Edexcel Further Maths A-level

Further Statistics 2

Formula Sheet

Provided in formula book	Not provided in formula book

Linear Regression

Equation of the regression line of y on x:	$y = a + bx$	$b = \frac{S_{xy}}{S_{xx}}$, $a = \bar{y} - b\bar{x}$

Summary Statistics

For a set of n pairs of values (x_i, y_i):	$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$
	$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$
	$S_{xx} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$

Product Moment Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\Sigma(x_i - \bar{x})^2\}\{\Sigma(y_i - \bar{y})^2\}}} = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\sqrt{\left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right)\left(\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}\right)}}$$

Residual Sum of Squares (RSS)

$$RSS = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = S_{yy}(1 - r^2)$$

Spearman's Rank Correlation Coefficient

$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$	$n =$ *number of pairs of observations* $d =$ *difference between ranks of each observation*

$r = -1$	$r = 0$	$r = +1$
Rankings are in exact reverse order	No correlation between rankings	Rankings in perfect agreement

Continuous Probability Distributions

For a continuous random variable X with probability density function $f(x)$:	$f(x) \geq 0$ for all $x \in \mathbb{R}$
	$P(a < X < b) = \int_a^b f(x)\,dx$
	$\int_{-\infty}^{+\infty} f(x)\,dx = 1$

Probability density function	$f(x) = \frac{dF(x)}{dx}$
Cumulative distribution function	$F(x_o) = P(X \leq x_o) = \int_{-\infty}^{x_o} f(x)\,dx$

Expectation (mean)	$E(X) = \mu = \int xf(x)\,dx$
Variance	$Var(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$
For a function $g(X)$:	$E\big(g(X)\big) = \int g(x)f(x)\,dx$
Median, m	$\int_a^m f(x)\,dx = 0.5$
Lower quartile, Q_1	$\int_a^m f(x)\,dx = 0.25$
Upper quartile, Q_3	$\int_a^m f(x)\,dx = 0.75$
n^{th} percentile	$\int_a^m f(x)\,dx = \frac{n}{100}$
Mode	$\frac{df(x)}{dx} = 0$ *(value at which the p.d.f is a maximum)*

$\mathrm{E}(aX+b) = a\mathrm{E}(X) + b$	$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$

Skewness

Positive skew	$mode < median < mean$
Negative skew	$mean < median < mode$

Continuous Uniform Distribution

Probability density function:	$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & otherwise \end{cases}$

Mean	$\frac{a+b}{2}$
Variance	$\frac{(b-a)^2}{12}$
Probability distribution function	$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$

Combination of Random Variables

If X, Y are random variables:	$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
	$\mathrm{E}(X-Y) = \mathrm{E}(X) - \mathrm{E}(Y)$
	$\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(Y)$
If X, Y are independent random variables:	$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(\mathrm{Y})$
	$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(\mathrm{Y})$
	$\mathrm{Var}(aX \pm bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(\mathrm{Y})$

If X and Y are independent random variables with $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$:

$$aY + bX \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

$$aY - bX \sim N(a\mu_x - b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

If $X_1, X_2, \ldots, X_n$ are independent identically distributed random variables with $X_i \sim N(\mu, \sigma^2)$:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

PMT •resources•tuition•courses

Estimates and Tests Using a Normal Distribution

S^2 – unbiased estimator for σ^2 *(random variable)*	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
s^2 – estimate *(observation from a random variable)*	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{S_{xx}}{n-1} = \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right) = \frac{1}{n-1}(\sum x^2 - n\bar{x}^2)$

$\bar{X}$ is an unbiased estimator of μ, with $Var(\bar{X}) = \frac{\sigma^2}{n}$

Standard deviation of an estimator is
$standard\ error = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

If the population is assumed to be normal, then, for large samples, the statistic $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ *has an approximate* $N(0, 1^2)$ *distribution.*

If the population is not normal, by assuming that s *is a close approximation to* σ*, then for large samples,* $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ *can be treated as having an approximate* $N(0, 1^2)$ *distribution.*

For a random sample of n *observations from* $N(\mu, \sigma^2)$,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

For a random sample of n_x *observations from* $N(\mu, \sigma^2)$ *and, independently, a random sample of* n_y *observations from* $N(\mu_y, \sigma_y^2)$,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$$

Confidence Intervals

	$(\bar{x} - z \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z \times \frac{\sigma}{\sqrt{n}})$ where z is the relevant percentage point from the standard normal distribution.
Example:	95% confidence interval for μ: $\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$

Variance of a Normal Distribution

If a random sample of n observations $X_1, X_2, \dots, X_n$ is selected from $N(\mu, \sigma^2)$ then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

For a probability of α that the variance falls outside the limits:	
The $100(1-\alpha)\%$ *confidence limits are:*	$\frac{(n-1)s^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}$ and $\frac{(n-1)s^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}$
The $100(1-\alpha)\%$ *confidence interval for the variance of a normal distribution is:*	$\left(\frac{(n-1)s^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{(n-1)s^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)$

F-distribution

For a random sample of n_x observations from an $N(\mu_x, \sigma_x^2)$ distribution and an independent random sample of n_y observations from an $N(\mu_y, \sigma_y^2)$ distribution,

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n_x-1, n_y-1}$$

If a random sample of n_x observations is taken from a normal distribution with unknown variance σ^2 and an independent random sample of n_y observations is taken from a normal distribution with equal but unknown variance, then

$$\frac{S_x^2}{S_y^2} \sim F_{n_x-1, n_y-1}$$

$$F_{\nu_1, \nu_2} = \frac{1}{F_{\nu_2, \nu_1}}$$

t-distribution

For a random sample of n observations from $N(\mu, \sigma^2)$:	$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

For a small sample of size n from a normal distribution $N(\mu, \sigma^2)$ with unknown mean and variance:	
The $100(1-\alpha)\%$ *confidence limits for the population mean are:*	$\bar{x} \pm t_{n-1}\frac{\alpha}{2} \times \frac{s}{\sqrt{n}}$
The $100(1-\alpha)\%$ *confidence interval for the population mean is:*	$\left(\bar{x} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}\right)$

Paired t-test

(two independent normal distributions X, Y with equal unknown variances)

In a paired experiment with a mean of the differences between the samples of $\bar{D}$:	$\frac{\bar{D} - \mu_D}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Pooled estimate for σ^2:
$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$
$s_x^2 = \frac{\sum x^2 - n_x\bar{x}^2}{n_x - 1}$, $s_y^2 = \frac{\sum y^2 - n_y\bar{y}^2}{n_y - 1}$

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}} \sim t_{n_x+n_y-2}, \text{ where } S_p^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2}$$

Confidence limits for the difference between the two means of X and Y:	$(\bar{x} - \bar{y}) \pm t_c s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$
Confidence interval:	$\left((\bar{x} - \bar{y}) - t_c s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}, (\bar{x} - \bar{y}) + t_c s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}},\right)$
	t_c - relevant value from t-tables