Mr G's Little Book on

Higher Level Statistics

Issue 2.0

Contents

Notation

a	gradient regression line y on x
c	gradient regression line x on y
m_i	residual of plot (x_i , y_i)
r	product moment correlation coefficient
R^2	coefficient of determination
s^2	variance of sample
$s^{\wedge 2}$	most efficient estimator of variance
s_{xy}	covariance of x and y
s_{xx}	sample variance of x
s_x	sample standard deviation of x
S_{xx}	sum of squares of diff. x and x_{bar}
S_{xy}	product of sum squares x and y
x_{bar}	arithmetic mean of x *(The author apologises for this nomenclature)*
μ	mean of parent population
σ	standard deviation parent population
σ^2	variance of parent population

Discrete Random Variables

S1.1 Introduction

If $P(X = x_1) = p_1$ If $P(X = x_2) = p_2$ …
If $P(X = x_n) = p_n$ etc.
then X is a discrete random variable if
$p_1 + p_2 \ldots p_n = 1$
written $\sum_{all\ x} P(X = x_i) = 1$
Random variables are denoted by
X, Y, R and particular values by x y r.
The function which allocates the probabilities is termed the probability density function (pdf) and the results the probability distribution (pd).

S1.2 Expectation E(X)

$E(X) = \sum_{all\ x} x\ P(X = x)$ or simply $\sum x_i\ p_i$.
$E(X) = \mu$ and also $E(x_{bar}) = \mu$
x_{bar} is a statistic whereas
μ is a parameter
X is the variable arising from a single observation, a sample of one from some population.
x_{bar} is the mean of several observations - a sample of n in a population.
Some pdf's have the important property of symmetry about a central value
E(X) is the midpoint value.
This definition can be extended to any function of the random variable.
If g(X) is any function of the discrete random variable
$E[\ g(X)\ ] = \sum_{all\ x} g(x)\ P(X = x)$

Corollaries

$E(a) = a$
$E(aX) = aE(x)$
$E(aX + b) = aE(X) + b$
$E[\ f_1(X) + f_2(X)\ ] = E[f_1(X)] + E[f_2(X)]$
$E(X + Y) = E(X) + E(Y)$
X and Y do not need to be independent.

S1.3 Variance

Experimental Approach

For a frequency distribution with mean x_{bar} and variance s^2.
$s^2 = \sum f(x - x_{bar})^2 / \sum f$
which can be simplified to
$\sum fx^2 / \sum f - x_{bar}^2$

Theoretical Approach

For discrete random variable X with
$E(X) = \mu$
the variance Var(X)
is defined as
$Var(X) = \sum(x_i - \mu)^2\ P(X = x_i)$
which can also be written
$Var(X) = \sum(x_i^2)P(X = x_i) - \mu^2$
Note as $\mu^2 = E[(X)]^2$
$Var(X) = E(X^2) - E^2[(X)]$

Corollaries

$Var(a) = 0$
$Var(aX) = a^2\ Var(X)$
$Var(ax + b) = a^2\ Var(X)$

S1.4 Cumulative Distribution Function

If X is a discrete random variable with probability density function
$P(X = x)$ for $x = x_1, x_2, \ldots x_n$
then the cumulative distribution function is given by
$F(t) = P(X \leq t) = \sum P(X = x)$
where $t = x_1, x_2, \ldots x_n$

S1.5 2 Independent Random Variables

If X and Y are any two random variables then $E(X + Y) = E(X) + E(Y)$

If X and Y are any two independent random variables then

$Var(X + Y) = Var(X) + Var(Y)$

For random variables X and Y

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

If X and Y are independent, then

$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$

If a = 1 and b = –1 then

$E(X – Y) = E(X) – E(Y)$

$Var(X – Y) = Var(X) + (–1)^2Var(Y)$

$= Var(X) + Var(Y)$

$Var(aX – bY) = a^2Var(X) + b^2 Var(Y)$

S1.6 The Distribution of $X_1 + X_2$

X_1 and X_2 are independent observations <u>from the same distribution</u> X

$E(X_1 + X_2) = 2 E(X)$

$Var(X_1 + X_2) = 2 Var(X)$

For n independent observations

$E(X_1 + X_2 + \dots X_n) = n E(X)$

$Var(X_1 + X_2 + \dots X_n) = n Var(X)$

Note the distributions of 2X and X_1 and X_2 are very different.

Multiples	**Sums**
$E(2X) = 2 E(X)$	$E(X_1 + X_2) = 2 E(X)$
$Var(2X)=4Var(X)$	$Var(X_1+X_2)= 2Var(X)$

In general

$E(nX) = n E(X)$

$E(X_1+ X_2+ \dots X_n) = n E(X)$

$Var(nX) = n^2 Var(X)$

$Var(X_1+ X_2 + \dots X_n) = n Var(X)$

Discrete Probability Distributions

S2.1 The Binomial pdf.

If successful outcome is p and failure is q = 1 – p and if X is the random variable of the number of successful outcomes in n independent trails, then the pdf of X is

$P(X=x) = {}^nC_x\, q^{n-x}\, p^x \quad x = 0, 1, 2, \dots n$

$(q + p)^n$

$= {}^nC_0q^np^0 + {}^nC_1q^{n-1}p^1 + {}^nC_2q^{n-2}p^2 + \dots {}^nC_nq^0p^n$

$= P(X=0) + P(X=1) + P(X=2) + \dots P(X = n)$

$X \sim Bin(n , p)$

$E(X)$ (mean) $= \sum xP(X – x) = np$

$Var(X) = E(X^2) – E^2(X) = npq$

which are easily proved.

In general

$P(X = r \mid X \sim Bin(n , p)) = P(X = n–r \mid X \sim Bin(n , p))$

$P(X \geq r \mid X \sim Bin(n , p)) = P(X \leq n–r \mid X \sim Bin(n , p))$

$P(X \leq r \mid X \sim Bin(n , p)) = P(X \geq n – r \mid X \sim Bin(n , p))$

ie the terms are mirror reflections.

For cumulative probabilities when tables are not available successive terms can be calculated from

$p_{x+1} = {}^{(n-x)p} / {}_{(x+1)q}\, p_x$

where $p_{x+1} = P(X = x + 1)$ and

$p_x = P(X = x)$

Probability Generating Function

$(1 – p + pt)^n$

S2.2 The Geometric pdf

In the binomial distribution, the number of trials is ***fixed***, and we count the number of **"*successes*"**.
In the geometric distribution, the number of **"*successes*"** is fixed, and we count the number of ***trials*** needed to obtain the desired number of "successes".

$P(X = x) = q^{x-1}p$ where $q = 1 - p$
is given as $X \sim Geo(p)$

$E(X) = 1/p$ and $Var(X) = q / p^2$

$P(X > r) = q^r$

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

Probability Generating Function
$pt / \{1 - (1 - p)t\}$

S2.3 The Poisson pdf

A Poisson distribution is a discrete probability distribution giving the probability of an event happening a certain number of times (k) within a given interval of time or space.
It has only one parameter, λ,
the mean number of events.

$P(X = x) = e^{-\lambda} \lambda^x / x!$ and $X \sim Po(\lambda)$
where the parameter λ can be any positive value.
We know $e^{\lambda} = 1 = 1 + \lambda^2/2! + \ldots$
so clearly $\sum P_i = 1$ so we have a valid pdf

$E(X)$ (mean) $= \lambda$ and $Var(X) = \lambda$

Mode
The mode is the value most likely to occur, that is the one with the highest probability.
If λ is an integer there are two modes at $x = \lambda - 1$ and $x = \lambda$.
This can readily be seen from tables.

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

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

Probability Generating Function

$$e^{\lambda(t-1)}$$

S2.4 Poisson approx. Binomial

A binomial distribution with parameters n and p can be approximated by a Poisson distribution,
with parameter $\lambda = np$, if
n is large say $n > 50$ and
p small say $p < 0.1$.

The approximation improves as
$n \to \infty$ and $p \to 0$.

For cumulative probabilities the following recurrence formula can be used
$P(X = x + 1) = \lambda P(X = x) / (x + 1)$

S2.5 The Distribution of $X_1 + X_2$

The sum of two independent Poisson variables with parameters m and n is a Poisson variable with parameters (m + n).

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

Continuous Random Variables

S3.1 Introduction

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

It is specified by its probability density function written f(x).

If X is a continuous random variable with pdf f(x) valid over the range $a \le x \le b$ then ${}_{\text{all } x}\int f(x)\,dx = 1$

The probability $P(c \le X \le d) = {}_{c}\int^{d} f(x)\,dx$ that is the area under the curve between c and d.

As we are concerned with a particular range we discount any difference between say $\le$ and $<$.

$E(X) = {}_{\text{all } x}\int x\,f(x)\,dx$ which is denoted μ and termed the mean of X.

If g(x) is any function of the continuous randon variable X having pdf f(x) then

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

and in particular $E(X^2) = {}_{\text{all } x}\int x^2\,f(x)\,dx$

Corollaries

$E(a) = a$

$E(aX) = aE(x)$

$E(aX + b) = aE(X) + b$

$E[\,f_1(X) + f_2(X)\,] = E[f_1(X)] + E[f_2(X)]$ as before.

Similarly $\text{Var}(X) = {}_{\text{all } x}\int (x - \mu)^2 x\,f(x)\,dx$

$= {}_{\text{all } x}\int (x^2\,f(x)\,dx - \mu^2$

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

The standard deviation $\sigma = \sqrt{\{\text{Var}(X)\}}$

Corollaries

$\text{Var}(a) = 0$

$\text{Var}(aX) = a^2\,\text{Var}(X)$

$\text{Var}(ax + b) = a^2\,\text{Var}(X)$ as before.

Mode

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

S3.2 Cumulative Distribution Function

If X is a continuous random variable with pdf f(x) defined for $a \le x \le b$ then the cumulative distribution function is given by $F(x_0) = P(X \le x_0) = {}_{a}\int^{x0} f(t)\,d$

and $P(x_1 \le X \le x_2) = F(x_2) - F(x_1)$

a is often $= -\infty$

The median splits the curve into two equal haves and ${}_{a}\int^{m} f(x)\,dx = 0.5$ i.e. $F(m) = 0.5$

The probability density function can be obtained from the cumulative distribution as follows.

Now $F(t) = {}_{a}\int^{t} f(x)\,dx$ for $a \le t \le b$

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

The gradient of the F(x) curve gives the value of f(x)

S3.3 The Rectangular Distribution

A continuous random variable X having pdf f(x) where $f(x) = {}^{1}/_{b-a}$ for $a \le x \le b$ where a and b are constants is said to follow a rectangular or uniform distribution.

We write $X \sim R(a\,,\,b)$

The probability of a variable lying in one particular range is exactly the same as the probability that it lies in another range of the same length.

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

$Var(X) = \frac{1}{12}(b - a)^2$

The '12' appears because we integrate x^2 the get $\frac{1}{3}$ and square the Expectation to get ¼ and the product give $\frac{1}{12}$

cf The variance of the first n integers can be shown to be $\frac{1}{12}(n^2 - 1)$.

$E(X^2) = (b^2 + ab + a^2) / 3$

S3.4 The Exponential Distribution

The continuous r.v. X having pdf f(x)

where $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$

where λ is a positive constant is said to follow an exponential distribution.

$$\int_{\text{all } x} f(x)\,dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = -[e^{-\lambda x}]_0^{\infty}$$

$$= -e^{-\infty} + e^0 = 1$$

$E(X)$ (the mean) $= \frac{1}{\lambda}$

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

$E(X^2) = 2 / \lambda^2$

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

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

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

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

The exponential distribution can be regarded as the waiting time between events following a Poisson distribution

$m(X) = \frac{\ln 2}{\lambda} < E(X)$

nb ln 2 = – ln (½)

The Normal Distribution

S4.1 Introduction

A continuous r.v. X having pdf f(x)

where $f(x) = \left(\frac{1}{\sigma\sqrt{(2\pi)}}\right) e^{\wedge} \frac{1}{2} \left(\frac{-(x-\mu)^2}{\sigma^2}\right)$

is said to follow a normal distribution

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

$E(X) = \mu$ and $Var(X) = \sigma^2$

The maximum value of f(x) occurs at

$x = \mu$

The points of inflexion occur at

$x = \mu + \sigma$ and $x = \mu - \sigma$

To standardise X, subtract μ and $\div \sigma$.

$Z = \frac{X - \mu}{\sigma}$ and $Z \sim N(0, 1)$

and rearranging $X = \mu + \sigma Z$

The pdf of the standard normal variable Z is denoted by $\phi(z)$ where

$\phi(z) = \left(\frac{1}{\sqrt{(2\pi)}}\right) e^{-\frac{1}{2}z^2}$

The cumulative distribution function

$$\Phi(z) = \int_{-\infty}^{z} \left(\frac{1}{\sqrt{(2\pi)}}\right) e^{-\frac{1}{2}z^2} dz$$

$$= P(Z < z)$$

This integral is very difficult to evaluate so we refer to tables

The central 95% of the distribution lies between $\pm$ 1.96

The central 99% of the distribution lies between $\pm$ 2.575

Under certain circumstances the Normal distribution can be used as an approximation to the Binomial distribution.

If $X \sim$ Bin (n,p) then

$E(X) = np$ and $Var(X) = npq$

For large n and p not too small or large

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

The Normal distribution can also be an approximation to the Binomial.

The larger n the more flexibility on p.

If $X \sim P_o(\lambda)$ then
$E(X) = \lambda$ and $Var(X) = \lambda$
For large λ $X \sim N(\lambda , \lambda)$ approximately

Continuity Corrections

As we are using a continuous distribution to approximate a discrete variable we make the continuity correction

eg $P(a \leq X \leq b) \rightarrow P(a - ½ \leq X \leq b + ½)$

S4.2 Summary of Approximations

$X \sim Bin(n,p)$ approximated by $P_o (np)$
if $n > 50$ and $p < 0.1$

$X \sim Bin(n,p)$ approximated by
$N (np , npq)$

If $n > 50$ and $p \cong 0.5$
$X \sim P_o(\lambda)$ approximated by $N (\lambda , \lambda)$
if $\lambda > 20$

R.V.s and Random Sampling

S5.1 Two Independent Variables

If X and Y are two independent normal variables such that
$X \sim N(\mu_1 , \sigma_1^2)$ and $Y \sim N(\mu_2 , \sigma_2^2)$
then $X + Y \sim N(\mu_1 + \mu_2 , \sigma_1^2 + \sigma_2^2)$
and $X – Y \sim N(\mu_1 – \mu_2 , \sigma_1^2 + \sigma_2^2)$

This can be expanded to n independent normal variables. In the special case
$X_i = \sim N(\mu , \sigma^2)$ for i = 1, 2, … n then
$X_1 + X_2 + … X_n \sim N (n\mu , n\sigma^2)$

We also have
if $X \sim N(\mu , \sigma^2)$ then $aX \sim N(a\mu , a^2\sigma^2)$
$aX + bY = N(a\mu_1 + b\mu_2 , a^2\sigma_1^2 + b^2\sigma_2^2)$
$aX – bY = N(a\mu_1 – b\mu_2 , a^2\sigma_1^2 + b^2\sigma_2^2)$

As before be careful to distinguish between
$2X = N(2\mu , 4\sigma^2)$ but
$X_1 + X_2 = N(2\mu , 2\sigma^2)$

In general
$nX = N(n\mu , n^2\sigma^2)$ but
$X_1 + X_2 + … X_n = N(n\mu , n\sigma^2)$

S5.2 The Sample Mean

If $X_1 + X_2 + … X_n$ is a random sample of size n taken from any infinite population (or infinite population if sampling is with replacement) with population mean μ and variance σ^2 then the sample mean X_{bar} is such that
$E(X_{bar}) = \mu$ and $Var(X_{bar}) = \sigma^2 / n$

If the random sample is taken *without replacement* then
$E(X_{bar}) = \mu$ and
$Var(X_{bar}) = {}^{\sigma^2}/_{n} ({}^{N-n}/_{N-1})$

S5.3 Distribution of the Sample Mean

If $X_1 + X_2 + \ldots X_n$ is a random sample of size n taken from a normal distribution with population mean μ and variance σ^2 such that $X \sim N$ then the distribution of X is $X \sim N(\mu , \sigma^2)$.

The distribution of X_{bar} is also normal and

$X_{bar} \sim N(\mu , \sigma^2 / n)$ where

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

The distribution of the sample mean X_{bar} is known as the sampling distribution of means and the standard deviation of this distribution $\sigma / \sqrt{n}$ is known as the **standard error of the mean**.

S5.4 Central Limit Theorem

If $X_1 + X_2 + \ldots X_n$ is a random sample of size n from <u>any</u> distribution with mean μ and variance σ^2 then, <u>for large</u> *n*, the distribution of the sample mean X_{bar} is approximately normal such that

$E(X_{bar}) = \mu$ and $Var(X_{bar}) = \sigma^2 / n$

If the random sample is taken <u>without replacement</u> then

$X_{bar} \sim N(\mu , \sigma^2 / n)$ where

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

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, <u>for large</u> *n*, the distribution of the sum of the random variables is approximately normal with mean $n\mu$ and variance $n\sigma^2$

S5.5 Distribution of Sample Proportion

Consider a binomial population in which p is the proportion of successes then the probability of success in any trial is p.
If a random sample of size n is taken from this population and X is the random variable - the 'number of successes' then

$X \sim Bin(n , p)$ and for large n

$X \sim N (np , npq)$

Now if P_s is the random variable 'the proportion of successes in a sample' then

$P_s \sim N(p , pq/n)$

This is known as the **sampling distribution of proportions** and the standard deviation of the sampling distribution $\sqrt{(pq / n)}$ is known as the **standard error of proportion**.

Continuity Correction

When considering the Normal approximation to the Binomial distribution, a continuity correction of $\pm$ ½ is used. Now since Ps = X/n, we use a continuity correction of $\pm$ ½ $({}^1/_n)$ = ${}^1/_{2n}$.

Estimation of Population Parameters

S6.1 Point Estimation

Consider a population with unknown parameter θ. If U is some statistic derived from a random sample taken from the population, then U is an unbiased estimator for θ if $E(U) = \theta$.

There are many estimators which could be formed but the most efficient is the one which is unbiased and has the smallest variance.

If U is an estimator for an unknown parameter q, then U is a consistent estimator for q if
$Var(U) \to 0$ as $n \to \infty$

From a population with unknown variance σ^2, take a random sample of size n and let
$s^2 = \sum(X_i - X_{bar})^2 / n$ where s^2 is the sample variance then the most efficient estimator written $s^{\wedge 2}$ is $^{n} / _{(n-1)}\, s^2$

The author recommends using s_{n-1} for $s^{\wedge}$ to emphasise and clarify the introduction of $n-1$.

For a binomial population in which p is the unknown proportion of successes if a random sample of size n is taken then the unbiased estimator for p is P_s.

S6.2 Pooled Estimators from 2 Samples

From a population with unknown mean μ and unknown variance σ^2 if we take two random samples then
$u^{\wedge} = (n_1 X_{bar1} + n_2 X_{bar2}) / (n_1 + n_2)$
$s^{\wedge 2} = (n_1 S_{1^2} + n_2 S_2^2) / (n_1 + n_2 - 2)$

S6.3 Pop. Proportion Pooled Estimator

$P^{\wedge} = (n_1 P_{s1} + n_2 P_{s2}) / (n_1 + n_2)$

S6.4 Interval Estimation

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

Consider a population with unknown parameter θ. If we can find an interval (a , b) such that
$P(a < \theta < b) = 0.95$ we say that (a , b) is a 95% confidence interval for θ.
It is NOT the probability that θ lies in the interval.

If x_{bar} is the mean of a random sample of size n taken from a normal population with known variance σ^2 , then a central confidence interval for μ, the population mean is given by
$(x_{bar} - 1.96\, \sigma / \sqrt{n}\, ,\, x_{bar} + 1.96\, \sigma / \sqrt{n})$
written as $x_{bar} \pm 1.96\, \sigma / \sqrt{n}$

If the population is not normal then we require n to be large - say $n \geq 30$.

If x_{bar} and s^2 are the mean and variance of a random sample of size n (where n is large) taken from a normal population with unknown mean μ and unknown

variance σ^2 , then a central 95% confidence interval for μ, the population mean is given by

$(x_{bar} - 1.96\ \hat{\sigma} / \sqrt{n}\ ,\ x_{bar} + 1.96\ \hat{\sigma} / \sqrt{n})$

where $\hat{s}^2 = (^n/_{n-1})s^2 \cong s^2$ for large n

written as $x_{bar} \pm 1.96\ \sigma / \sqrt{n}$

See Appendix D for further z values.

S6.5 t-distribution

The random X is said to follow the t-distribution if the pdf of X is

$f(x) = C_v\ (1 + x^2/v)^{-½\ (v-1)}$

where $C_v = \Gamma(½\ (v + 1)) / \sqrt{(\pi v)}\ \Gamma(½\ v)$

and v = number of degrees of freedom.

v represents the number of variables – the number of restrictions involved in calculating the statistics.

$X \sim t(v)$

We use printed tables as this is a complicated function.

If $T = (X_{bar} - \mu) / (S / \sqrt{(n - 1)})$

then T follows a t-distribution with (n – 1) degrees of freedom.

$T \sim t(n - 1)$

If x_{bar} and s^2 are the mean and variance of a random sample of size n where n is *small* taken from a normal population with unknown mean μ and unknown variance σ^2 , then a central 95% confidence interval for μ, the population mean is given by

$(x_{bar} - t\ s/\sqrt{(n - 1)}\ ,\ x_{bar} + t\ s/\sqrt{(n - 1)})$

$P(X_{bar} - t\ S/\sqrt{(n - 1)} \le \mu \le X_{bar} + t\ S/\sqrt{(n - 1)} = 0.95$

i.e. (–t , t) encloses 95% of the t(n – 1) distribution.

S6.6 Prop. Success Confidence Interval

Recalling $P_s \sim N\ (p\ ,\ pq / n)$

it is reasonable to assume that an estimator for pq / n is $P_s\ Q_s / n$ so we have

$P_s \sim N\ (p\ ,\ P_sQ_s / n)$

If in a random sample of size n where $n \ge 30$ the proportion with a particular property is p_s, the 95% confidence interval for the population proportion p is given by

$p_s - 1.96\ \sqrt{(p_sq_s / n)}\ ,\ p_s + 1.96\ \sqrt{(p_sq_s / n)}$

or simply $p_s \pm 1.96\ \sqrt{(p_sq_s / n)}$

S6.7 Summary of Confidence Intervals

For 95% confidence interval

Population mean μ variance σ^2 known

$x_{bar} \pm 1.96\ \sigma / \sqrt{n}$

Population mean μ variance σ^2 unknown n large

$x_{bar} \pm 1.96\ \hat{\sigma} / \sqrt{n}$ where $\hat{\sigma}^2 = {}^n/_{n-1}\ s^2$

Population mean μ variance σ^2 unknown n small

$x_{bar} \pm t\ s/\sqrt{(n - 1)}$ where (–t , t) encloses 95% of the t(n – 1) distribution.

Population proportion p for n large

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

Author's note

The author recommends using s instead of $\hat{\sigma}$

Significance Testing

S7.1 Testing a single sample value

The methodology is

- State H_0 and H_1
- Decide if one-tailed or two-tailed
- Consider the distribution given by H_0
- Decide on the level of the test.
- Decide on rejection criteria
- Calculate the value of the test statistic
- Make conclusion

S7.2 Testing a Mean

We may wish to investigate whether a random sample of size n with mean x_{bar} could have been drawn from a normal population with mean μ.

Case 1

Population variance σ^2 known

Use the test statistic $Z = X_{bar} - \mu / (\sigma/\sqrt{n})$

which is distributed as N(0 , 1) under the null hypothesis H_0 that the true population mean is μ.

Case 2

Population variance σ^2 unknown sample size n large – say $n \geq 30$

We have the sampling distribution of means where $X_{bar} \sim N(\mu , s^2/n)$

Now since s^2 is unknown we use an estimator $s\hat{}^2$ for it. We use the test statistic

$Z = (X_{bar} - \mu) / (\sigma\hat{} / \sqrt{n})$ which is distributed as N(0 , 1) under the null hypothesis H_0 that the true population mean is μ with $\sigma\hat{} \cong S$

Case 3

Population variance σ^2

unknown sample size $n < 30$

As $s\hat{} / \sqrt{n} = \sqrt{n}\, S / \sqrt{n}\sqrt{(n-1)} = S / \sqrt{(n-1)}$

the test statistic becomes

$$Z = (X_{bar} - \mu) / (S / \sqrt{(n-1)}$$

and the distribution is t(n – 1) under the null hypothesis H_0 that the true population mean is μ.

S7.3 Testing difference between means

Consider two unpaired independent samples of sizes n_1 and n_2 such that $X_1 = N(\mu_1 , \sigma_1^2)$ $X_2 = N(\mu_2 , \sigma_1^2)$ then $X_{bar1} - X_{bar2} \sim N(\mu_1 - \mu_2 , \sigma_1^2/n_1 + \sigma_2^2 / n_2)$

This distribution is known as the sampling distribution of the difference between means. The following may be used to test whether there is a significant difference between means.

1) If $\sigma_1^2 + \sigma_2^2$ are known we use the test statistic

$Z = X_{bar1} - X_{bar2}$

$\sim (\mu_1 - \mu_2 / \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$

which is distributed as N(0 , 1)

2) If there is a known common population variance such that $\sigma_1^2 = \sigma_2^2 = \sigma$ then

$X_{bar1} - X_{bar2} \sim N(\mu_1 - \mu_2 , \sigma^2({}^1/_{n1} + {}^1/_{n2}))$

and we use the test statistic

$Z =$

$\{X_{bar1} - X_{bar2} - (\mu_1 - \mu_2)\} / \{\sigma\sqrt{({}^1/_{n1} + {}^1/_{n2})}\}$

3) If there is a unknown common population variance σ^2 then we use an estimate $\hat{\sigma}^2$ for it where

$\hat{\sigma}^2 = (n_1 s_1^2 + n_2 s_2^2) / (n_1 + n_2 - 2)$

where s_1^2 $s_2^2$are the sample variances.

For large samples we use the test statistic

$Z = \{X_{bar1} - X_{bar2} \sim (\mu_1 - \mu_2)\} / \{\hat{\sigma}\sqrt{(1/n1 + 1/n2)}\}$

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

For small samples we use the test statistic

$T = \{X_{bar1} - X_{bar2} \sim (\mu_1 - \mu_2)\} / \{\hat{\sigma}\sqrt{(1/n1 + 1/n2)}\}$

where $T \sim t (n_1 + n_2 - 2)$

S7.4 Testing a proportion

We may wish to test whether a random sample of size n, with proportion of successes p could have been drawn from a population with proportion of successes p.

The sampling distribution of proportions gives

$Ps \sim N(p , pq/n)$ for n large

The test statistic used is

$Z = (P_s - p) / \sqrt{(p\ q / n)}$

which is distributed as N(0 , 1) under the null hypothesis H_0 that the proportion of successes in the population is p.

S7.5 Testing difference in proportions

Consider two random samples sizes n_1 and n_2 with proportion of successes p_{s1} and $p_{S1.}$ If the population proportions are p_1 and p_2 and the sample sizes are large then

$P_{s1} \sim N(p_1 , p_1q_1/n_1)$

$P_{s2} \sim N(p_2 , p_2q_2/n_2)$ so

$P_{s1} - P_{s2} \sim N(p_1 - p_2 , p_1q_1/n_1 + p_2q_2 / n_2)$

The distribution is known as the sampling distribution of the difference between proportions. We usually wish to test whether the samples have been drawn from populations which have a common proportion p. In this case

$P_{s1} - P_{s2} \sim N(0 , p\ q\ (1/n1 + 1/n2))$

Case 1 If p is known we use the test statistic

$Z = \{P_{s1} - P_{s2} - (0)\} / \sqrt{(pq\ (1/n1 + 1/n2))}$

This is distributed as N(0 , 1) under the null hypothesis H_0 that the common population proportion is p.

Case 2 If p is unknown we use an estimate $\hat{p}$ for it where

$\hat{p} = (n_1p_{s1} + n_2\ p_{s2}) / (n_1 + n_2)$

and we use the test statistic

$Z = \{P_{s1} - P_{s2} - (0)\} / \sqrt{(\hat{p}\hat{q}\ (1/n1 + 1/n2))}$

This is distributed as N(0 , 1) under the null hypothesis H_0.

S7.6 Tests for Binomial n not large

$X \sim Bin(n , p)$. Level of test α%.

Test the single value x = r

1 tailed test

H_0 $p = p_o$ Reject H_0 if $p(X \geq r) < \alpha/100$

H_1 $p > p_o$

H_0 $p = p_o$ Reject H_0 if $p(X \leq r) < \alpha/100$

H_1 $p < p_o$

2 tailed test

H_0 $p = p_o$ Reject H_0 if $p(X \leq r) < ½ \alpha/100$

H_1 $p \neq p_o$ or if $p(X \geq r) < ½ \alpha/100$

S7.7 Tests for Poisson n not large

$X \sim Po(\lambda)$. Level of test α%.

Test the single value $x = r$

1 tailed test

H_0 $\lambda = \lambda_o$ Reject H_0 if $p(X \geq r) < \alpha/100$

H_1 $\lambda > \lambda_o$

H_0 $\lambda = \lambda_o$ Reject H_0 if $p(X \leq r) < \alpha/100$

H_1 $\lambda < \lambda_o$

2 tailed test

H_0 $\lambda = \lambda_o$ Reject H_0 if $p(X \geq r) < ½ \alpha/100$

H_1 $\lambda \neq \lambda_o$ or if $p(X \leq r) < ½ \alpha/100$

Note in this last test inequalities are reversed from 1 tailed test. Is this correct or a misprint?

S7.8 Type I and Type II Errors

A Type I error occurs if we reject H_0 when in fact it is true.

A Type II error occurs if we accept H_0 when in fact it is false.

The χ^2 Chi-Squared Distribution

S8.1 Introduction

Consider the random variable X with probability density function f(x) where

$$f(x) = K_v (½ x)^{½ v - 1} e^{-½ x}$$

X has one parameter v and the constant K_v depends on this parameter.

We write $X \sim \chi^2(v)$

It is very difficult to evaluate so tables are used.

This test is used to decide how well the observed data fits the expected data and consider whether any difference can be attributed to chance.

Now the comparison between the observed frequencies (O_i) and the Expected frequencies (E_i) for I = 1, 2, … n (that is for n pairs or values or classes) is made by considering the statistic

$$\sum_{i=1}^{n} (O_i - E_i)^2 / E_i$$

For large values of O_i and E_i this statistic approximates to the chi-squared distribution.

We define $\chi^2_{calc} = \sum_{i=1}^{n} (O_i - E_i)^2 / E_i$

$\chi^2_{calc} = 0$ then there is exact agreement between the observed and the expected data. $\chi^2_{calc} > 0$ then O_i and E_i do not agree exactly and for a given value of v the larger the value of χ^2_{calc} the greater the discrepancy.

S8.2 Goodness of Fit Tests

The chi-squared test can be used to investigate whether an observed distribution fits a well-known distribution.

The following tests can be considered
Case 1a Binomial Distribution p known
Case 1b Binomial Distribution
p unknown
Case 2 Poisson Distribution
Case 3a Normal Distribution
mean and variance known
Case 3b Normal Distribution
mean and variance unknown

S8.3 Contingency Tables - χ^2 Tests

Sometime situations arise when individuals are classified according to two sets of attributes. We may then wish to investigate whether the two sets of attributes are independent or whether there is evidence of an association between them.
This is best explained by study of specific examples.

Regression and Correlation

S9.1 Introduction

Regression is a statistical method that determines the strength and character of the relationship between one dependent variable, denoted Y and a series of other independent variables.

If the least squares regression line y on x is y = ax + b, then the values of a and b are found by solving the simultaneous equations

$\sum y = a \sum x + n\, b$ and
$\sum xy = a \sum x^2 + b \sum x$

These equations are called the normal equations for y on x – see Appendix A.

There should be an insert sheet with this booklet showing example calculations on a large data set.

S9.2 Least Squares Regression Line

We are looking for a potential relationship between x and y that is y = f(x) and we seek to determine the function f. Given only the points we have to "work backwards" or regress to the original function f. Hence the function is called the
regression function.

Define
$S_{xy} = \sum(x_{bar} - x_i)(y_{bar} - y_i)$ and
$S_{xx} = \sum(x_{bar} - x_i)^2$
$S_{yy} = \sum(y_{bar} - y_i)^2$

Let the least squares regression line be
y = a x + b where
$a = S_{xy} / S_{xx}$ and $b = y_{bar} - a\, x_{bar}$

So the residual sum of squares RSS is

$$\begin{aligned} RSS &= \sum(y_i - f(x_i))^2 \\ &= \sum(y_i - (a\, x_i + b))^2 \\ &= \sum(y_i - (a\, x_i + y_{bar} - a x_{bar}))^2 \\ &= \sum(a(x_{bar} - a x_i) - (y_{bar} - y_i))^2 \\ &= a^2 S_{xx} - 2a\, S_{xy} + S_{yy} \\ &= S_{yy} - a\, S_{xy} \\ &= S_{yy}\,(1 - S^2_{xy} / S_{xx}\, S_{yy}) \end{aligned}$$

S9.3 Location x_{bar} and y_{bar}

Now as the regression line is $y = a x + b$

$\sum y = a \sum x + nb$ and dividing through by n

$\sum y / n = a \sum x / n + b$

$y_{bar} = a x_{bar} + b$

so plot (x_{bar} , y_{bar}) lies on line $y = a x + b$

$y - y_{bar} = S_{xy} / S_{xx} (x - x_{bar})$

is the regression line y on x and

$x - x_{bar} = S_{xy} / S_{yy} (y - y_{bar})$

is the regression line x on y and

The notation of S_y^2 and S^2_y for S_{yy} is pointless and confusing. It should be avoided.

S9.4 Product Moment Corr. Coeff.

Covariance is defined as

$s_{xy} = {}^1/_n \, (\sum (x - x_{bar}) (y - y_{bar})$

see Appendix B for further clarification.

The magnitude is difficult to interpret directly so we normalise to produce the correlation coefficient r, a dimensionless measure.

Again consider

$y - y_{bar} = (S_{xy} / S_{xx}) (x - x_{bar})$ and

$x - x_{bar} = (S_{xy} / {}^2_y) (y - y_{bar})$

We standardise as follows

$(y - y_{bar}) / S_y = (S_{xy} / S_x S_y) (x - x_{bar}) / S_x$

$(x - x_{bar}) / S_x = (S_{xy} / S_x S_y) (y - y_{bar}) / S_y$

Now we move the axis to (x_{bar} , y_{bar}) with X graduated in S_x and Y graduated in S_y

$Y = (y - y_{bar}) / S_y$.

$X = (x - x_{bar}) / S_x$.

The regression line can now be written

$Y = r X$ and $X = r Y$

where $r = S_{xy} / \sqrt{(S_{xx} S_{yy})}$

r is the product moment correlation coefficient also known as Pearson's product moment correlation coeff.

Note the gradients are complementary.

$X = r Y$ makes the angle θ with the Y axis

$Y = r X$ makes the angle θ with the X axis

S9.5 Alternative expressions for r

$r = S_{xy} / \sqrt{(S_{xx} S_{yy})}$

$RSS = S_{yy} (1 - S^2_{xy} / S_{xx} S_{yy})$

$RSS = S_{yy} (1 - r^2)$

$a = S_{xy} / S_{xx}$ and

$c = S_{xy} / S_{yy}$ so

$a \times c = S^2_{xy} / (S_{xx} S_{yy})$

$a \times c = r^2$ and often written R^2

S9.6 Coefficient of Determination R^2

The coefficient of determination R^2 is a measure of the strength of association between two variables X and Y. For linear regression it is the proportion explained by the model

The following descriptors can be used

0	$\le R^2 <$	0.25	very weak
0.25	$\le R^2 <$	0.5	weak
0.5	$\le R^2 <$	0.75	moderate
0.75	$\le R^2 <$	0.9	strong
0.9	$\le R^2 <$	1.0	very strong

$R^2 = 1$ is perfect correlation

For linear models $(R^2) = (r)^2$ where (R^2) is the coefficient of determination, a measure in its own right whereas $(r)^2$ is the square of the product moment correlation coefficient.

There are special cases where R^2 can yield negative values.

S9.7 Coefficients of Rank Correlation

For the data $(x_1 , y_1), (x_2 , y_2), \ldots (x_n , y_n)$ the product-moment correlation coefficient $r = S_{xy} / \sqrt{(S_{xx} S_{yy})}$

Now suppose that instead of using precise values of the variables, we rank the numbers in order of size.

S9.8 Rank Correlation Coefficients

These rank coefficients are often used as a test statistic in a statistical hypothesis test to establish whether two variables may be regarded as statistically dependent. These tests are non-parametric, as they does not rely on any assumptions on the distributions of X or Y or the distribution of (X , Y).
Goodness of fit test and contingency tables : $\sum(O_i–E_i)^2 / E_i \;\; \chi_v^2$

S9.9 Spearman's Coefficient

$r_s = 1 – 6\sum d^2 / n (n^2 – 1)$
d is the individual rank differences.
It is easier to calculate r_s than r but less reliable than r.
To test the significance of the calculated value of r_s we calculate the probability of obtaining a given value of $\sum d^2$.
When we test r_s for significance, a suitable null hypothesis is $H_0 : \rho = 0$ where ρ is the true population correlation coefficient indicating there is no predictable correlation between the rankings. Tables are published showing probabilities $\sum d^2$ exceeds or is less than certain values for different values of n.

S9.10 Kendall's tau Coefficient

r_τ is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities.

$r_\tau = S / ½ n (n – 1)$
where S is the score value obtained by examining all of the ranking pairs. The score value is determined by making repeated passes through the data, comparing ranking.

Specifically $S = {}_{i<j}\sum \text{sgn} (x_i – x_j) \text{ sgn } (y_i – y_j)$

In general when there are n pairs of ranking the maximum S score is given by
$S = 1 + 2 + 3 + \ldots n = ½ n (n – 1)$
so Kendall's coefficient is the ratio of the actual score to the maximum possible value.

In order to test the significance of the calculated value or r_τ it is necessary to calculate the probability of obtaining a given value of S. Tables are available.

r_τ is a suitable coefficient for small data sets say $4 \le n \le 10$. As there are n! calculations, r_k become impracticable to calculate for $n > 10$ as no Excel routine exists.

For $n > 10$ the coefficient can be considered an approximation to the normal with the z-statistic having mean zero and variance
$\{2 (2n + 5)\} / \{9n (n – 1\}$

Appendix A Regression line gradient

Traditional approach to calculating the least squares regression line.

Consider a number of plots, (x_1 , y_1), (x_2 , y_2) … (x_n , y_n) where the line of best fit $y = a x + b$ has already been drawn.

Consider point plot P_1 at (x_1 , y_1) and the corresponding point Q_1 on the line of best fit.

Consider m_1 distance between P and Q

$m_1 = Q_1P_1 – P_1R_1 = a x_1 + b – y_1$

$m_1^2 = (a x_1 + b – y_1)^2$

$\sum m_1^2 = \sum(a x_1 + b – y_1)^2$

This is the sum of the squares of the residuals and the line is the least squares regression line y on x.

m_1^2 is now termed the residual sum of squares RSS

Step 1

Vary b to set $\sum m_1^2$ a minimum.

by differentiating w.r.t. b.

$d\sum m_1^2 / db = \sum 2 \sum(a x_1 + b – y_1)$

$= 2 (a \sum x_1 + n b – \sum y_1)$

At minimum $d\sum m_1^2 / db = 0$ so

$\sum y_1 = a\sum x_1 + n b$ equation ❶

Step 2

Vary a to set $\sum m_1^2$ a minimum

by differentiating w.r.t. a

$d\sum m_1^2 / da = \sum 2 (a x_1 + b – y_1) x_i$

At minimum $d \sum m_1^2 / d b = 0$ so

$\sum x_i y_i = a \sum x_i^2 + b \sum x_i$ equation ❷

So we now solve these two conditional equations.

Solving for b

$b = (\sum y \sum x^2 – \sum x y \sum x) / (\sum x^2 – (\sum x)^2)$

Solving for a

$a = (n \sum x y – \sum x \sum y) / (n \sum x^2 – (\sum x)^2)$

and dividing through by n^2

$a =$

$(\sum xy/n – \sum x/n \sum y/n) / (\sum x^2/n – (\sum x/n)^2)$

dividing through by n

$a = (\sum xy – {}^1/_n \sum x \sum y) / (\sum x – {}^1/_n (\sum x)^2)$ ❸

a is the gradient of the least squares regression line.

Appendix B Covariance

We define $s_{xy} = {}^1/_n (\sum(x – x_{bar}) (y – y_{bar})$ as the covariance of (x_1 , y_1), (x_2 , y_2) … (x_n , y_n)

and by the usual expansion and simplification

$s_{xy} = (\sum x y) / n – x_{bar} y_{bar}$

$s_{xy} = (\sum x y) / n – \sum x / n \sum y / n$ ❹

By extension we define

$s_{xx} = {}^1/_n (\sum(x – x_{bar})^2$

$s_{xx} = \sum x_i^2 – (\sum x_i^2) / n$ ❺

$s_{yy} = {}^1/_n (\sum(y – y_{bar})^2$

$s_{xx} = \sum y_i^2 – (\sum y_i^2) / n$

and by extension

$s_{xy} = {}^1/_n (\sum(x – x_{bar}) (y – y_{bar})$

substituting ❹ and ❺ into ❸ we deduce $a = s_{xy} / s_{xx}$ the gradient of the line termed the regression coefficient of y on x.

However modern terminology drops the divisor by n and write S_{xx}, S_{yy} and S_{xy}.

Robert Goodhand April 2024