Mr G's Little Book on

Probability Distributions

Preface

MS Word supplies over 10 000 characters ranging from the sublime to the ridiculous but not those key expressions required in statistics. $\bar{x}$ is in Arial Unicode and looks reasonable but capital x ie $\bar{X}$ is undersized. In both cases Word then shifts open the line spacing.

Placing a circumflex above a Greek letter is all but impossible – the best I could realise was $\sigma^{\wedge}$.

Determined not to use equation editor a definite integral between a and b is ${}_a\int^b$ and the result in square brackets is [some $f(x)]_a^b$. Where the limits have suffixes the result looks a bit clumsy ${}_{x1}\int^{x2}$.

Robert Goodhand

Basic Definitions

Mean $\bar{x} = {}^1/_n \sum_{i=1}^{n} x_i$

or more simply just $\bar{x} = {}^1/_n \sum x_i$

For ungrouped data $\bar{x} = {}^{\sum f(x)} / {}_{\sum f}$

for grouped date we can only estimate the mean by taking the midpoint of each interval to represent that interval

midpoint = ½ (lower class boundary + upper class boundary)

so statistically we do live in a classless society!

Range

range = highest value – lowest value ignoring correction factors.

Mean Deviation

mean deviation = $\sum | x_i - \bar{x} | / n$

and for a frequency distribution

mean deviation = $\sum | f_i (x_i - \bar{x} | / \sum f_i$

This measure is not widely used.

Standard Deviation

$s = \sqrt{\{ \sum (x_i - \bar{x})^2 / n\}}$

Variance

variance = (standard deviation)^2

$s^2 \quad = {}^1/_n \sum (x_i - \bar{x})^2$

$= {}^1/_n \sum (x_i^2 + \bar{x}^2 - 2\, x_i\, \bar{x})$

$= {}^1/_n (\sum x_i^2 + \sum \bar{x}^2 - 2\, \bar{x} \sum x_i)$

$= \sum x_i^2 / n + \sum \bar{x}^2 / n - 2\, \bar{x} \sum x_i / n$

$= \sum x_i^2 / n + \bar{x}^2 - 2\, \bar{x}^2$

Hence we derive the key relationship

$s^2 \quad = \sum x_i^2 / n - \bar{x}^2$

Worked Example

Find the mean and standard deviation of the first n integers

$\bar{x} \quad = {}^1/_n \sum x_i$

$= {}^1/_n \, ½\, n\, (n+1)$

$= ½\, (n+1)$

This result is used when determining the midpoint term of n terms

$s^2 \quad = \sum x_i^2 / n - \bar{x}^2$

$= {}^1/_6\, n\, (n+1)(2n+1) / n - [½(n+1)]^2$

which reduces to ${}^1/_{12}\, (n^2 - 1)$

Scaling

If each number in a data set in increased by a constant c then

$\bar{x}_n = \bar{x}_0 + c$ but

$s_n = s_0$

so adding c affects the mean but leaves the standard deviation unchanged

If each number in the data set is multiplied by constant a

$\bar{x}_n = a\,\bar{x}_0\ s_o$

$s_n = a\,s_0$

To summarise if

$y = ax + b$ then

$\bar{y} = a\,\bar{x} + b$

$s_y = a\,s_x$

Coding

if $y = (x – a) / b$

$x = a + by$

$\bar{x} = a + b\,\bar{y}$

$s_x = bs_y$

Interquartile Range (IQR)

$IQR = Q_3 – Q_1$

On a normal distribution this occurs at approximately $\pm\, ^2/_3\ s$

$Q_1 = ¼\ (n + 1)^{th}$ value

$Q_3 = ¾\ (n + 1)^{th}$ value

Semi Interquartile Range

$SIQR = ½\ (Q_3 – Q_1)$

Median

The median value is the $½\ (n+1)^{th}$ observation.

Histograms

Frequency density

= frequency / class width

if the intervals are of equal width

height "bar" = frequency

if the intervals are of unequal width

height "bar" = frequency density

and finally note that area α frequency

Skew

For positive skew in general

<u>mode</u> ‹ median ‹ <u>mean</u>

For negative skew in general

<u>mode</u> › median › <u>mean</u>

Principle comparators are mode and mean

Coefficient of Skew

mean – mode ≈ 3 (mean – median)

which leads to

Pearson's coefficient

$= {}^{3\text{ (mean – median)}} / {}_{\text{std. deviation}}$

$= {}^{\text{(mean – mode)}} / {}_{\text{std. deviation}}$

The quartile coefficient of skew is given by $\{(Q_3 – Q_2) – (Q_2 – Q_1)\} / (Q_3 – Q_1)$

Outliers

Outliers are defined as values that lie more than $^3/_2\ (Q_3 – Q_1)$ above Q_3 or below Q_1

For a normal distribution this approximates to $\pm 2\,^2/_3 \times$ std. dev.

Chebychev's Theorems

1) The mean and median cannot differ by more than one standard deviation. This sets the coefficient of skew measure at $\pm$ 3

2) For ANY distribution the proportion of the population that lies outside k standard deviations is less than $^1/_{k^2}$

This sets the boundary condition for any unknown distribution.

In probabilistic terms we say

$\Pr(|X - \mu|) \geq k\sigma \leq {}^1/_{k^2}$

where s is the population std. dev.

Arrangements

Number of ways of arranging n unlike objects in a line is n!

Number of ways of arranging n objects of which p are alike is n! / p!

Number of ways of arranging n objects of which p are of one type, q are of another type etc. is n! / p! q! …

Number of ways of arranging n unlike objects in a ring where clockwise/anticlockwise are taken as different arrangements is (n – 1)!

Number of ways of arranging n unlike objects in a ring where clockwise/anticlockwise are taken as the same arrangement is ½ (n – 1)!

Permutations

Permutations are ordered subsets

$^nP_r = n! / (n - r)!$

Combinations

Combinations are unordered subsets

$^nC_r = n! / (n - r)!\, r!$

Probability Distributions

we use capital letters to represent the variable X

Discrete Random Variables (drv)

we use lower case letters to represent specific discrete values that the variable X can take

$x_1, x_2, x_3, \ldots x_n$

with probabilities of occurrence as

$p_1, p_2, p_3, \ldots p_n$

X is defined as a discrete random variable if and only if

$p_1 + p_2 + p_3 + \ldots p_n = 1$

$\sum p_i = 1$ where $i = 1, 2, 3, \ldots n$

$_{\text{all } x}\sum P(X=x) = 1$

The function that allocates the probabilities is termed the probability density function (pdf).

Expectation

The expectation of expected value is defined as

$E(X) = {}_{\text{all } x}\sum xP(X=x)$ or $\sum x_i P_i$

For symmetrical or uniform distributions E(X) is the midpoint value.

The concept can be extended to any function of the random variable.

Let g(X) be any function

$E[g(X)] = {}_{\text{all } x}\sum g(x)\, P(X=x)$

This relationship has the following colloraries

Cor1: $E(a) = a$

Cor.2: $E(aX) = a\,E(X)$

Cor.3 $E(aX+b) = aE(X) + b$

Cor.4

$E(f_1(X)+f_2(X) = E[f_1(x)]+E[f_2x]$

or simply $E(X+Y) = E(X) + E(Y)$

X and Y need not be independent.

Variance

Taking the experimental approach

For a frequency distribution with mean $\bar{x}$ and variance s^2 where

$s^2 = {}^{\sum f(x-\bar{x})^2} / {}_{\sum f}$

which can be written as

$s^2 = \sum {}^{fx^2} / {}_{\sum f} - \bar{x}^2$

Taking the theoretical approach

For a discrete random variable X with $E(X) = \mu$ the variable is defined as

$V_{ar}(X) = E(X-\mu)^2 \quad = E(X^2) - \mu^2$

By convention we write this as

$V_{ar}(X) = E(X^2) - E^2(X)$ *pretty neat*

Cor1: $V_{ar}(a) = 0$

Cor.2: $V_{ar}(aX) = a^2\, V_{ar}(X)$

Cor.3 $V_{ar}(aX + b) = a^2\, V_{ar}(X)$

where a and b are any constants.

Cumulative Distribution Function

Let X be a drv with P(X=x)

then the cdr is given by

$F(t) = P(X \le t) = {}_{x1}\sum^{t} P(X=x)$

$t = x_1, x_2, x_3, \ldots x_n$

Worked example

X is the score from an unbiased die

$F(t) = {}^{t}/_{6}$ where t = 1, 2, 3, 4, 5, 6

It is not necessary that there be a specific formula for the cdf.

Two Random Variable

For X and Y independent

$V_{ar}(X+Y) = V_{ar}(X) + V_{ar}(Y)$

$E(aX+bY) = aE(X) + bE(Y)$ but

$V_{ar}(aX \pm bY) = a^2E(X) \pm b^2E(Y)$.

If X_1 and X_2 are independent observations from distribution X then

$E(X1 + X2) = 2\,E(X)$

$V_{ar}(X1 + X2) = 2\,Var(X)$

For n independent observations

$E(X_1+X_2+X_3\ldots X_n) = nE(X)$ and

$V_{ar}(X_1+X_2+X_3\ldots X_n) = nVar(X)$

In any problem we need to think carefully about whether we are investigating multiples or sums.

Multiples	Sums
$E(2X) = 2E(X)$	$E(X1+X2) = 2E(X)$
$V_{ar}(2X) = 4E(X)$	$V_{ar}(X_1+X_2) = 2V_{ar}(X)$

and the relationship n measurements follows on.

The Binomial Distribution

Where we conduct n independent trials and p is the probability of success in the outcome of any one trial

$X \sim B_{in}(n,p)$

$P(X=x) = {}^nC_x p^x q^{n-x}$ where $p+q = 1$

$P(X=x)$ is thus given by the binomial expansion

$(p+q)^n = {}^nC_0p^0q^n + {}^nC_1p^1q^{n-1} + \ldots + {}^nC_rp^rq^{n-r} + \ldots + {}^nC_np^nq^0$

Hence the following relationships apply

$E(X) = np$

$V_{ar}(X) = npq$

$P(X=r \mid X \sim B_{in}(n,p)$

has the same value as

$P(X=n-r \mid X \sim B_{in}(n,1-p))$

For cumulative distributions

$P(X \leq r \mid X \sim B_{in}(n,p)\ )$

has the same value as

$P(X \geq n-r \mid X \sim B_{in}(n,1-p))$

The mode is the highest value term in the binomial expansion.

Geometric Distribution

Consider performing a number of independent trials with constant probability of success and q of failure $(p+q)=1$ we define

$X \sim G_{eo}(p)$

where the drv X has pdf of the form

$P(X=x) = q^{x-1}p$

Cor.1 $P(X > x) = q^{x-1}p$

Cor,2 $P[(X > a+b) \mid (x > a)] = P(X > b)$

If $X \sim G_{eo}(p)$ then

$E(X) = {}^1/_p$ and

$V_{ar}(X) = {}^q/_{p^2}$

The Poisson Distribution

If an event is randomly scattered in time or space and has a mean number of occurrences λ in a given interval and X is the random variable

"the number of occurrences"

then $X \sim P_o(\lambda)$ and

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

$x = 0, 1, 2, 3, \ldots\infty$ and $l \in R^+$

we know $e^\lambda = 1 + \lambda + {}^{\lambda^2}/_{2!} \ldots$

then clearly $\sum P_i = e^{-l} \times e^l = 1$

so we have a valid pdf.

In general we define

$E(X) = {}_{\text{all } x}\sum xP(X=x)$

so where $X \sim P_o(\lambda)$

$E(X) = 0\times e^{-\lambda}\lambda^0/0! + 1\times e^{-\lambda}\lambda^1/1!$

$\qquad +2\times e^{-\lambda}\lambda^2/2! + 3\times e^{-\lambda}\lambda^3/3! \ldots$

$\qquad = \lambda e^{-\lambda} (1+\lambda+\lambda^2/2!+\lambda^3/3! \ldots)$

$E(X) = \lambda$ (ie the mean)

$V_{ar}(X) = E(X^2) - E^2(X)$

It can be shown $E(X^2) = \lambda + \lambda^2$

$V_{ar}(X) = \lambda + \lambda^2 - \lambda^2 = \lambda$

(ie both mean and variance are λ)

The Poisson distribution approximates to the Binomial distribution under certain conditions

If $X \sim B_{in}(n,p)$

$P(X=x) = {}^nC_x p^x q^{n-x}$

Now $p = {}^{\lambda}/_{n}$ $q = 1 - {}^{\lambda}/_{n}$

so $P(X=x) = {}^nC_x({}^{\lambda}/_{n})^x(1- {}^{\lambda}/_{n})^{n-x}$

$\qquad = \{{}^nC_x/n^x\}\, \lambda^x\, (1- {}^{\lambda}/_{n})^{n-x}$

which we wish to equate to ${}^{1}/_{x!}\, \lambda^x e^{-\lambda}$

For large n $(1- {}^{\lambda}/_{n})^{n-x} \approx e^{-\lambda}$

and ${}^nC_x / n^x \approx {}^{1}/_{x!}$

and hence the approximation is valid

Examples for $n = 100$ $p = 0.01$ $\lambda = 1$

x	$P(X=x) \sim B_{in}(n,p)$	$P(X=x) \sim P_o(\lambda)$
0	0.3660	0.3697
1	0.3697	0.3697
2	0.1849	0.1839
3	0.0160	0.0613
4	0.0149	0.0153

which is close agreement

Examples for $n = 50$ $p = 0.1$ $\lambda = 5$

x	$P(X=x) \sim B_{in}(n,p)$	$P(X=x) \sim P_o(\lambda)$
0	0.00515	0.00674
1	0.0286	0.0337
2	0.0779	0.0842
3	0.139	0.140
4	0.181	0.175

which isn't quite so close

One must also question whether all this is quite so relevant with modern computing power.

Mode of Poisson Distribution

The mode is the value most likely to occur, that is the one with the highest probability. If λ is an integer the distribution is bimodal with modes at $x = \lambda - 1$ and $x = 1$

This can readily be seen from Poisson tables.

If λ is not an integer then the mode m occurs at $\lambda - 1 < m < 1$

Two Independent Variables

If $X \sim P_o(m)$

and $Y = P_o(n)$

then $X + Y \sim P_o(m+n)$

Probability Distributions

A continuous random variable is a theoretical representation of a continuous variable such as height, mass or time.

A continuous random variable is specified by its probability density function f(x).

The pdf is represented by a "curve" and the probabilities are the areas under the curve.

We are therefore concerned with some particular range and we discount any difference between say $\leq$ and $<$.

$\int_{\text{all } x} f(x)\,dx = 1$

$P(x_1 \leq X \leq x_2) = \int_{x1}^{x2} f(x)\,dx$

Mean

$E(X) = \int_{\text{all } x} f(x)\,dx = \mu.$

If f(x) is symmetrical then μ will be the x-value on the line of symmetry.

$E[g(x)] = \int_{\text{all } x} g(x)\,f(x)\,dx$

$E[X^2] = \int_{\text{all } x} x^2\,f(x)\,dx$

Cor.1 $E(a) = a$

Cor.2 $E(aX) = a\,E(x)$

Cor.3 $E(aX+b) = aE(x) + b$

Cor.4 $E[f_1(X)+f_2(X)]$

$= E[f_1(x)] + E[f_2(x)]$

Variance

If X is a continuous random variable with probability density function f(x) then

$V_{ar}(X) = \int_{\text{all } x} x^2\,f(x)\,dx - \mu^2$

where $\mu = E(X) = \int_{\text{all } x} x\,f(x)\,dx$

Standard Deviation

$\sigma = \sqrt{\{V_{ar}(X)\}}$

Cor.1 $V_{ar}(a) = 0$

Cor.2 $V_{ar}(aX) = a^2V_{ar}(X)$

Cor.3 $V_{ar}(aX+b) = a^2V_{ar}(x)$

Mode

The mode is the value of X for which f(x) is a maximum.

Cumulative Distribution Function

If X is a continuous random variable with probability density function f(x) defined for $a \leq x \leq b$

then the cumulative distribution function is given by

$F(t) = P(X \leq t) = \int_a^t f(x)\,dx$

a and /or b may be $-\infty$ to $+\infty$

$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$

Median

The median splits the curve into two equal area halves

so $\int_a^m f(x)\,dx = 0.5$ or simply $F(m) = 0.5$

Relationship pdf to cdf

$f(x) = {}^{d}/_{dx}\, F(x) = F'(x)$

ie the gradient of F(x) gives the value f(x)

Rectangular Distribution

The distribution has constant height ${}^{1}/_{b-a}$ between values a and b giving an area of 1.

So $f(x) = {}^{1}/_{b-a}$ for $a \le x \le b$

$X \sim R(a,b)$

Hence ${}_{\text{all } x}\int f(x)\,dx = {}_{a}\int^{b} {}^{1}/_{b-a}\,dx$

If $X \sim R(a,b)$ then

$E(X) = \frac{1}{2}(a+b)$

$Var(X) = {}^{1}/_{12}(b-a)^2$ deduced earlier

$E(X^2) = {}^{1}/_{3}(b^2 + ab + a^2)$

Exponential Distribution

A continuous random variable X having probability density function f(x) where

$f(x) = \lambda e^{-\lambda x}$

is said to be an exponential distribution.

${}_{\text{all } x}\int f(x)\,dx = {}_{0}\int^{\infty} \lambda e^{-\lambda x}\,dx$

$= -[e^{-\lambda x}]_0^{\infty} = e^{-\infty} + e^0 = 1$

$P(X < a) = {}_{0}\int^{a} \lambda e^{-\lambda x}\,dx$

$= -[e^{-\lambda x}]_0^{a} = -e^{-\lambda x} + e^0$

Hence $P(X < a) = 1 - e^{-\lambda a}$

and $P(X > a) = e^{-\lambda a}$

$F(x) = P(X \le x) = 1 - e^{-\lambda x}$ for $x \ge 0$

Also $P((X > a+b) \mid (X > b)) = P(X > b)$

$E(X) = {}^{1}/_{\lambda}$

$Var(X) = {}^{1}/_{\lambda^2}$

$E(X^2) = {}^{2}/_{\lambda^2}$

The waiting times between successive events in a Poisson distribution follow an exponential distribution.

Median

Let the median lie at $x = m$

$F(m) = 0.5 = 1 - e^{-\lambda m}$

so $e^{\lambda m} = 2$ and hence $m = {}^{\ln 2}/_{\lambda}$

The Normal Distribution

A continuous variable X having pdf f(x) where

$f(x) = {}^{1}/_{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$

has a normal distribution and we write

$X \sim N(\mu, \sigma^2)$

$E(X) = \mu$

$Var(X) = \sigma^2$

Proof

$E(X) = {}_{\text{all } x}\int x\, f(x)\,dx$

$= {}^{1}/_{\sigma\sqrt{2\pi}}\ {}_{-\infty}\int^{+\infty} x e^{-(x-\mu)^2/2\sigma^2}\,dx$

let $t = {}^{x-\mu}/_{s}$

$x = t\sigma + \mu$

${}^{dx}/_{dt} = \sigma$

$E(X) = = {}^{1}/_{\sigma\sqrt{2\pi}}\ {}_{-\infty}\int^{+\infty} (\mu + \sigma t) e^{-\frac{1}{2}t^2} \sigma\,dt$

$= {}^{\mu}/_{\sqrt{(2\pi)}}\ {}_{-\infty}\int^{\infty} e^{-\frac{1}{2}t^2}\,dt + {}^{\sigma}/_{\sqrt{(2\pi)}}\ {}_{-\infty}\int^{\infty} t e^{-\frac{1}{2}t^2}\,dt$

$= \mu + {}^{\sigma}/_{2\pi}\,[e^{-\frac{1}{2}t^2}]_{-\infty}^{\infty}$

Here I assume ${}^{1}/_{\sqrt{(2\pi)}}\ {}_{-\infty}\int^{\infty} e^{-\frac{1}{2}t^2}\,dt = 1$

$E(X) = \mu$

A similar extended proof will show

$V_{ar}(X) = \sigma^2$

If $X \sim N(\mu,\sigma^2)$ the maximum value $f(x)$ occurs at $x = \mu$ and $f(x)$ has points of inflexion at $x = \mu \pm \sigma$

Standardised Normal Variable Z

$Z \sim N(0, 1)$

Tables give cumulative probabilities – that is areas under the curve.

The symbol for cumulative probability is $\Phi(z)$ *(pronounced phi)*

so $\Phi(z) = P(Z < z)$

if $X \sim N(\mu, \sigma^2)$

and $Z = {}^{X-\mu}/_{\sigma}$

then $Z \sim N(0,1)$

So we have standardised the variable so we only need one universal table.

If $Z = {}^{(X-\mu)}/_{\sigma}$

then $X = \mu + \sigma Z$

Approximation Normal to Binomial

if $X \sim B_{in}(n,p)$ with $n \geq 50$ and $p \approx 0.5$

then $X \sim N(np, npq)$

The greater n the more flexibility on p.

Approximation Normal to Poisson

if $X \sim P_o(\lambda)$

then $E(X) = \lambda$

$V_{ar}(X) = \lambda$

for $\lambda > 20$ $\quad X \sim N(\lambda, \lambda)$

Random Variables and Sampling

Let X and Y be two independent normal variables.

$X \sim N(\mu_1, \sigma_1^2)$

$Y \sim N(\mu_2, \sigma_2^2)$

then

$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$X - Y = N(\mu_1 - \mu_2, \sigma_1^2+\sigma_2^2)$

and this may be extended to any number of normal independent variables.

Now take the special case where $X_1, X_2, X_3 \ldots X_n$ are independent observations from same normal distribution

$X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, 3, \ldots n$ then

$X_1 + X_2 + X_3 + \ldots X_n \sim N(n\mu, n\sigma^2)$.

Remembering if X is a normal variable $X \sim N(\mu, \sigma^2)$ then

$aX \sim N(a\mu, a^2\sigma^2)$

and it therefore must follow that

$aX \pm bY \sim N(a\mu_1 \pm b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

The Sample Mean

Let $X_1, X_2, X_3, \ldots Xn$ be a random <u>sample</u> of n independent observations from a population mean μ variance σ^2.

Consider the sample mean $\bar{X}$

$\bar{X} = {}^1/_n(X_1 + X_2 + X_3 + \ldots X_n)$

$= {}^1/_n \sum X_i \qquad i = 1, 2, 3, \ldots n$

$\sum(\bar{X}) = E[\,{}^1/_n(X_1 + X_2 + X_3 + \ldots X_n)]$

$= {}^1/_n\,[E(X_1) + E(X_2) \ldots E(X_n)]$

$= {}^1/_n\,(n\mu)$

$= \mu$

$Var(\bar{X}) = Var[\,{}^1/_n(X_1 + X_2 + X_3 + \ldots X_n)]$

$= {}^1/_{n^2}\,[Var(X_1)+Var(X_2) \ldots Var(X_n)]$

$= {}^1/_{n^2}\,(n\sigma^2)$

$Var\,(\bar{X}) = {}^{\sigma^2}/_n$

If we sample without replacement

$E(\bar{X}) = \mu$

$Var\,(\bar{X}) = {}^{\sigma^2}/_n\,({}^{N-n}/_{N-1})$

To summarise if we take a random sample of n items from a normal distribution then

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

This is a key result which has even wider implications because under the Central Limit Theorem the distribution can be of <u>any form</u> and still the standard deviation of $\bar{X}$ is ${}^{\sigma}/_{\sqrt{n}}$ and is termed the standard error.

Distribution of the Sample Proportion

Consider a population in which the proportion of successes is p. If the random sample size is n and X is the random variable “number of successes” then $X \sim Bin(n,p)$.

If n is large (say $n \geq 50$)

$X \sim N\,(np, npq)$

Let P_s be the proportion of successes

$P_s = {}^X/_n$

$E(P_s) = E\,({}^X/_n)$

$= {}^1/_n\,E(X)$

$= {}^1/_n\,(np)$

$=p$

$Var(P_s) = Var\,({}^X/_n)$

$= {}^1/_{n^2} Var(X)$

$= {}^1/_{n^2}\,(npq)$

$= {}^{pq}/_n$

Therefore $P_s \sim N\,(p, {}^{pq}/_n)$

Estimation Population Parameters

Distribution	**Parameter**
discrete	
Binomial	n and p
Poisson	λ
Geometric	P
continuous	
Rectangular	a and b
Exponential	λ
Normal	μ and σ^2

If the parameters are unknown we take a sample from the population and use that to make estimates.

This statistic is called an ESTIMATOR

(U,T etc).

The numerical value in any particular instance is called an ESTIMATE (u,t).

Consider a population with unknown parameter θ

If U is some statistic derived from a random sample taken from the popuation, then U is an unbiased estimator for θ if

$E(U) = \theta$

The most efficient estimator is the one which is unbiased and has the smallest variance.

The three important population parameters are mean, variance and proportion ("successes")

Parameter	**Symbol**	**Estimator**
mean	μ	$\mu^\wedge$
variance	s^2	$\sigma^{\wedge 2}$
proportion	p	$p^\wedge$

Parameter	**Estimator**	**Estimate**
mean	$\bar{X}$	$\bar{x}$
variance	S^2	s^2
proportion	P_s	p_s

For mean $\mu^\wedge = \bar{X}$

as $E(\bar{X}) = \mu$ the estimate is unbiased.

For variance

$\sigma^{\wedge 2} = {}^{nS^2}/_{n-1} = {}^{1}/_{n-1} \sum(X - \bar{X})^2$

For proportion

$p^\wedge = P_s$

as $E(P_s) = p$ and estimate is unbiased.

Pooled Estimators

	Size	Mean	Var.
sample 1	n_1	$\bar{X}_1$	S_1^2
sample 2	n_2	$\bar{X}_2$	S_2^2

$\mu^\wedge = \{n_1 \bar{X}_1 + n_2 \bar{X}_2\} / \{n_1 + n_2\}$

This calculation is exact for any set of figures representing sample 1 and sample 2.

$\sigma^{\wedge 2} = \{n_1 S_1^2 + n_2 S_2^2\} / \{n_1 + n_2 - 2\}$

This calculation however is dependent on the assumption that the samples are drawn from the same parent population.

	Size	Proportion
sample 1	n_1	P_{s1}
sample 2	n_2	P_{s2}

$p^\wedge = \{n_1 P_{s1} + n_2 P_{s2}\} / \{n_1 + n_2\}$

	one sample	two sample
$\mu^\wedge$	$\bar{X}$	$\{n_1 \bar{X}_1 + n_2 \bar{X}_2\} / \{n_1 + n_2\}$
$s^{\wedge 2}$	${}^{n}/_{n-1} S^2$	$\{n_1 S_1^2 + n_2 S_2^2\}/\{n_1 + n_2 - 2\}$
$p^\wedge$	P_s	$\{n_1 P_1^2 + n_2 P_2^2\}/\{n_1 + n_2\}$

Interval Estimation

An interval estimation of an unknown population parameter is an interval constructed so that it has a given probability of including the parameter

eg $P(a < \theta < b) = 0.95$

is termed

the 95% confidence interval for θ.

What it is <u>not</u> is the probability that θ lies in the interval because it is θ that is fixed – albeit unknown – and the interval that varies.

There are three cases to consider

Case 1

Find the confidence interval for μ when the variance σ^2 is known.

If $X \sim N(n, \sigma^2)$

then for any n

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

If we do not know if X follows a normal distribution then we need $n > 30$. We then standardise

$Z = \{\bar{X} - \mu\} / \{{}^{\sigma}/_{\sqrt{n}}\}$

so $Z \sim N(0,1)$

For 95% confidence we set 1.96.

So $P(\bar{X} - 1.96\ {}^{s}/_{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\ {}^{s}/_{\sqrt{n}}$

or in brief we write

95% confidence interval is $\bar{x} \pm 1.96\ {}^{\sigma}/_{\sqrt{n}}$

Case 2

Find the confidence interval for μ when the variance σ^2 is unknown.

So first we have to generate an estimate for $\hat{\sigma}^2$

$\hat{\sigma}^2 \quad = {}^{n}/_{n-1}\ S^2$ *(sample variance)*

$= \sum^{(x-\bar{x})^2} / {}_{(n-1)}$

95% confidence interval is

$\bar{x} \pm 1.96\ \hat{\sigma} / \sqrt{n}$

Case 3

Confidence interval for p is

$P_s \pm 1.96 \sqrt{(p_s q_s/n)}$

so we don't have any prior knowledge of the parent population, the trade off being we require n large – say $n > 30$.

Significance Testing

The relevant standard deviations for given confidence limits are

	One Tailed	Two Tailed
90%	1.28	1.64
95%	1.64	1.96
99%	2.33	2.58
99.5%	2.58	2.91
1 sd includes	84.1%	68.3%
2 sd includes	97.7%	95.5%
3 sd includes	99.9%	99.7%

Cumulative Distribution Function $\Phi(z) - \Phi(0) = {}^{\mu}/_{\sqrt{(2\pi)}}\ {}_{0}\!\int^{z} e^{-\frac{1}{2}t^2}\,dt$ (*ie one tailed*)

Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.30	0.1179	0.11217	0.11255	0.11293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.365	0.3389
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.30	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.40	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.50	0.4332	0.4354	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.60	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.70	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.80	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.90	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.10	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.20	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.30	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.40	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.50	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.60	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.70	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.80	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.90	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.10	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.20	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.30	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.40	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.50	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998

Back Page