$as $\boldsymbol{\bar{x} \sim N(\mu, \sigma^2/n)}$; whenever two conditions are met:

- The population is normally distributed, or the sample size is sufficiently large.
- The population standard deviation σ is known.

2. Sampling Distribution of the Proportion

Let the probability of getting a success is P; and the probability of a failure is Q in a population. From this population of size N, suppose that we draw all possible samples of size n. And finally, within each sample, suppose that we determine the proportion of successes p and failures q. In this way, we create a sampling distribution of the proportion.

Let us take p as proportion of successes in a sample of size n.

Suppose there are$m$ number of such samples drawn from this large population.

If you take the mean of the sample proportions by

$\frac{\sum_{i=1}^{m} p}{m}$; $\frac{\sum_{i=1}^{m} p}{m} = \boldsymbol{\mu_p = P}$ **(Population proportion of success)**

And, the standard error of the sampling distribution

$\sqrt{\sigma^2{}_p} = \sigma_p = \sqrt{\sigma^2(1/n - 1/N)} = \sqrt{PQ(1/n - 1/N)} = \sqrt{PQ/n}$as N→∝

Thus, $\boldsymbol{\sigma_p = \sqrt{PQ/n}}$

Therefore, we can specify the sampling distribution of the proportion $p \sim N(\mu_p, \sigma^2{}_p)$ as

$\boldsymbol{p \sim N(P, PQ/n)}$; whenever the sample size is sufficiently large and the population probability of success (P) is known.

Example:

1. Suppose that a biased coin has probability p=0.4of heads. In 1000 tosses, what is the probability that the number of heads exceeds 410?

1. Find the probability that of the next 120 births, no more than 40% will be boys. Assume equal probabilities for the births of boys and girls. Assume also that the number of births in the population (N) is very large, essentially infinite.

Exercises:

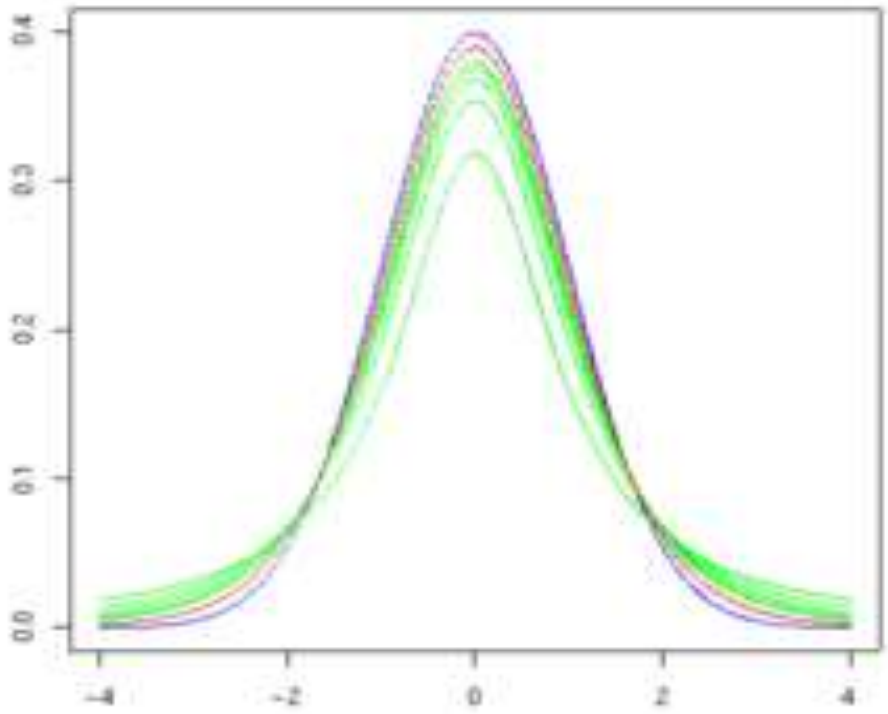
1. A true-false examination has 48 questions. Jane has probability 3/4 of answering a question correctly. Ama just guesses on each question. A passing score is 30 or more correct answers. Compare the probability that Jane passes the exam with the probability that Ama passes it. Jane's score has distribution B(48,0.75), so the probability that Jane's score is 30 or more is 1-P(X<=29) = 0.9627. In case your calculator doesn't give an answer, you will have to use a normal approximation to the Binomial distribution (based on the Central Limit Theorem)

2. A restaurant feeds 400 customers per day. On the average 20 percent of the customers order apple pie.
   (a) Give a range for the number of pieces of apple pie ordered on a given day such that you can be 95 percent sure that the actual number will fall in this range.
   (b) How many customers must the restaurant have, on the average, to be at least 95 percent sure that the number of customers ordering pie on that day falls in the 19 to 21 percent range?

3. A rookie is brought to a baseball club on the assumption that he will have a 0.3 batting average. (Batting average is the ratio of the number of hits to the number of times at bat.) In the first year, he comes to bat 300 times and his batting average is 0.267. Assume that his at bats can be considered Bernoulli trials with probability 0.3 for success. Could such a low average be considered just bad luck or should he be sent back to the minor leagues?

3) Student's t Distribution

A 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.



In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (which, in sample variance, is one, since the sample mean is the only intermediate step).

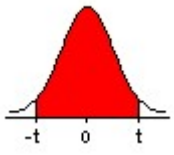
Lane, David M.."Degrees of Freedom".*HyperStatOnline*.StatisticsSolutions.http://davidmlane.com/hyperstat/A42408.html. Retrieved 2008-08-21.

Suppose we have a simple random sample of size n drawn from a Normal population with mean μ and standard deviation σ. Let $\bar{x}$denote the sample mean and s, the sample standard deviation.

Then the quantity $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a *t* distribution with *n*-1 degrees of freedom.

The t score 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 $\bar{x}$, given a random sample of size *n*.

The notation t_α represents the t-score that has a cumulative probability of (1 - α).



Example: $t_{0.05}$ = 2.92, then $t_{0.95}$ = -2.92 for *df*=3

Properties of the t Distribution

- The **mean** of the distribution is equal to **0** .
- The **variance** is equal to ***v* / (*v* - 2)**, where *v* is the degrees of freedom 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). i.e. when the population size is large but the sample sizes are small and the standard deviation of the population is unknown t-Distribution can be applied.

Example: $\boldsymbol{t = (p - P)/\sqrt{(PQ\ /\ n)}}$ **has a *t* distribution with *n*-1 degrees of freedom**

When not to use the t-distribution

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

Example:

1. A random sample of 12 observations from a normal population with mean 48 produced the following
   Estimates: $\bar{x} = 47.1$and $s^2 = 4.7$. Find the probability of getting a sample of the same size with its mean less than or equal to the population mean.

2. The MD of Orrange light bulb manufactures claims that an average of their light bulbs lasts 300 days. An investigator randomly selects 15 bulbs for testing and those bulbs last an average of 290 days, with a standard deviation of 50 days. Assuming MD's claim as true, determine the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

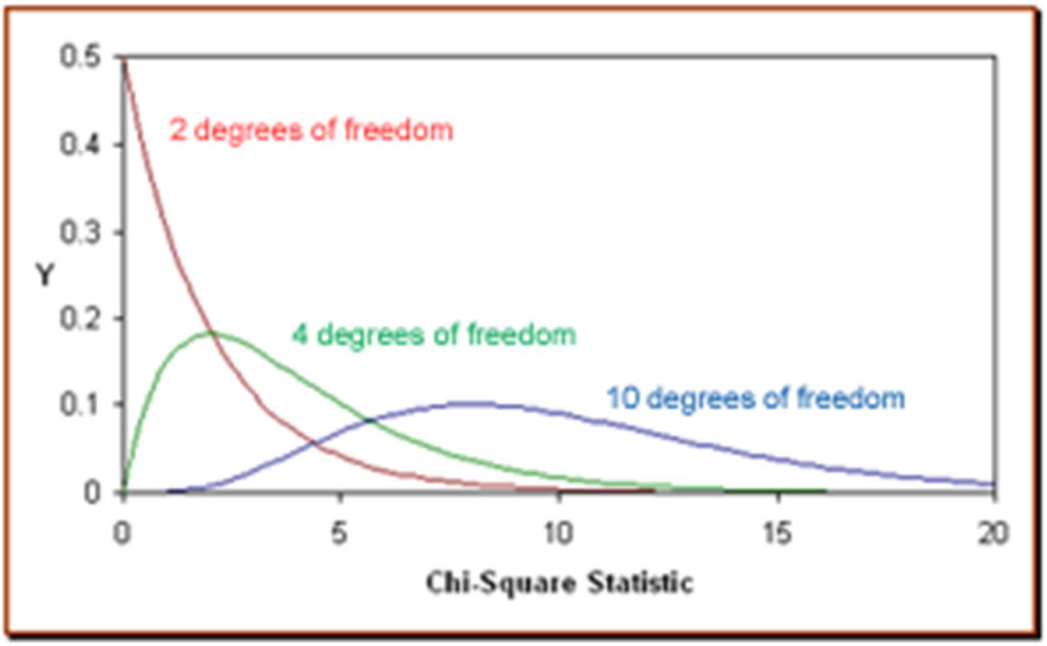
4. Chi-square Distribution

The chi-square statistic can be calculated from a sample of size $n$drawn from a population, which is normal, using the following equation:

$$\chi^2 = {(n-1)s^2}/{\sigma^2}$$

When sampling is done for an infinite number of times, and by calculating the chi-square statistic for each sample, the sampling distribution for the chi-square statistic can be obtained. It is then called the chi-square distribution.

The **chi-square distribution** also depends on the degrees of freedom; $(n-1)$.



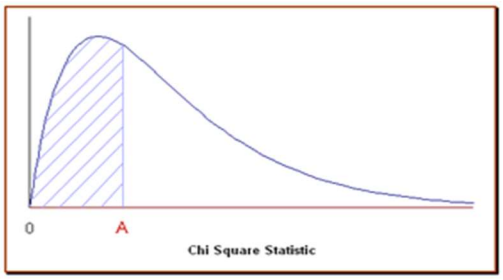
Properties of the chi-square distribution:

- The mean of the distribution is equal to the number of degrees of freedom: $\boldsymbol{\mu = \nu}$.
- The variance is equal to two times the number of degrees of freedom: $\boldsymbol{\sigma^2 = 2\nu}$
- When the degrees of freedom are greater than or equal to 2, the maximum value for $f(x), the\ pdf\ of\ chi-square$ occurs.

- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

Cumulative Probability of the Chi-Square Distribution

The chi-square distribution is constructed so that the total area under the curve is equal to 1. The probability that the value of a chi-square statistic will fall between 0 and A; $P(\chi^2_{n-1} \leq A)$ is illustrated by the following diagram.



Using the following Chi-Square Distribution table, one can find thecriticalχ^2value, when the probability of exceeding the critical value is given.

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Example: My Cell company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 5 minutes. Suppose the manufacturing department runs a quality control test. They randomly select 10 batteries. The standard deviation of the selected batteries is 6 minutes.

a) What is the chi-square statistic which represents this test?

b) What is the probability that the standard deviation of any sample of size 10 would be greater than 6 minutes?

5. 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 ***f* statistic**, also known as an ***f* value**, is a random variable that has an F distribution.

How to compute an ***f* statistic**:

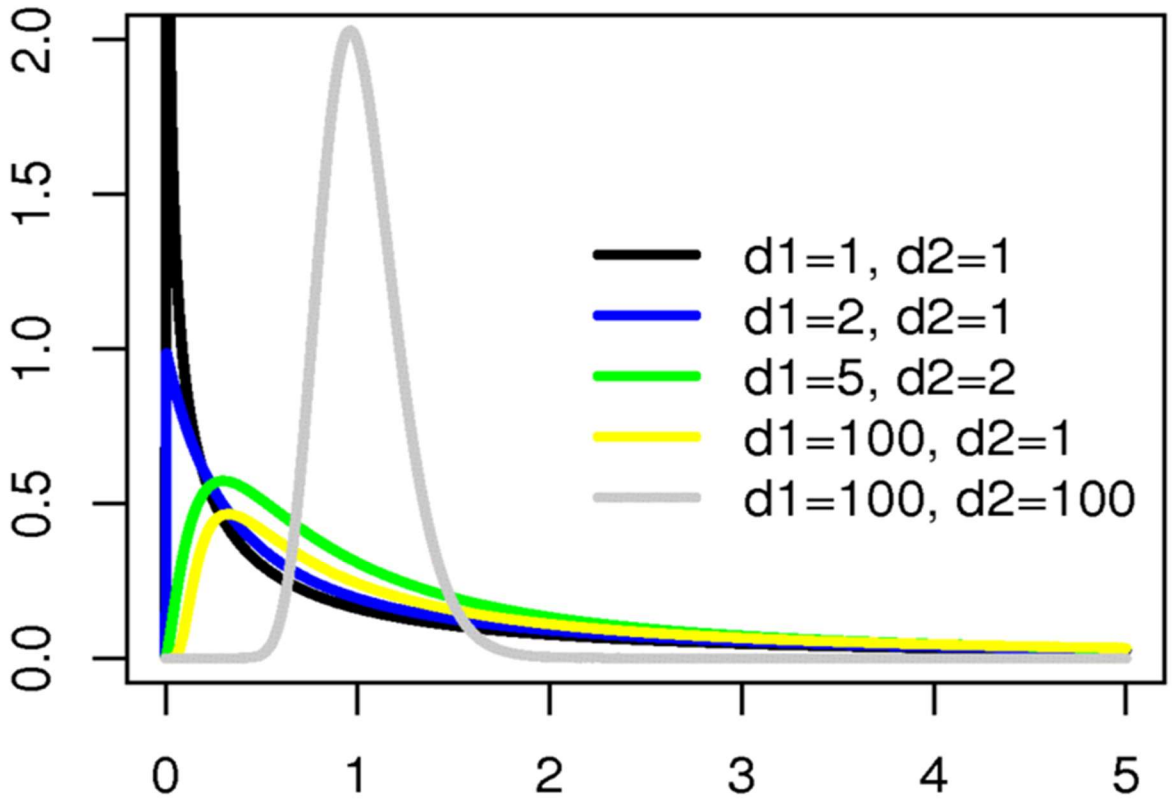
- Select a random sample of size n_1 from a normal population, having a standard deviation equal to σ_1.
- Select an independent random sample of size n_2 from a normal population, having a standard deviation equal to σ_2.
- The *f* statistic is the ratio of s_1^2/σ_1^2 and s_2^2/σ_2^2.

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

$$f(v_1, v_2) = [\, s_1^2/\sigma_1^2 \,] / [\, s_2^2/\sigma_2^2 \,]$$
$$f(v_1, v_2) = [\, s_1^2 . \sigma_2^2 \,] / [\, s_2^2 . \sigma_1^2 \,]$$

$$f(v_1, v_2) = [\, \chi^2_1 / v_1 \,] / [\, \chi^2_2 / v_2 \,]$$
$$f(v_1, v_2) = [\, \chi^2_1 . v_2 \,] / [\, \chi^2_2 . v_1 \,]$$

The curve of the F distribution depends on the degrees of freedom, v_1 and v_2.



Properties of the F distribution:

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

Cumulative Probability of the F Distribution

This cumulative probability represents the likelihood that the *f* statistic is less than or equal to a specified value.

F-distribution table can be used to find the value of an *f* statistic having a cumulative probability of (1 - α); represented by f_{α}.

Thus, $f_{0.05}(v_1, v_2)$ refers to value of the f statistic having a cumulative probability of (1-0.05)= 0.95, with $v_1$and $v_2$degrees of freedom.

Example:

Suppose a sample of 11 of cows was selected at random from a population of them having the population standard deviation of their weight is 5 kg and the estimated sample sd is 4.5 kg. Another sample of size 7 of bulls was taken in a similar way with their population sd is 3.5 kg and sample sd is 4 kg.

a) Compute an f-statistic.

b) Determine the associated cumulative probability by finding an approximate f-value to the above answer from the f-tables available for different significance levels (α).

c) Interpret the probability you found.

Reference for f-table: http://www.socr.ucla.edu/Applets.dir/F_Table.html

Upgrade your knowledge by:

- Finding the *pdf* 's of the above sampling distributions.
- Studying the patterns of *cdf* 's of the above sampling distributions.

Confidence Intervals

A confidence interval will give you a range of values for a given population parameter, within which the parameter falls in 100(1-α)% of the time.

In general a Confidence Interval (CI) is

$$\boldsymbol{Statistic \pm margine\ of\ error\ at\ 100(1-\alpha)\%\ confidence\ level}$$

Confidence interval for mean and variance

The following table includes the standard errors of some statistics to help you in finding the confidence intervals for the respective population parameters.

Statistic	Standard Error (SE)
Sample mean, $\bar{x}$	s / √ n
Sample proportion, p	√ [p(1 - p) / n]
Difference between means, $\bar{x}_1 - \bar{x}_2$	√ [s^2_1 / n_1 + s^2_2 / n_2]
Difference between proportions, p_1 - p_2	√ [$p_1(1-p_1)$ / n_1 + $p_2(1-p_2)$ / n_2]

$100(1-\alpha)$% confidence interval for population mean (μ)

$$\bar{x} \pm Z\alpha_{/2}(SE)$$

Margin of Error

$100(1-\alpha)$% confidence interval for population proportion of success (*P*)

$$\hat{p} \pm Z\alpha_{/2}(SE)$$

$100(1-\alpha)$% confidence interval for population variance(σ^2)

You know that;

$$Chi-square\ statistic\ = {(n-1)s^2}/{\sigma^2}\ and\ {(n-1)s^2}/{\sigma^2} \sim \chi 2_{(n-1)}$$

Thus, $\left[\chi 2_{\alpha/2,(n-1)} \leq {(n-1)s^2}/{\sigma^2} \leq \chi 2_{1-\alpha/2,(n-1)}\right]$

$= \left[\chi 2_{\frac{\alpha}{2},(n-1)}/(n-1)s^2 \leq {1}/{\sigma^2} \leq \chi 2_{1-\frac{\alpha}{2},(n-1)}/(n-1)s^2\right]$

$= \left[(n-1)s^2/\chi 2_{1-\frac{\alpha}{2},(n-1)} \leq \sigma^2 \leq (n-1)s^2/\chi 2_{\frac{\alpha}{2},(n-1)}\right]$

Computational Exercise

Breakdown voltage is a characteristic of an insulator that defines the maximum voltage difference that can be applied across the material before the insulator collapses and conducts. In solid insulating materials, this usually creates a weakened path within the material by creating permanent molecular or physical changes by the sudden current. Within rarefied gases found in certain types of lamps, **breakdown voltage** is also sometimes called the "striking voltage". [Wikipedia]

The breakdown voltage of a material is observed on 17 experimental units as it is not a definite value because it is a form of failure. Thus we have *n* = 17 and s^2= 137324:3. Find the 95% confidence interval for σ^2, to describe more about the population variance (variance of the breakdown voltage of the material)?

Hypothesis Testing

A **statistical hypothesis** is an intelligent educated guess/assumption about a population parameter, which may or may not be true. There are two forms of statistical hypotheses.

- **Null hypothesis**: This is denoted by H_0, is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis:** This is denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

The process of hypothesis testing needs

- ✓ A hypothesis on a population parameter
- ✓ A test statistic under the null hypothesis
- ✓ The p value of the test statistic
- ✓ Decision rule based on a significance level

Decision Errors

	Reject H_0	**Do not reject H_0**
H_0True	**Type I error (α)**	
H_0 False		**Type II error (β)**

- **Significance level =α=P(Type I error)**
- **Power of the test = 1- β**

Rejection Criterion for null hypothesis

- P-value: The strength of evidence in support of a null hypothesis is measured by the **P-value**.

- Region of rejection in One-Tailed and Two-Tailed Tests

Mean test (One sample)

Null Hypothesis	Alternative Hypothesis	Test Statistic, the type of test & rejection criterion	
$H_0: \mu = A$ Normal Population has mean "A"	$H_1: \mu \neq A$	$Z = (x - \mu)/\sigma$; σ^2 known; testing on a single sample point $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$; testing on a single sample mean	Two tail test
		$T = (X - \mu)/s$ or $T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$; σ^2 unknown	
$H_0: \mu \leq A$ $H_0: \mu \geq A$	$H_1: \mu > A$ $H_1: \mu < A$		One tail test

Examples:

1. Bon Air Elementary School has 300 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01.
2. An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. Suppose a simple random sample of 50 engines is tested. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

Mean Test (Two Sample)

<table>
<tr><th>Hypotheses</th><th colspan="2">Test Statistic, the type of test & rejection criterion</th></tr>
<tr><td rowspan="3">$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$</td><td>When the two population variances are known and not equal

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When the two population variances are known and equal</td><td rowspan="3">Two tail test</td></tr>
<tr><td>When the two population variances are unknown but equal</td></tr>
<tr><td>When the two population variances are unknown and not equal

df = $(s_1^2/n_1 + s_2^2/n_2)^2$ /

{ [$(s_1^2 / n_1)^2 / (n_1 - 1)$] + [$(s_2^2 / n_2)^2 / (n_2 - 1)$] }

or smaller of $n_1 - 1$ & $n_2 - 1$</td></tr>
</table>

Proportion Tests

Null Hypothesis	Alternative Hypothesis	Test Statistic, the type of test & rejection criterion	
			Two tail test
			One tail test
			One tail test

Paired t-test

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$$

$\bar{d}$ = *sample mean difference*

μ_{d_0} = *hypothesized population mean difference*

$s_{\bar{d}} = s_d / \sqrt{n}$

n = *number of sample differences*

s_d = *standard deviation of sample differences*

Example:

The weights of 9 obese women before and after 12 weeks on a very low calorie diet were as follows:

Before	After	Difference
117.3	83.3	-34.0
111.4	85.9	-25.5
98.6	75.8	-22.8
104.3	82.9	-21.4
105.4	82.3	-23.1
100.4	77.7	-22.7
81.7	62.7	-19.0
89.5	69.0	-20.5
78.2	63.9	-14.3

Test whether the expected weight loss is at least 20kg for obese women after the treatment of this low-calorie diet for 12 weeks. Use 5% significance level.

Variance Tests

Population μ	Estimation of σ^2	Test Statistic & Distribution
μ Known	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_n$
	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_n$
μ Unknown	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$
	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

Acknowledged **http://www.xycoon.com/**

Example:

1. The data 159.9, 187.2, 180.1, 158.1, 225.5, 163.7, and 217.3 consists of the weights, in pounds, of a random sample of seven individuals taken from a population that is normally distributed. The variance of this sample is given as 753.04.

   Test the null hypothesis H_0: $\sigma^2 = 750.0$ against the alternative hypothesis H_1: $\sigma^2 \neq 750.0$ at a level of significance of 0.3.

2. Students have collected the data 27, 29, 22, 21, 26, 28, 24, and 29 from one population and the data 19, 18, 24, 18, 22, and 15 from another. The variance of the first sample is 9.64286 and the variance of the second sample is 10.2667. The ratio of the first variance to the second is 0.939239. Test the null hypothesis that the two variances are equal, H_0: $\sigma_1^2 / \sigma_2^2 = 1$, against the alternative hypothesis that the two variances are not equal, H_1: $\sigma_1^2 / \sigma_2^2 \neq 1$, where σ_1^2 is the variance of the first population and σ_2^2 is the variance of the second population, at a significance level of .10.

Exercises on hypothesis testing:

1. An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was $1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculated a sample mean of $1,950. Assuming that the standard deviation of claims is $500, and set α= 0.05, test to see if the insurance company should be concerned.

2. Trying to encourage people to stop driving to campus, the university claims that on average it takes people 30 minutes to find a parking space on campus. I do not think it takes so long to find a spot. In fact I have a sample of the last five times I drove to campus, and I calculated $\bar{x}$= 20. Assuming that the timeit takes to find a parking spot is normal, and that s^2= 6 minutes, then perform a hypothesis test with levelα= 0.10 to see if my claim is correct.

3. A sample of 40 sales receipts from a grocery store has $\bar{x}$= $137 and s^2= $30.2. Use these values to test whether or not the mean value of a receipt at the grocery store is different from $150.

4. The actual proportion of families in a certain city who own, rather than rent their home is 0.70. If 84 families in this city are interviewed at random and their response to the question of whether they own their home, are recorded. 61 of them have responded saying that they own the home. Using a suitable test statistic test the claim that the population proportion of owning a home is 0.7.

Chi-square tests

1. Chi-square test of Association

2× 2 Contingency Table

The outcome of a certain IQ test is tabularized as follows.

	Pass		Fail		Total
Male	**O=28**	**E=**	**O=12**	**E=**	**40**
Female	**O=34**	**E=**	**O=26**	**E=**	**60**
Total	**62**		**38**		

The two variables here are "Gender of Candidate" and "Results of the Candidate".

When you have two categorical variables from the same population, you may test whether there is a significant association between the two variables using the Chi-square Test.

Hypothesis

H_0 : *There is no relationship between gender and results*

H_1 : *There is a significant relationship between gender and results*

Test Statistic

$$\chi^2_{cal} = \sum \frac{(O_i - E_i)^2}{E_i} \quad \sim \quad \chi^2_{df,\ \alpha\%} \quad ; df = (r-1)(c-1)$$

Under H_0

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

$\chi^2_{cal} =$

$\chi^2{}_{1,\ 5\%} = 3.84$

Decision

$h \times k$ Contingency Table

Example: A survey of 200 families, known to the regular television viewers was undertaken. They were asked which of the TV channels they watched most during a common week, and the observations are as follows.

TV channel watched most	Region			
	North	East	South	West
1	19	16	42	23
2	6	11	26	7
3	15	3	12	10

Test the hypothesis that there is no association between the TV channel watched most and the Region, using the Chi-square test.

2. Chi-square Goodness of Fit test

Example 1

From a list of 500 digits the occurrence of each distinct digit is observed. Test at 5% significance level, whether the sequence is a random sample from **the Uniform distribution.**

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	40	58	49	53	38	56	61	53	60	32

Example 2

The table below gives the number of heavy rainstorms reported by 330 weather stations over a one year period.

a) Find the expected frequencies of rainstorms given by the Poisson distribution having the same mean and the total as the observed distribution.

b) Use the Chi-square test to check the adequacy of the **Poisson distribution** to model these data.

# rainstorms	0	1	2	3	4	5	More than 5
# weather stations	102	114	74	28	10	2	0

Moments and Moment Generating Function

Moments

In Statistics, the **mathematical expectation** is called the **moments** of the distribution of a random variable.

Definition(r[th] moment of ar.v.)

The r[th] moment of a random variable X denoted by $\boldsymbol{\mu_r'}$ is the expected value of the random variable's r[th] power; i.e.$\boldsymbol{E(X^r)}$.

$$For\ r\ = 1,2,3,\ldots$$

$$\mu_r' = E(X^r) = \sum x^r P(x)\ ;\ \text{when } X \text{ is discrete}$$

$$\mu_r' = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x)dx\ ;\ \text{when } X \text{ is continuous}$$

***Special Application*: $\boldsymbol{\mu_1' = E(X^1) = E(X) = \mu}$ (mean of r.v. X)**

Definition (r[th] moment about the mean of ar.v.)

The r[th] moment about the mean of a random variable X denoted by $\boldsymbol{\mu_r}$ is the expected value of $(X-\mu)^r$; i.e.$\boldsymbol{E[(X-\mu)^r]}$.

$$For\ r\ = 1,2,3,\ldots$$

$$\mu_r\ = E[(X-\mu)^r] = \sum (x-\mu)^r P(x)\ ;\ \text{when } X \text{ is discrete}$$

$$\mu_r\ = E[(X-\mu)^r] = \int_{-\infty}^{+\infty} (x-\mu)^r f(x)dx\ ;\ \text{when } X \text{ is continuous}$$

***Special Application*: $\boldsymbol{\mu_2 = E[(X-\mu)^2] = V(X) = \sigma^2}$ (variance of r.v. X)**

Theorem:

$\boldsymbol{\sigma^2 = \mu_2' - \mu^2}$; i.e. $\boldsymbol{V(X) = E(X^2) - [E(X)]^2}$

Proof:

Moment Generating Function (MGF)

- The moments of most distributions can be determined directly by evaluating the respective integrals or sums.
- MGF is an alternative procedure, which sometimes provides considerable simplifications to find the moments.
- MGF can be used to find the expected value of ar.v. and its variance.

Definition

$M_X(t)$is the value which the function M_X assumes for the real variable (t).

The MGF of ar.v. X is given by;

$$M_X(t) = E(e^{tx}) = \sum e^{tx}P(x); \quad \text{when} X \text{ is discrete}$$

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{+\infty} e^{tx}f(x)dx; \quad \text{when} X \text{ is continuous}$$

$$where, M_X(t) = E(e^{tx}) = E[1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^r}{r!} + \cdots$$

$$M_X(t) = E(e^{tx}) = 1 + tE(x) + \frac{t^2E(x^2)}{2!} + \frac{t^3E(x^3)}{3!} + \cdots + \frac{t^rE(x^r)}{r!} + \cdots$$

$$and, \quad M'_X(t) = \frac{d\, M_X(t)}{dt}$$

$$= \frac{d\, E(e^{tx})}{dt} = 0 + E(x) + \frac{2t \quad E(x^2)}{2!} + \frac{3t^2E(x^3)}{3!} + \cdots + \frac{rt^{r-1}E(x^r)}{r!} + \cdots$$

$$M''_X(t) = \frac{d^2\, M_X(t)}{dt^2}$$

$$= \frac{d^2\, E(e^{tx})}{dt^2} = 0 + 0 + \frac{2E(x^2)}{2!} + \frac{6t \quad E(x^3)}{3!} + \cdots + \frac{r(r-1\_t^{r-2}E(x^r)}{r!} + \cdots$$

Properties of the MGF

1. $M'_X(t)|_{t=0} = E(X) = \mu$
2. $M''_X(t)|_{t=0} = E(X^2)$
3. $V(X) = M''_X(t)|_{t=0} - (M'_X(t)|_{t=0})^2$

Example 1

Find the MGF and E(X) and V(X) for the r.v.X whose *pdf* is given by

$$f(x) = \begin{cases} e^{-x}, & for\ x > 0 \\ 0, & otherwise \end{cases}$$

Example 2

Suppose Y~ Bin (n,p). Find E(X) and V(X) using its MGF.

Exercises

1. Let Y be a continuous r.v with *pdf*f(y)=$2e^{-3y}$; y≥ 0, Find the mean and the variance of Y.
2. Given that the probability distribution of a r.v. is 1/8 ${}^{3}_{r}C$ for r=1,2,3. Find the MGF, mean and variance for this random variable.

Linear Regression

The simplest way to predict values of a random variable in Statistics can be considered as Linear Regression technique.

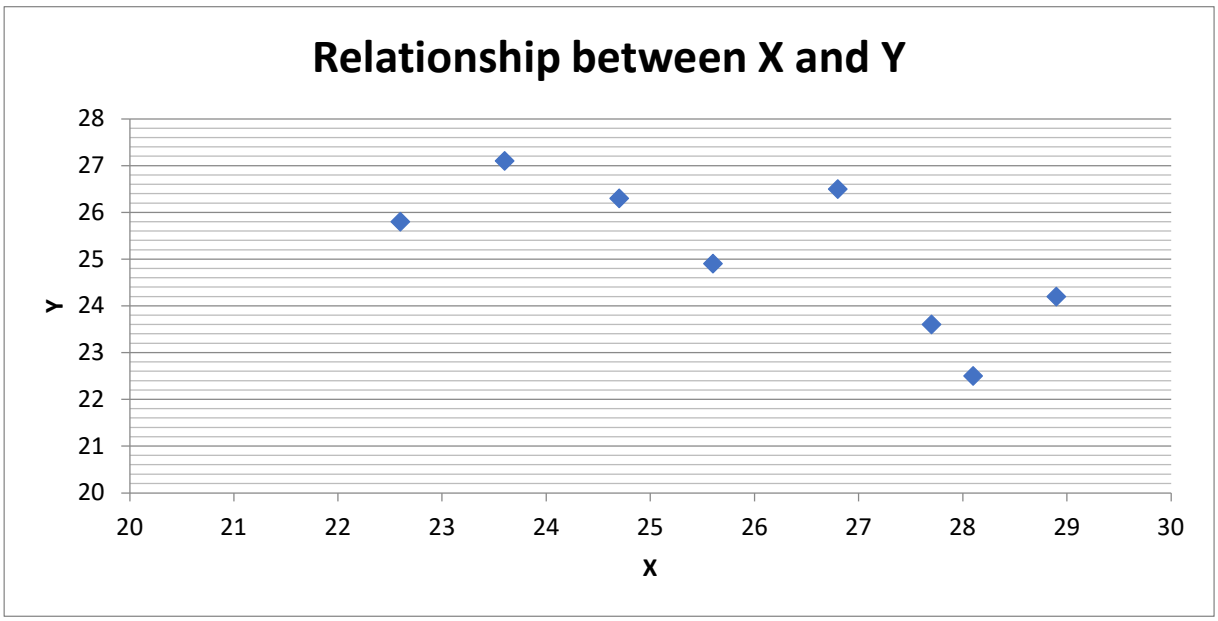
Simple Linear Regression

Mainly the relationship between two continuous variables, which can be measured simultaneously on an experimental unit in an experiment is considered, where one variable will be taken as the dependent variable (y) and the other is said to be the independent or the explanatory variable (x).

Eg: Temperature and Pressure

Indications

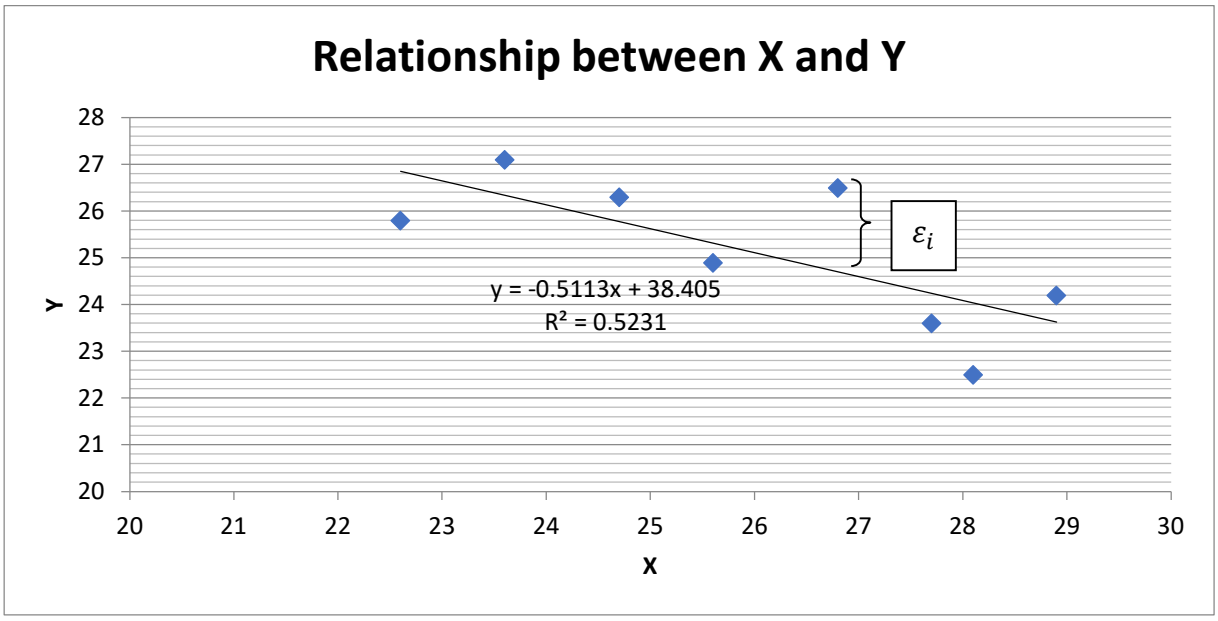
1. Scatter Diagram
2. Correlation Coefficient



Simple Linear Regression

In simple linear regression, we allow only one independent variable to predict the dependent variable. Under multiple linear regression there can be many independent variables predicting a sing dependant variable.

In here, a set of measurements(x_i, y_i); i=1,2,3…*n*on *n* individuals are taken and if evidence is available on a scatter diagram for a linear relationship between x and y, a regression function will be established to model the relationship.



The **coefficient of determination** (denoted by $\mathbf{R^2}$) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

General Model

$$\boldsymbol{Y = \alpha + \beta X}$$

For an existing individual

$$y_i = \alpha + \beta x_i + \varepsilon_i;$$

$$\varepsilon_i = the\ observation's\ deviation\ from\ the\ model\ Y = \alpha + \beta X$$

OR

$\varepsilon_i = error\ in\ prediction$

Fitting a Linear Regression Function

First Step : Plot the scatters and look for any evidence for a linear relationship.

Assumption:

Error terms are independently and identically distributed as Normal with mean zero and varianceσ^2.

i.e.$\varepsilon \sim N(0, \sigma^2)$

Parameter Estimation (estimating $\boldsymbol{\alpha}$ &$\boldsymbol{\beta}$):

Estimation will be carried out based on the principle of **"least squares".**

In least square estimation, the "sum of squares of errors" will be minimized.

i.e we will find α, β such that $\sum \varepsilon_i^2$ is at minimum.

Let *ESS* = Error Sum of Squares

$$\boldsymbol{ESS} = \sum_{i=1}^{n} \varepsilon_i^2 \quad = \sum_{1}^{n} (y_i - \alpha - \beta x_i)^2$$

ESS is at minimum when $\frac{\partial(ESS)}{\partial \alpha} = 0$ and $\frac{\partial(ESS)}{\partial \beta} = 0$

Computing $\hat{\beta}$; the least square estimate of β (regression coefficient)

$$\frac{\partial(ESS)}{\partial \beta} = \frac{\partial(\sum_1^n (y_i - \alpha - \beta x_i)^2)}{\partial \beta} = 2 \sum_{1}^{n} x_i (y_i - \alpha - \beta x_i)^1 = 0$$

$$\sum_{1}^{n} x_i y_i - \alpha \sum_{1}^{n} x_i - \beta \sum_{1}^{n} x_i^2 = 0$$

$$\sum_{1}^{n} x_i y_i = \alpha \sum_{1}^{n} x_i + \beta \sum_{1}^{n} x_i^2 \; ---------(1)$$

Let $\hat{\alpha}$ be the least square estimate of α and

$$\frac{\partial(ESS)}{\partial\alpha} = \frac{\partial(\sum_1^n (y_i - \alpha - \beta x_i)^2)}{\partial\alpha} = 2\sum_1^n (y_i - \alpha - \beta x_i) \quad = 0$$

$$\frac{\partial(ESS)}{\partial\alpha} = \sum_1^n y_i - n\alpha - \beta\sum_1^n x_i = 0$$

$$\sum_1^n y_i = n\alpha + \beta\sum_1^n x_i \text{ - - - - - - - - -(2)}$$

Equation (1) and (2) are called the Normal Equations.

From (2)

$$\hat{\alpha} = \frac{\sum_1^n y_i}{n} - \hat{\beta}\frac{\sum_1^n x_i}{n}$$

$$\boldsymbol{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

Substituting for $\hat{\alpha}$ in equation (1)

$$\sum_1^n x_i y_i = \left(\frac{\sum_1^n y_i}{n} - \hat{\beta}\frac{\sum_1^n x_i}{n}\right)\sum_1^n x_i + \hat{\beta}\sum_1^n x_i^2$$

$$\sum_1^n x_i y_i = \frac{\sum_1^n y_i \sum_1^n x_i}{n} - \hat{\beta}\frac{(\sum_1^n x_i)^2}{n} + \hat{\beta}\sum_1^n x_i^2$$

$$\hat{\beta} = \frac{n\sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{n\sum_1^n x_i^2 - (\sum_1^n x_i)^2}$$

$$\boldsymbol{\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad and \qquad \hat{\beta} = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2}$$

Example:

X	y
26.8	26.5
28.9	24.2
23.6	27.1
28.1	22.5
22.6	25.8
27.7	23.6
24.7	26.3
25.6	24.9

Confidence Interval for $\boldsymbol{\beta}$

It can be proved that under the assumption of $\varepsilon \sim N(0, \sigma^2)$; $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$.

When $\sigma^2 is\ unknown$ it can be estimated by $\widehat{\sigma^2} = \frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}$.

$$P\left(t_{\frac{\alpha}{2}, n-2} \leq \frac{\hat{\beta} - \beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \leq t_{1-\frac{\alpha}{2}, n-2} \right) = (1 - \alpha)$$

Hypothesis Testing on Regression Coefficient

$H_0 : \beta = 0$

$H_1 : \beta \neq 0; regression\ coefficient\ is\ significantly\ different\ from\ 0$

Test Statistic and its distribution

$$\frac{\widehat{\beta}-\beta}{\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

Analysis of Variance (ANOVA)

H_0 : *Regression Line does not fit the data well*

H_1 : *Regression Line fits the data well*

Source of variation	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**Mean Sum of Squares(MS)**	**F-ratio**	**p-value (prob.>F)**
REGRESSION (estimation via reg. line)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	**1**	**RSS/1**	$F_{cal} = \frac{\text{RSS}/1}{\text{ESS}/(n-2)} \sim F_{1,n-2}$	$P(F_{1,n-2} \geq F_{cal})$
ERROR(Residual) (error in estimation)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	**n-2**	**ESS/(n-2)**		
TOTAL (estimation + error)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	**n-1**			

Example :

Description: These data are on the production of power from wind mills. Direct Current (DC) output was measured against wind speed (in miles per hour).

Number of observations: 25

Variable	**Description**
output	Current output produced by the wind mill
speed	Windspeed (in miles per hour)

Source: Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) Lack-of-fit testing when replicates are not available, *American Statistician*, 43, pp. 135-143.

(speed,output)$\equiv (x, y)$

0.123,2.45	1.582,5.00	2.166,8.15
0.500,2.70	1.501,5.45	2.112,8.80
0.653,2.90	1.737,5.80	2.303,9.10
0.558,3.05	1.822,6.00	2.294,9.55
1.057,3.40	1.866,6.20	2.386,9.70
1.137,3.60	1.930,6.35	2.236,10.00
1.144,3.95	1.800,7.00	2.310,10.20
1.194,4.10	2.088,7.40	
1.562,4.60	2.179,7.85	

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134.282	1	134.282	160.257	.000[b]
	Residual	19.272	23	.838		
	Total	153.554	24			

a. Dependent Variable: Electricity Production at the Wind Mill

b. Predictors: (Constant), Wind Speed (in miles per hour)