Statistical measures & Distributions

Contents:

1. STATISTICAL MEASURES
   - (a) Mean
   - (b) Median
   - (c) Mode
   - (d) Range
   - (e) Standard Deviation
   - (f) Variance
2. DISTRIBUTIONS
   - (a) Binomial
   - (b) Poisson
   - (c) Normal
3. EXAMPLES

- The type of measures of MEAN, MEDIAN and MODE is called an *average* or *measure of location*.
- The type of measures of STANDARD DEVIATION and VARIANCE is called *measure of dispersion*.

The Mean

The ARITHMETIC MEAN (or just MEAN) of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is denoted by $\bar{x}$ and is defined by:

$$\boxed{\bar{x} = \frac{1}{n}(x_1 + x_2 \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

Consider a discrete frequency distribution taking values $\{x_1, x_2, \ldots, x_n\}$ with corresponding frequencies $\{f_1, f_2, \ldots, f_n\}$. The mean $\bar{x}$ is given by:

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}.}$$

<u>Example</u>

- Find the mean of the set $\{-3, -1, 0, 2, 3, 4\}$.
  $\bar{x} = (-3 - 1 + 0 + 2 + 3 + 4)/6 = 0.83$.
- Find the mean of the following frequency distribution:

x_i	-3	-2	-1	0	1	2	3
f_i	6	5	4	3	2	1	1
$f_i x_i$	-18	-10	-4	0	2	2	3

$\sum f_i = 22, \quad \sum f_i x_i = -25, \quad \bar{x} = -25/22 = -1.14.$

Let the set $\{x_1, x_2, \ldots, x_n\}$ be transformed to $\{X_1, X_2, \ldots, X_n\}$, where $X_i = (x_i - a)/b$. The $\bar{x}$ can be obtained by

$$\boxed{\bar{x} = a + b\bar{X}}$$

Proof

From $X_i = (x_i - a)/b$ it follows that $x_i = bX_i + a$. Now,

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum x_i = \frac{1}{n}\sum(a + bX_i) \\ &= \frac{1}{n}\sum a + \frac{b}{n}\sum X_i \\ &= a + b\bar{X}. \end{aligned} \qquad \square$$

Example

Find the mean of the set $\{2678, 4678, 8678, 5678, 6678\}$.

Let $X_i = (x_i - a)/b$, where $a = 2678$ and $b = 1000$. This gives the new set $X = \{0, 2, 6, 3, 4\}$ with mean $\bar{X} = 3$. Thus,

$$\bar{x} = 2678 + 1000 \times 3 = 5678$$

The Median

The mean has the disadvantage of taking extreme values into account, especially for a small set of numbers.

The MEDIAN of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is defined as the middled value of the set when arranged in size order. If the set has an even number of items, then the median is taken as the mean of the two middle two.

Example

1. The wages arranged in size order are: $\{28, 29, 32, 35, 36, 38, 41, 103\}$.

   The Mean is $\bar{x} = 41.89$ and the MEDIAN $x^* = 35.5$.

2. Find the median of the set: $\{65, 68, 68, 66, 64, 65, 65, 67\}$.

   Arranging the set in order: $\{64, 65, 65, 65, 66, 67, 68, 68\}$.

   The Median is given by: $(65 + 66)/2 = 65.5$.

Consider the discrete frequency distribution taking the values $\{x_1, x_2, \ldots, x_n\}$ with corresponding frequencies $\{f_1, f_2, \ldots, f_n\}$. The median is given by the

$$\boxed{\left(\frac{1+\sum f}{2}\right)\textbf{th}}$$

value when the values are ranked.

Example

Find the median of the following discrete distribution:

x_i	0	1	2	3	4	5	6
f_i	5	5	10	20	30	20	10
Cum f	5	10	20	40	70	90	100

$\sum f_i = 100, \quad (1+\sum f_i)/2 = 50.5.$

The 50.5th item falls at $x = 4$ using the Cumulative frequency (Cum f). Hence the *Median is* 4.

The Mode

The MODE of a set of values is defined as the one which occurs with the greatest frequency.

Note that for a set that has no repeated values the mode does not exist.

Example

The mode of the set $\{2,3,3,1,3,2,4,5,8,3,2,4,4,3\}$ is 3.

It should be noted that there should be more than one mode in a set of numbers.

Example

The set $\{8,6,8,5,5,7,6,8,6,9\}$ has the two modes 6 and 8.

The Range

The RANGE of a set of numbers $S = \{x_1, x_2, \ldots, x_n\}$ is given by:

$$\boxed{\textbf{Range} = \max(S) - \min(S).}$$

The Range is the simplest of all measures of dispersion and can be calculated very quickly and easily.

It is not a serious measure of dispersion since it uses the only extreme values.

Examples

- The set $\{6, \boxed{5}, 7, \boxed{10}, 8, 9\}$ has Range $= 10 - 5 = 5$.
- The set $\{600, 610, 620, \boxed{600}, 610, \boxed{650}, 640, 650, 650\}$ has Range $= 650 - 600 = 50$.
- The set $\{600, 610, 620, \boxed{200}, 610, \boxed{1000}, 640, 650, 650\}$ has Range $= 800$.

The Standard Deviation and Variance

The standard deviation is the measure of dispersion used most widely in statistics. It is based on the arithmetic mean.

The standard deviation of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with mean $\bar{x}$ is denoted by S and defined as

$$S = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2}$$

Example

The set $\{3, 4, 6, 2\}$ has $\bar{x} = 15/4 = 3.75$, $\bar{x}^2 = 14.063$ and $\sum_i^n x_i^2 = 65$.

Thus,

$$S = \left(\frac{65}{4} - (3.75)^2\right)^{\frac{1}{2}} = (16.25 - 14.063)^{\frac{1}{2}} = 1.48$$

For a discrete frequency distribution the standard deviation is defined as:

$$S = \sqrt{\frac{\sum_i f_i (x_i - \bar{x})^2}{\sum_i f_i}} = \sqrt{\sum_i \frac{f_i x_i^2}{\sum_i f_i} - \bar{x}^2}$$

where $\bar{x} = \sum_i f_i x_i / \sum_i f_i$ is the mean of the frequency distribution.

Example

Find the median of the following discrete distribution:

							Σ
x_i	0	1	2	3	4	5	
f_i	1	3	11	9	5	2	31
$f_i x_i$	0	3	22	27	20	10	82
x_i^2	0	1	4	9	16	25	
$f_i x_i^2$	0	3	44	81	80	50	258

$$S = \sqrt{\sum_i \frac{f_i x_i^2}{\sum_i f_i} - \left(\frac{\sum_i f_i x_i}{\sum_i f_i}\right)^2} = \sqrt{\frac{258}{31} - \left(\frac{82}{31}\right)^2}$$
$$= 1.15$$

The VARIANCE of a set (or distribution) of numbers is defined as the square of the standard deviation and is denoted by S^2.

For a set of numbers:

$$S^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n} = \frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2.$$

For a frequency distribution:

$$S^2 = \frac{\sum_i f_i (x_i - \bar{x})^2}{\sum_i f_i} = \sum_i \frac{f_i x_i^2}{\sum_i f_i} - \bar{x}^2.$$

Let the set $\{x_1, x_2, \ldots, x_n\}$ be transformed to $\{X_1, X_2, \ldots, X_n\}$, where $X_i = (x_i - a)/b$. If the Standard deviations of X and x are denoted, respectively, by S_X and S_x, then

$$\boxed{S_x = bS_X.}$$

Proof

From $X_i = (x_i - a)/b$ it follows that $x_i = bX_i + a$. Now,

$$\begin{aligned} S_x &= \sqrt{\frac{(\sum x_i - \bar{x})^2}{n}} = \sqrt{\frac{(\sum bX_i + a - (a + b\bar{X}))^2}{n}} \\ &= b\sqrt{\frac{(\sum X_i - \bar{X})^2}{n}} \\ &= bS_X. \end{aligned}$$

$\square$

Example

Find the standard deviation of the set $\{2678, 4678, 8678, 5678, 6678\}$.

Let $X_i = (x_i - a)/b$, where $a = 2678$ and $b = 1000$. This gives the new set $X = \{0, 2, 6, 3, 4\}$ with mean $\bar{X} = 3$ and standard deviation $S_X = \sqrt{20/5} = 2$. Thus,

$$S_x = bS_X = 1000 \times 2 = 2000.$$

Statistical measures & Distributions

Lecture 3

PROBABILITY DISTRIBUTIONS

1. Binomial
2. Poisson
3. Normal

Binomial Distribution

<u>COMBINATIONS</u>

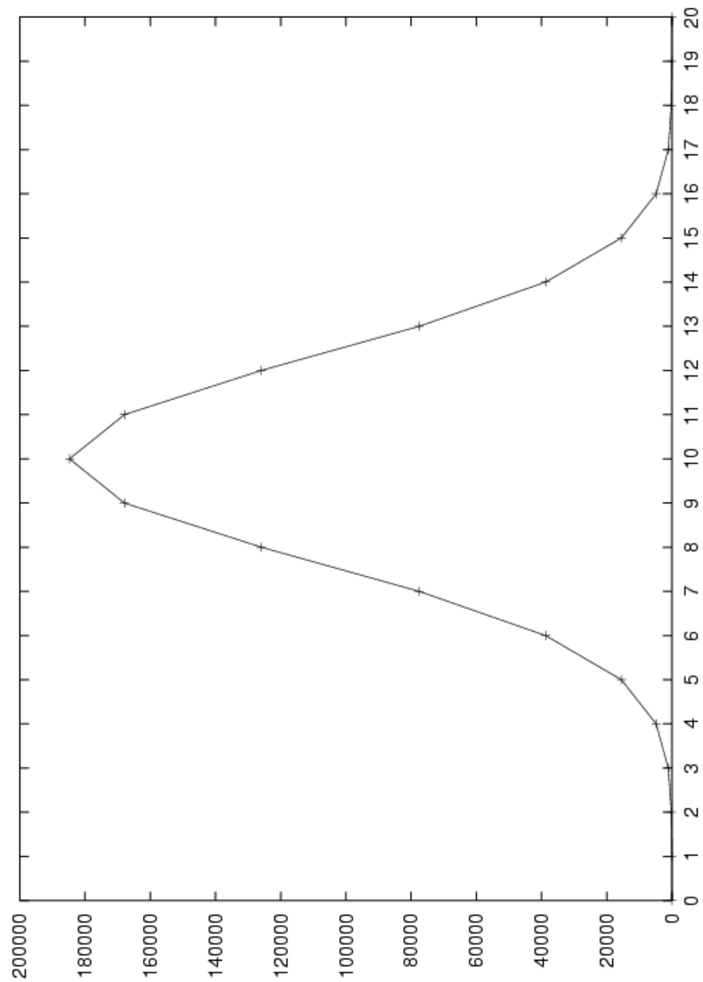
A COMBINATIONS is a non-ordered subset of a set of elements. Element order does not matter when determining Combinations.

The number of combinations of r elements taken from a set of n elements is given by:

$$\boxed{\binom{n}{r} = \frac{n!}{r!(n-r)!} \quad \textbf{where} \quad r \leq n.}$$

The $n!$ (n factorial) is given by $n! = 1 \times 2 \times \cdots \times n$ and $0! = 1$.

The expression $\binom{n}{r}$ can be written as C_r^n, or ${}_nC_r$, or $C(n,r)$.



$\binom{20}{x}$

Example

Four assets designated A, B, C and D are considered by a fund manager. However he is only allowed to invest in three of them. How many different portfolios are possible?

Solution

The fund manager makes a selection of 3 assets from A, B, C and D without taking order in to account. He has 4 possible selections:

$$ABC, \quad ABD, \quad ACD, \quad BCD.$$

Otherwise, $n = 4$ and $r = 3$ so that

$$\binom{n}{r} = \binom{4}{3} = \frac{4!}{3!(1)!} = 4.$$

The BINOMIAL distribution arises in many applications where you are counting events.

The binomial distribution arises from the following assumptions:

1. It has a fixed number of trials, say n.
2. A random variable can take only two values (*success* or *failure*.)
3. Each trial is independent.
4. The probability of success, say p, is constant throughout the experiment.

Example

Consider an investor who is taking a buy position in 3 assets. If the price of the asset goes up (down), then the investor makes a profit (loss). Assume that the price distribution of the 3 assets included in the portfolio are independent. Now, let p ($q = 1 - p$) be the probability of the price going up (down). That is, $P(U) = p$ and $P(D) = q = 1 - p$. The probability of the first two assets going up and one going down is given by:

$$P(UUD) = P(U)P(U)P(D) = p\,p\,q = p^2\,q.$$

Note that $P(UUD) = P(UDU) = P(DUU)$.

Let $B(x;3;p)$ denote the probability that the prices of x assets rise. That is,

Port.	# of rises	Prob.
DDD	0	q^3
DDU	1	pq^2
DUD	1	pq^2
UDD	1	pq^2
DUU	2	p^2q
UDU	2	p^2q
UUD	2	p^2q
UUU	3	p^3

or

x	$B(x;3;p)$
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

or

$$B(x;3;p) = \binom{3}{x} p^x (1-p)^{3-x} \quad \textbf{where} \quad x = 0,1,2,3.$$

Note that
$\sum_1^3 B(x;3;p) = q^3 + 3pq^2 + 3p^2q + p^3 = (p+q)^3 = 1.$

The Binomial distribution has the form:

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

Here the integer $n \geq 0$ and the probability p $(0 \leq p \leq 1)$ are the parameters.

The mean and variance of the Binomial distribution are given by:

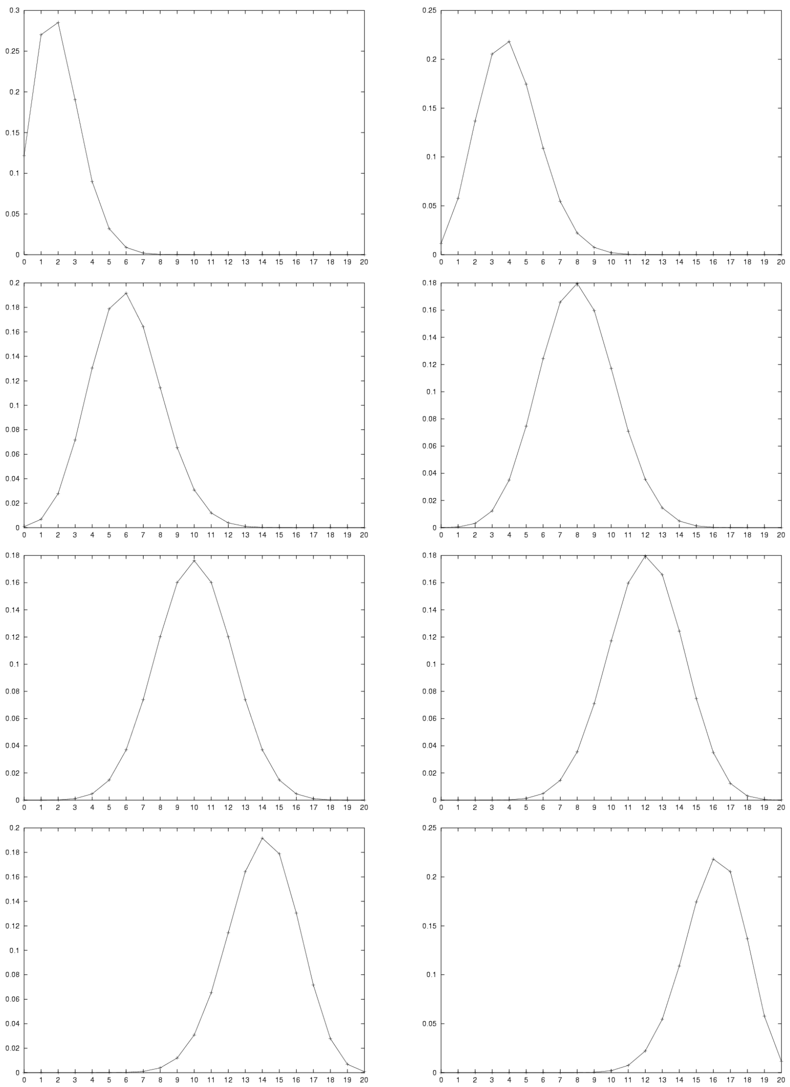
$$\mathbf{E}(x) = n\,p$$
$$\mathbf{Var}(x) = n\,p\,(1-p).$$

The relation

$$\boxed{P(X=x+1) = \frac{p}{1-p}\,\frac{n-x}{x+1}\,P(X=x)}$$

is known as the *Binomial recurrence formula*.

$B(x, 20, p)$, where $p = 0.1, 0.2, \ldots, 0.8, 0.9$.



Example

In tossing a coin the probability of a head is 0.5. If the coin is tossed 5 times, then what is the probability of (a) exactly 2 heads and (b) more than one head.

Let x denote the number of heads, $n = 5$ and $p = q = 1 - p = 1/2$.

1. $P(X = 2) = B(2; 5; p) = \binom{5}{2} p^2 (1-p)^3 = 5/16$.

2.

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - \Big(P(x = 0) + P(x = 1)\Big) \\ &= 1 - B(0; 5; 0.5) - B(1; 5; 0.5) \\ &= 13/16. \end{aligned}$$

Given a frequency distribution the binomial distribution can be fitted by:

1. Deriving the values of the parameters n and p either
   (a) From a know binomial situation
   (b) Computing the mean $\bar{x}$ of the frequency distribution and using the relation $\bar{x} = np$.
2. Generating a Binomial probability distribution using n and p.
3. Generate the *Expected* frequencies by multiplying the total frequencies $\sum_i f_i$ by the probability.

Example

A *biased* die is thrown 5 times as an experiment. The experiment is repeated 250 times. The number of even numbers shown on the die in each experiment is recorded giving the results:

# of *evens*	0	1	2	3	4	5	Total
Observed freq.	11	41	83	73	36	6	250

FIT A BINOMIAL DISTRIBUTION TO THIS DATA.

Step 1

The number of trials $n = 5$. The relationship $\bar{x} = np$ can be used in order to compute p.

Now,

x_i	0	1	2	3	4	5	Total
f_i	11	41	83	73	36	6	250
$f_i x_i$	0	41	166	219	144	30	600

It follows that $\bar{x} = \sum_i f_i x_i / \sum_i f_i = 600/250 = 2.4$.

Thus, from $\bar{x} = np$ it implies $2.4 = 5p$, i.e. $p = 0.48$.

Step 2

Let the random variable X which represent the *number of evens in the experiment* to have the binomial distribution:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \textbf{for } x = 0, 1, 2, 3, 4, 5,$$

where $n = 5$ and $p = 0.48$.

Now $P(X=0) = \binom{5}{0} p^0 (1-p)^5 = 0.038$ and

$$P(X=x+1) = \frac{p}{1-p}\,\frac{n-x}{x+1}\,P(X=x)$$
$$= 0.9231\,\frac{5-x}{x+1}\,P(X=x).$$

From the latter it follows that $P(X=1)=0.1754$, $P(X=2)=0.324$, $P(X=3)=0.2990$, $P(X=4)=0.1380$ and $P(X=5)=0.0255$.

<u>Summary</u>

x_i**:** The number of *evens*.

f_i**:** Actual frequency.

$P(X=x_i)$**:** Probability of obtaining x_i *evens*.

$P(X=x_i)\sum_i f_i$**:** Expected frequency.

x_i	0	1	2	3	4	5	Total
f_i	11	41	83	73	36	6	250
$P(X=x_i)$	0.04	0.18	0.32	0.3	0.14	0.03	0.9999
$P(X=x_i)\sum_i f_i$	10	44	81	75	34	6	250

Poisson Distribution

The discrete random variable X is said to have a Poisson distribution if it has a pdf of the form

$$\boxed{P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \textbf{for } x = 0,1,\ldots,n,}$$

where $\lambda > 0$ is the parameter.

The mean and variance of the Poisson distribution are given by:

$$\boxed{\mathbf{E}(x) = \mathbf{Var}(x) = \lambda.}$$

The Poisson distribution can be defined as the limiting case of the binomial distribution for $n \to \infty$ but with constant $np = \lambda$. Thus, it describes the behavior of a large number n of independent experiments of which only a very small fraction np is expected to yield events of a given type.

Example

The claim experience of 5000 policies, each expose to risk for a year, is summarized in the table below:

x_i	0	1	2	3	Total
f_i	3695	1120	160	25	5000

Here x_i is the number of claims and f_i the observed number of policies.

Questions

1. Calculate the average number of claims.
2. Assume a Poisson distribution. Compare the predictions of the theoretical distribution with the observed number of policies.

Answer

1. The average number of claims is calculated by:

$$\lambda = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{1515}{5000} = 0.303$$

2. First calculate $P(X = x)$ for $x = 0, 1, 2, 3$.

- $P(X = 0) = \dfrac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = 0.7386.$
- $P(X = 1) = \dfrac{e^{-\lambda}\lambda^1}{1!} = P(X = 0) \times \dfrac{\lambda}{1} = 0.2238.$
- $P(X = 2) = \dfrac{e^{-\lambda}\lambda^2}{2!} = P(X = 1) \times \dfrac{\lambda^2}{2} = 0.0339.$
- $P(X = 3) = P(X = 2) \times \dfrac{\lambda^3}{3} = 0.0034.$

Now the predicted number of policies for x claims is given by $\hat{f}_i = P(X = x_i)\sum_i f_i$. Thus, $\hat{f}_1 = P(X = x_i)\sum_i f_i = 0.7386 \times 5000 = 3693.$

Number of claims	Probab. of claims per policy	Number of policies Predicted	Actual
x_i	p_i	$\hat{f}_i$	f_i
0	0.7386	3693.0	3695
1	0.2238	1119.0	1120
2	0.0339	169.5	160
3	0.0034	17.0	25
TOTAL	0.9997	4998.5	5000

Comparing the theoretical with the empirical values we observe that a Poisson distribution with parameters value of $\lambda = 0.3$ describes this particular random variable very well.

Observations

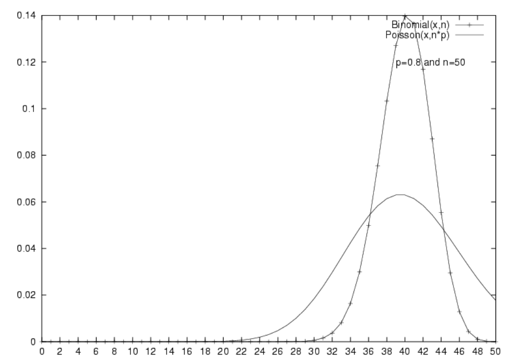
1. $\sum_i p_i = 0.9997 \approx 1$.
2. $\mathrm{E}(x) = \sum_i x_i p_i = 0.3018 \equiv \lambda$.
3. $\mathrm{E}(x^2) = \sum_i x_i^2 p_i = 0.39$.
4. $\mathrm{Var}(x) = \mathrm{E}(x^2) - (\mathrm{E}(x))^2 = 0.299 \approx \lambda$. .

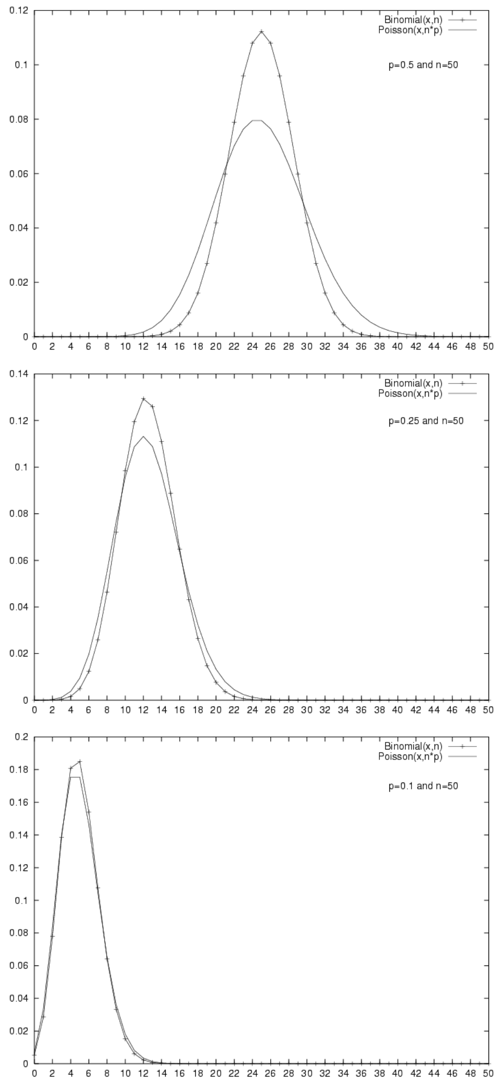
Binomial approximation to Poisson distribution

A Binomial distribution with parameters n and p can be approximated by a Poisson distribution with parameter $\lambda = n\,p$ if n is large and p is small. That is, $n \to \infty$ and $p \to 0$.

Note that if $p \to 0$, then $q = 1 - p \approx 1$. Thus, the variance of the Binomial distribution is given by:

$$\boxed{\mathbf{Var}(x) = n\,p\,(1-p) \approx n\,p\,1 = n\,p = \mathbf{E}(x).}$$





Example

Consider an individual who has insured his car against theft. The probability of a theft in any 24-hour period, leading to an insurance claim is 0.005. The probabilities of claims on successive days are independent. In addition it is not possible to have more than one theft (leading to a claim) on the same day.

Calculate the probability that a policyholder makes at least 3 claims in a year.

The probability of theft is $p = 0.005$. The probability of the car not to be stolen (i.e. No theft) is $q = 1 - p = 0.995$. The total number of trials in a year is the number of days, that is, $n = 365$.

The probability of at least 3 claims is given by:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - \Big(P(X=0) + P(X=1) + P(X=2)\Big). \end{aligned}$$

Using the Binomial distribution and recurrence formula:

$$P(X=0) = \binom{365}{0} p^0 q^{365} = 0.1605,$$
$$P(X=1) = 0.2944 \quad \textbf{and}$$
$$P(X=2) = 0.2692.$$
$$\textbf{Thus,} \quad P(X \geq 3) = 0.276\textbf{.}$$

Now, Using an approximation to the Poisson Distribution. Let $\lambda = np = 365 \times 0.005 = 1.825$. Note that $npq = 1.815$.

$$P(X=0) = \frac{e^{-\lambda}\lambda^0}{0\textbf{!}} = 0.1612,$$
$$P(X=1) = 0.2942 \quad \textbf{and}$$
$$P(X=2) = 0.2685.$$
$$\textbf{Thus,} \quad P(X \geq 3) = 0.2761\textbf{.}$$

The difference using a Poisson distribution has relative error 0.04% which is extremely small.

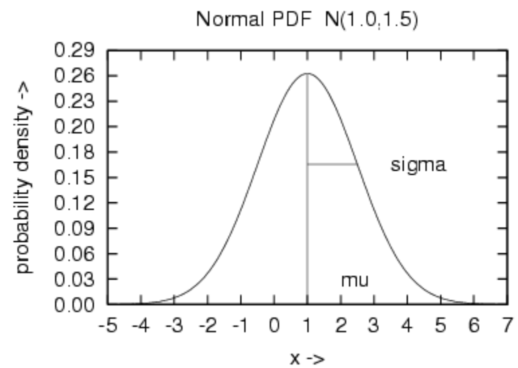
The Normal Distribution

The Binomial and Poisson distributions were important examples of special distributions of the discrete kind. The NORMAL distribution can be described as the single most important continues distribution in statistics.

The Normal distribution has two parameters: the mean μ and the standard deviation σ. Its shorthand notation is $N(\mu,\sigma)$. The pdf of $N(\mu,\sigma)$ is bell shaped and is symmetrical about the mean. The formula is:

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2} \quad \textbf{for } -\infty < x < \infty\textbf{,}}$$

where $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$.



A normal distribution having $\mu = 0$ and $\sigma^2 = 1$, i.e. $N(0,1)$, is called a STANDARD NORMAL DISTRIBUTION. The random variable associated with this distribution is usually denoted by Z. That is, $Z \sim N(0,1)$. The pdf of Z is given by:

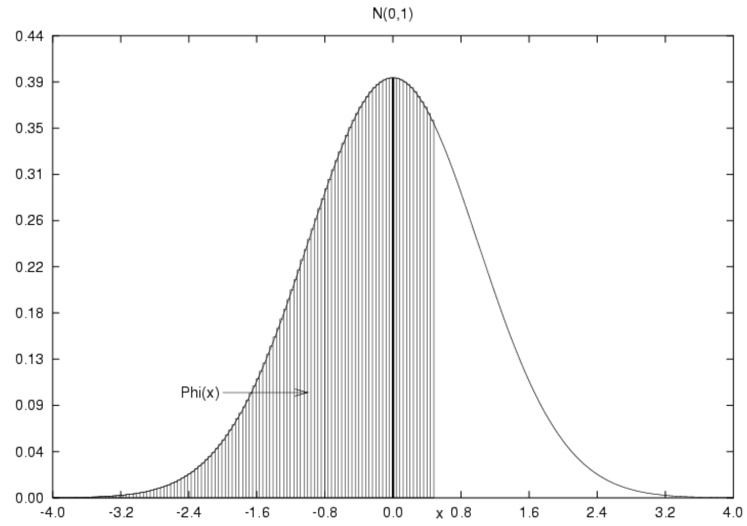
$$\boxed{f(x) = \frac{1}{\sqrt{2\Pi}} e^{-\frac{x^2}{2}} \quad \textbf{for } -\infty < x < \infty.}$$

The distribution function of a standard normal variable Z is denoted by:

$$\Phi(x) = P(Z < x) = \int_{-\infty}^{x} f(x)\,dx$$

$$= \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}}\,dx.$$

Close form expression for the integral does not exist. Hence its evaluation can only be obtained by approximate procedures. Therefore, areas under the normal density function are presented in tables.



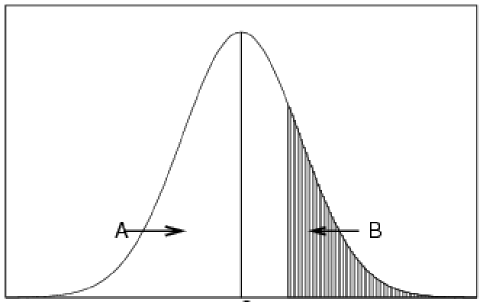
<u>Example</u>

$P(Z < 0.1) = \Phi(0.1) = 0.5398.$

The Normal distribution tables only give values of $\Phi(x)$ for $x \geq 0$. The probabilities such as $P(Z < -0.1)$ and $P(Z \geq 0.3)$ have to be *transformed* into probabilities of the type $P(Z < x)$, where $x \geq 0$.

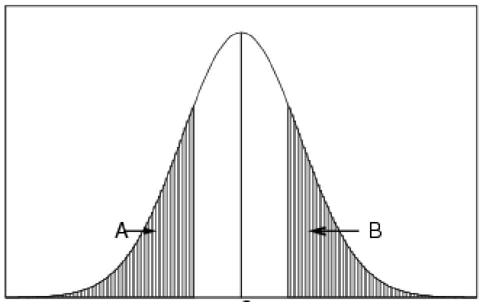
Let $Z \sim N(0,1)$ and $\Phi(x) = P(Z < x)$. If $a \geq 0$, then

1. $P(Z > a) = 1 - P(Z < a) = 1 - \Phi(a)$.



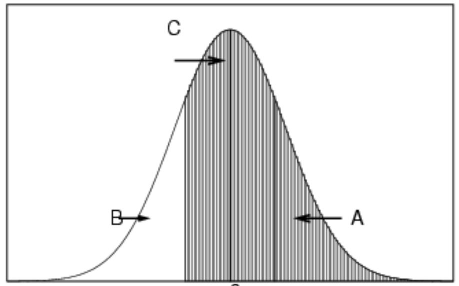
$P(Z > a) = \textbf{Area(B)} = \textbf{Total Area - Area(A)} = 1 - \Phi(a)$.

2. $P(Z < -a) = \Phi(-a) = 1 - \Phi(a)$.



$\Phi(-a) = \textbf{Area(A)} = \textbf{Area(B)} = P(Z > a) = 1 - \Phi(a)$.

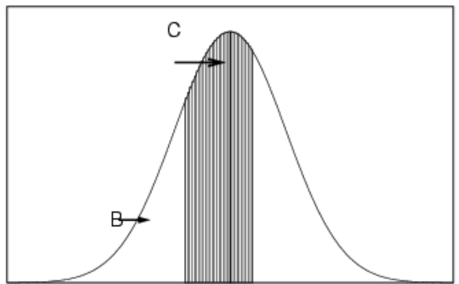
3. $P(Z > -a) = \Phi(a)$.



$P(Z > -a) = \textbf{Area(C)} + \textbf{Area(A)}$

$= \textbf{Area(C)} + \textbf{Area(B)} = P(Z < a) = \Phi(a)$.

4. If b and c are any positive or negative numbers such that $b \leq c$, then $P(b < Z < c) = \Phi(c) - \Phi(b)$.



$P(b < Z < c) = \textbf{Area(C)} = \textbf{Area(B+C)} - \textbf{Area(B)}$

$= P(Z < c) - P(Z < b) = \Phi(c) - \Phi(b)$.

Example 1

The random variable $X \sim N(\mu, \sigma^2)$ denotes the number of claims per year, where $\mu = 100$ and $\sigma = 4$. Find the probabilities of the number of claims to be:

1. Less that 90.
2. More than 108.
3. Between 96 and 104 (including).

Solution

1.

$$P(X < 90) = P\Big(\frac{X-100}{4} < \frac{90-100}{4}\Big)$$
$$P(Z < -2.5) \quad \textbf{since} \quad (X-\mu)/\sigma \sim N(0,1)$$
$$1 - \Phi(2.5) = 0.0062$$

2.

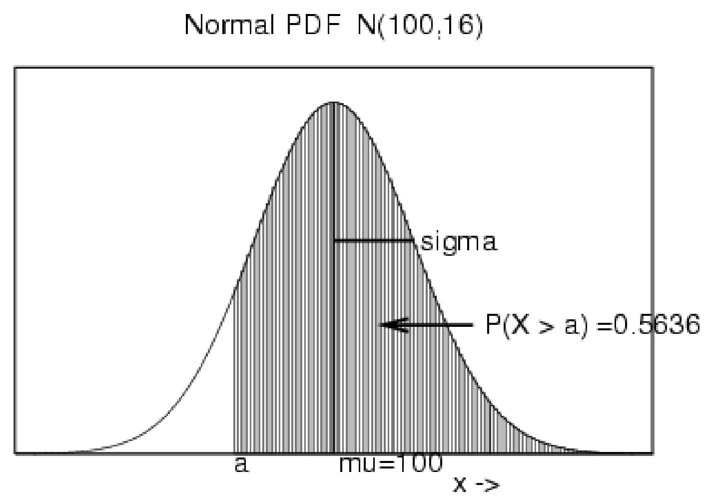
$$P(X > 108) = P\Big(\frac{X-100}{4} > \frac{108-100}{4}\Big)$$
$$P(Z > 2.0) = 1 - \Phi(2.0)$$
$$1 - 0.9772 = 0.0228.$$

3.

$$P(95 < X < 105) = P\Big(\frac{95-100}{4} < \frac{X-100}{4} < \frac{105-100}{4}\Big)$$
$$= P(-1.25 < Z < 1.25)$$
$$= \Phi(1.25) - \Phi(-1.25)$$
$$= 2 \times \Phi(1.25) - 1 = 0.7888.$$

Example 2

The probability of having more than a claims is 0.5636. What is the value of a, when $X \sim N(100, 16)$.



Since the probability given is greater than 0.5 then a must be less than the mean $\mu = 100$.

Now, $P(X > a) = 0.5636$ and thus,

$$P\Big(\frac{X-100}{4} > \frac{a-100}{4}\Big) = 0.5636$$

$$\textbf{or} \quad P\Big(Z > \frac{a-100}{4}\Big) = 0.5636.$$

Since a is less than the mean, then $(a-100)/4 < 0$. Therrefore,

$$P\Big(Z > \frac{a-100}{4}\Big) = P\Big(Z > -(\frac{100-a}{4})\Big) = \Phi\Big(\frac{100-a}{4}\Big).$$

Hence,

$$\Phi\Big(\frac{100-a}{4}\Big) = 0.5636$$

From the tables it follows that

$$\frac{100-a}{4} = 0.16 \quad \textbf{and thus,} \quad a = 99.36.$$

Example 3

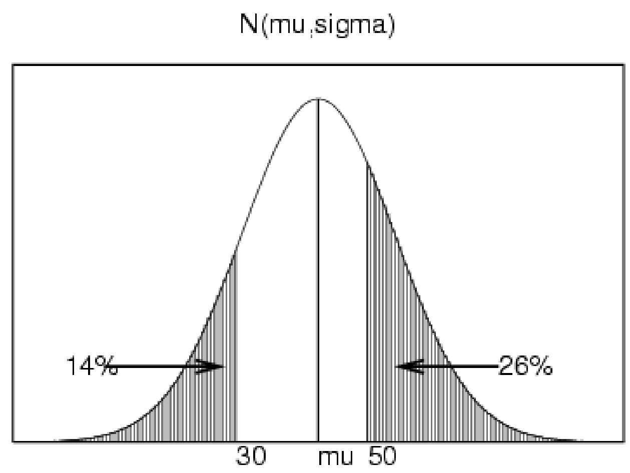
A policyholder has a large number of policies. If the policies are known to be normally distributed and

1. 14% of the policies gave rise to less than 30 claims;
2. 26% gave rise to more than 50 claims.

FIND THE MEAN AND VARIANCE OF THE CLAIMS.

Answer

Let X denotes the claims, i.e. $X \sim N(\mu, \sigma^2)$. The information given by (1) and (2) can be written as $P(X < 30) = 0.14$ and $P(X > 50) = 0.26$. Graphically this can be illustrated as:



Now, $P(X < 30) = 0.14$ can equivalently be written as

$$P\Big(\frac{X-\mu}{\sigma} < \frac{30-\mu}{\sigma}\Big) = 0.14$$

$$\textbf{or} \quad P\Big(Z < -(\frac{\mu-30}{\sigma})\Big) = 0.14$$

$$\textbf{or} \quad \Phi\Big(-(\frac{\mu-30}{\sigma})\Big) = 0.14$$

$$\textbf{or} \quad 1-\Phi\Big(\frac{\mu-30}{\sigma}\Big) = 0.14 \quad \textbf{or} \quad \Phi\Big(\frac{\mu-30}{\sigma}\Big) = 0.86.$$

From the tables it follows that $\Phi(1.08) = 0.8599$. Thus,

$$\frac{\mu-30}{\sigma} = 1.08 \quad \textbf{or} \quad \mu - 1.08\times\sigma = 30. \qquad (1)$$

Clearly, from $P(X > 50) = 0.26$ it follows that

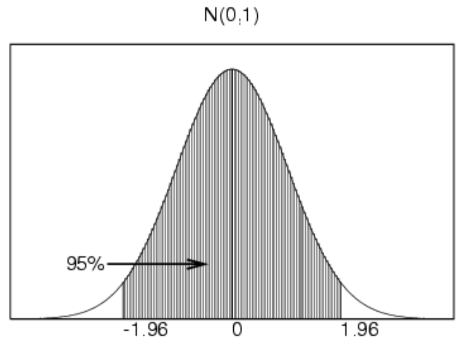
$$P\Big(Z > \frac{50-\mu}{\sigma}\Big) = 0.26, \quad \textbf{where} \quad Z = \frac{X-\mu}{\sigma}.$$

Thus,
$1-\Phi\Big(\frac{50-\mu}{\sigma}\Big) = 0.26$ or $\Phi\Big(\frac{50-\mu}{\sigma}\Big) = 0.74$. From the tables it follows that

$$\frac{50-\mu}{\sigma} = 0.643 \quad \textbf{or} \quad \mu + 0.643\times\sigma = 50. \qquad (2)$$

From (1) and (2) it follows that the mean $\mu = 42.54$ and variance $\sigma^2 = (11.61)^2 = 134.8$.

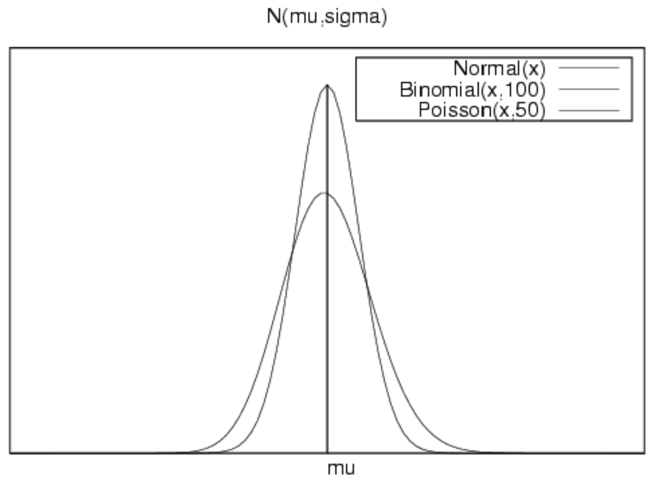
The central 95% of a standard normal distribution lies between the limits ± 1.96. Alternatively, the central 95% of any normal distribution lies within 1.96 standard deviations of its means.



The central 99% (99.8%) of a standard normal distribution lies between the limits ± 2.58 (± 3.09).

Normal Approximation to the Binomial and Poisson

- If X is distributed binomially with parameters n and p, i.e. $X \sim Bin(n,p)$, then for large n and not too small (or too large) p we can consider $X \sim N\big(np, np(1-p)\big)$.
  The number of trial n should, in general, $n > 50$ and $p \approx 0.5$.
- If X has a Poisson distribution with parameter μ, i.e. $X \sim Po(\mu)$, then for large μ, we canconsider $X \sim N(\mu,\mu)$.
  In general the parameter μ>20 for good approximations.



Central limit theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from *any distribution* with mean μ and variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has an approximate normal distribution with mean μ and variance σ^2/n. That is, approximately

$$\boxed{\bar{X} \sim N(\mu, \sigma^2/n).}$$

The approximation gets better for large n.

Example

Let $X_1, X_2, \ldots, X_n$ denote a random sample of claims from Poisson distribution with parameter $\mu = 3$, where $n = 20$. Using the central limit theorem find the approximation that the sample mean will be greater than 4.

Solution

If X is Poisson with parameter 3, then $\mathrm{E}(X) = \mu = 3$ and $\mathrm{Var}(X) = \sigma^2 = \mu = 3$. From the Central Limit Theorem it follows that $\bar{X} \sim N(\mu, \mu/n) = N(3, 0.15)$ approximately.

Hence,

$$\begin{aligned}P(\bar{X} > 4) &= P\Big(\frac{\bar{X}-3}{\sqrt{0.15}} > \frac{4-3}{\sqrt{0.15}}\Big)\\ &= P(Z > 2.58) \quad (\text{where } Z = (\bar{X}-3)/\sqrt{0.15})\\ &= 1-\Phi(2.58) = 0.005.\end{aligned}$$

Example

A population of insured clients has a mean claim (in pounds) of $\mu = 69$ and a standard deviation $\sigma = 3.22$. If a random sample of $n = 10$ insured is drawn, then what is the chance that the mean $\bar{X}$ will be within £2 of the population mean μ ?

Solution

We want to find

$$\begin{aligned}P(|\bar{X}-\mu| < 2) &= P(-2 < \bar{X}-\mu < 2)\\ &= P(67 < \bar{X} < 71)\\ &= 1-P(\bar{X} < 67) - P(\bar{X} > 71).\end{aligned}$$

Now, according to the central limit theorem, $\bar{X} \sim N(\mu, \sigma^2/n) = N(69, 1.04)$.

Let $Z = (\bar{X}-69)/\sqrt{1.02}$, such that after standardization the $P(|\bar{X}-\mu| < 2)$ can be written as

$$\begin{aligned}P(|\bar{X}-\mu| < 2) &= 1-P(Z < -1.96) - P(Z > 1.96)\\ &= 1-0.025-0.025\\ &= 0.95.\end{aligned}$$

Thus, there is a 95% chance that the sample mean will be within £2 of the population mean.

Example

Assume that a large class in quantitative methods has marks normally distributed around mean of 72 with a standard deviation of 9.

1. Find the probability that an individual student drawn at random will have a mark over 80.
2. Find the probability that a random sample of 10 students will have an average mark over 80. What will be this probability when the population is NOT normal.

Solution

1. Let X denote the marks of the students, where $\mu = 72$ and $\sigma^2 = 81$. Thus, $X \sim N(72, 81)$.

   We want the probability $P(X > 80)$.

   Let $Z = (X - 72)/9$. Notice that $(80 - 72)/9 = 0.89$. Thus, after standardizing

   $P(X > 80) = P(Z > 0.89) = 0.187$.

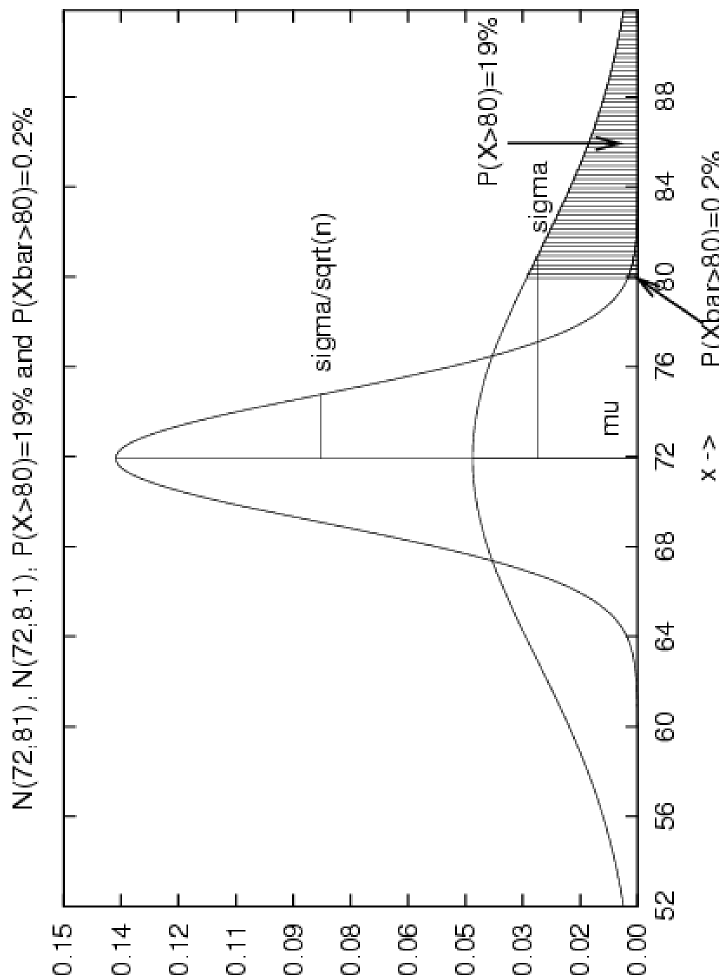
2. From the central limit theorem we have that $\bar{X} \sim N(\mu, \sigma^2/n) = N(72, 8.1)$, where $\bar{X}$ denotes the average mark of a random sample of 10 students.

   We want the $P(\bar{X} > 80)$.

   Now let, $Z = (\bar{X} - 72)/2.85$ and notice that $(80 - 72)/2.85 = 2.81$. Thus, after standardizing:

   $P(\bar{X} > 80) = P(Z > 2.81) = 0.002$.

   The CLT implies that for sufficient large n the $\bar{X} \sim N(\mu, \sigma^2/n)$ (approximately) no matter what the distribution of the parent population. Thus, $P(\bar{X} > 80)$ is approximately 0.2%.



Interval estimation

The purpose of interval estimation is to construct ranges of values within which the population parameters are expected to lie within a given probability based on the results of a random sample.

A *B% confidence interval* (C.I.) for some unknown parameter θ is an interval constructed based on the results of a random sample so that the probability θ lies in this interval is $B/100$.

The most commonly used is a 95% C.I. If (a,b) constitute a 95% C.I. for some parameter θ, then we have in probability terms: $P(a \leq \theta \leq b) = 0.95$.

The construction of the intervals that we consider will be based on the sample values of unbiased estimators for the particular parameters that we are interested.

The mean (known variance)

If $\bar{X}$ is the mean of a random sample of size n from a normal distribution with known variance σ^2, the a central 95% C.I. for μ, the population mean, is given by $\bar{X} \pm 1.96\sigma/\sqrt{n}$. That is,

$$\boxed{P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.} \qquad (3)$$

Exercise

An experiment was carried out in which it was found that the height in cm of a plan were

$\{12.3, 11.8, 11.6, 12.6, 13.4, 12.8, 11.1, 12.2, 14.8, 13.1\}$

Given that the height of the plans are approximately normally distributed with variance $1.44cm$, construct a 95% C.I. for the mean of the population heights μ.

Solution

$P(11.83 \leq \mu \leq 13.31) = 0.95$.

The mean with unknown variance (small samples)

Let $\bar{X}$ and S^2 denote the mean and variance of a random sample of size n drawn from a normal population with unknown mean μ and unknown variance σ^2. The central $B\%$ C.I. for μ is given by $\bar{X} \pm t\, S/\sqrt{n-1}$, where t is such that the interval $(-t, t)$ encloses $B\%$ of a $T(n-1)$ distribution. That is,

$$\boxed{P\Big(\bar{X} - t\,\frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t\,\frac{S}{\sqrt{n-1}}\Big).} \qquad (4)$$

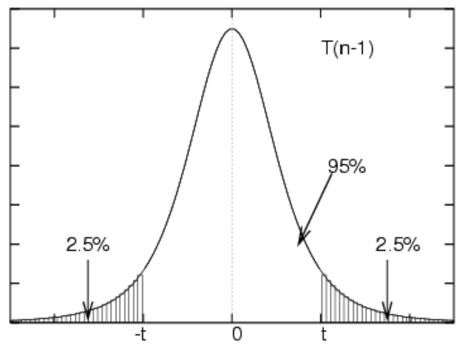
Note that the latter form is also used for samples from populations that are approximately normal.

Example

Eight shares of a certain PLC are reported to be (in £) $\{10.6, 11.2, 10.4, 12.2, 11.3, 10.2, 10.3, 12.5\}$.

Find a 95% confidence limits for the mean price of a sample of the PLC assuming these prices are coming from a normal distribution.

Solution



The confidence limits take the form $\bar{X} \pm t\, S/\sqrt{n-1}$, where $(-t, t)$ is the interval of a $T(n-1)$ distribution enclosing the central 95% of the distribution. From T tables with $n = 7$ we have that $t = 2.365$. That is, $T_{2.5\%}(7) = 2.365$.

From the share prices we have
$\bar{X} = (\sum X_i)/n = 88.7/8 = 11.09$ and $\sum X_i^2 = 988.7$.
Thus, $S^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = 0.68$ and $S = 0.82$.

From the latter it follows that 95% C.I. is

$$11.09 \pm \frac{2.365 \times 0.82}{\sqrt{7}} = 11.09 \pm 0.73 = (10.36, 11.82).$$

Thus, we have a probability of 95% that the mean share price of the PLC will lie between 10.36 and 11.82.

The mean with unknown variance (Large sample approx.)

For large values of n, the $T(n)$ closely resembles the standard Normal distribution, i.e. $N(0,1)$. For example $Z_{2.5\%} = 1.96$ which implies that $(-1.96, 1.96)$ encloses the central 95% of a $N(0,1)$ distribution. Similarly, from column $P = 2.5\%$ and $n = 120$ of T-tables we obtain $t = 1.98$. That is, the central 95% of a $T(120)$ distribution lies within the interval $(-1.98, 1.98)$.

Let $\bar{X}$ and S^2 denote the mean and variance of a random sample of size n (large) from a normal population with unknown mean μ and unknown variance σ^2. A central $B\%$ confidence interval for μ is given (approximately) by $\bar{X} \pm zS/\sqrt{n}$, where z is the $0.5(100-B)\%$ point of a $N(0,1)$ distribution. That is,

$$\boxed{P\Big(\bar{X} - z\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{S}{\sqrt{n}}\Big) = \frac{B}{100}.}$$

Note that $n \geq 30$ will be considered adequately large.

Example

One hundred shares from the same kind of business are taken at random and their prices are found to have mean 69 and variance 7. Find 95% and 98% confidence limits for the mean price of the shares.

Solution

From the large sample $n = 100$ we have the mean $\bar{X} = 69$ and variance $S^2 = 7$. Since n is large we compute $\bar{X} \pm zS/\sqrt{n}$. For 95% C.I., the $z = 1.96$ and gives the limits as: $69 \pm 1.96(\sqrt{7/100}) = 69 \pm 0.52$.

That is, the 95% C.I. is given by $(68.48, 69.52)$.

For 98% C.I., the $z = 2.33$. Thus, ...