Revision Sheet

FM Unit 2 WJEC

Probability Distributions

The expected value of a constant is the constant: $E(a) = a$

The **variance** of X is:
$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2$

The **standard deviation** of X is:
$\sigma = \sqrt{Var(X)}$

$P(A \cap B) = P(A)P(B)$ for **independent** A and B.

Combinations of Random Variables

Use the results:
$E(aX + b) = aE(X) + b$
$Var(aX + b) = a^2Var(X)$
$E(aX + bY) = aE(X) + bE(Y)$

For **independent** X and Y, use the results:
$E(XY) = E(X)E(Y)$
$Var(aX + bY)$
$= a^2Var(X) + b^2Var(Y)$

Discrete Probability Distributions

Finding the **mean** and **variance** using:
(a) $E(X) = \mu = \sum x_i p_i$
(b) $Var(X) = \sigma^2 = \sum {x_i}^2 p_i - \mu^2$
where $p_i = P(X = x_i)$.

For a function $g(X)$:
$E\big(g(X)\big) = \sum g(x_i)p_i$

Use the result: $\sum_x P(X = x) = 1$

Changing probabilities from $<$ to $\leq$ e.g.
$P(Y < 2) = P(Y \leq 1)$

Changing $\leq$ to $\geq$ and back:
$P(Y \leq 3) = 1 - P(Y > 3)$
$= 1 - P(Y \geq 4)$

Continuous Probability Distributions

Finding the **mean** and **variance** using the **probability density function** (pdf) $f(x)$ of X:
(a) $E(X) = \mu = \int xf(x)\,dx$
(b) $Var(X) = \sigma^2 = \int x^2 f(x)\,dx - \mu^2$

For a function $g(X)$:
$E\big(g(X)\big) = \int g(x)f(x)\,dx$

Use the result:
$\int_a^b f(x)\,dx = 1$ where $a \leq x \leq b$ and $f(x) = 0$ for $x < a$ and $x > b$

Calculating **probabilities**:
$P(c < X < d) = \int_c^d f(x)\,dx$

Cumulative distribution function:
$F(x) = P(X \leq x)$

Changing probabilities from $<$ to $\leq$ e.g.
$P(Y < 2) = P(Y \leq 2)$

Changing $\leq$ to $\geq$ and back:
$P(Y \leq 2) = 1 - P(Y > 2)$
$= 1 - P(Y \geq 2)$

Converting between probability density and cumulative distribution functions.

$f(x) = \frac{dF(x)}{dx}$ and
$F(x) = \int_{-\infty}^{x} f(t)dt$

Finding the **median**: $F(m) = 0.5$

The Poisson Distribution

Suppose $X \sim \text{Po}(\lambda)$ is the number of events occurring per 1 unit of time/space, then if Y is the number of these events occurring per t units of time/space, then: $Y \sim \text{Po}(\lambda t)$
$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$.
[$x! = x \times (x-1) \times (x-2) \times \cdots \times 3 \times 2 \times 1$.]

Conditions

(a) The events take place randomly and are independent of each other.
(b) The events happen one at a time.
(c) On average, the events occur at a constant rate.

Mean: $E(X) = \lambda$
Variance: $Var(X) = \lambda$

The sum of **independent** Poisson random variables also has a Poisson distribution.
If $X_i \sim \text{Po}(\lambda_i)$ for $i = 1, \ldots, n$ then $\sum_1^n X_i \sim \text{Po}(\sum_1^n \lambda_i)$

The Binomial Distribution

$X \sim \text{B}(n, p)$.
n = number of trials.
p = probability of success.
$q = 1 - p$.

$P(X = x) = \binom{n}{r}p^r(1-p)^{n-r}$.

Conditions

(a) Two possible outcomes in each trial.
(b) A fixed number of trials.
(c) Independent trials.
(d) Identical trials (p is the same for each trial).

Mean: $E(X) = np$
Variance: $Var(X) = np(1-p)$

Poisson Approximation to the Binomial

If $X \sim \text{B}(n, p)$ with large n (say > 50) and small p (say < 0.1) then we can approximate:
$X \sim \text{Po}(np)$

Exponential Distribution

$X \sim \text{Exp}(\lambda)$ if

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

Mean: $E(X) = \frac{1}{\lambda}$
Variance: $Var(X) = \frac{1}{\lambda^2}$

Cumulative distribution function:

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

Results:
$\frac{d}{dx}(e^{kx}) = ke^{kx}$ $\quad \int e^{kx}dx = \frac{e^{kx}}{k}\ (+c)$

If the number of events in t units of time, X, has a Poisson distribution with mean λt, then the waiting times between the events, T, are exponentially distributed with parameter λ.

$X \sim \text{Po}(\lambda t) \Leftrightarrow T \sim \text{Exp}(\lambda)$

PROOF

Least Squares Regression

The least squares regression line/model of Y on X:

$$y = a + bx$$

where $a = \bar{y} - b\bar{x}$ and $b = \frac{S_{xy}}{S_{xx}}$

Interpret a and b in context.

Interpolation: estimating *between* data points.
Extrapolation: estimating *outside* data points.

Pearson's Product Moment Correlation Coefficient

Denoted r.
Measures **linear** association between two Normally distributed continuous random variables.
For a sample of n pairs (x_i, y_i):
$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, $S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$

Interpret the value of r.

Spearman's Rank Correlation Coefficient

Denoted r_s
Measures the **monotonic relationship** between two variables.
For a sample of n pairs (x_i, y_i):

$r_s = 1 - \frac{6\sum_i {d_i}^2}{n(n^2-1)}$

where d_i is the **difference** between the ranks of the ith pair.

Interpret the value of r_s.

Deciding when r and r_s are appropriate.

Testing the Significance of Correlation Coefficients

ρ: Population correlation coefficient
r/r_s: Sample correlation coefficient

$H_0: \rho = 0$	$H_0: \rho = 0$	$H_0: \rho = 0$
$H_1: \rho < 0$	$H_1: \rho > 0$	$H_1: \rho \neq 0$
One-tailed	One-tailed	Two-tailed

Compare the test statistic r/r_s with the critical value, r_{crit} from tables.

Conclude (in *context*):
$|r| \geq r_{\text{crit}}$: Evidence to reject H_o in favour of H_1.
∴ Significant ... relationship.
$|r| < r_{\text{crit}}$: Insufficient evidence to reject H_o in favour of H_1.
∴ No significant ... relationship

Chi-squared Goodness of Fit Test

H_0: Data **does** follow specified distribution
H_1: Data **does not** follow specified distribution

n: Number of categories (include every possibility).
O_i: Observed frequency of ith category.
E_i: Expected frequency of ith category, assuming H_0 is true.

Calculate expected frequencies for **Binomial**, **Poisson** and **Discrete Uniform** distributions.

If $E_i < 5$ then **pool** categories.

Calculate the chi-squared **test statistic**

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

Degrees of Freedom (ν) = # of categories after pooling -1

Compare the test statistic X with the critical value, X_{crit} from tables.

Conclude (in *context*):
$X \geq X_{\text{crit}}$: Evidence to reject H_o in favour of H_1.
∴ Specified distribution is not a good fit.
$X < X_{\text{crit}}$: Insufficient evidence to reject H_o in favour of H_1.
∴ Specified distribution is a good fit.

Chi-squared No Association Test

H_0: Variables are **independent**
H_1: Variables are **not independent**

$O_{i,j}$: observed frequency of ith row and jth column categories.
$E_{i,j}$: Expected frequency of ith row and jth column categories, assuming H_0 is true.

$$E_{i,j} = \frac{i^{th} \text{ row total} \times j^{th} \text{ column total}}{\text{sample size}}$$

If $E_{i,j} < 5$ then **pool** 'logical' categories.

Calculate the chi-squared **test statistic**

$$X = \sum_{i=1,j=1} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$\nu = (c-1) \times (r-1)$ (after pooling)
Compare the test statistic X with the critical value, X_{crit} from tables.

Conclude (in *context*):
$X \geq X_{\text{crit}}$: Evidence to reject H_o in favour of H_1.
∴ Variables are not independent
$X < X_{\text{crit}}$: Insufficient evidence to reject H_o in favour of H_1.
∴ Variables are independent

The Exam: Length: 1 hour 30 minutes. 75 marks. $13\frac{1}{3}$% of the A Level. $33\frac{1}{3}$% of the AS Level. **Checklist** ☐ Attempted all past paper questions. ☐ Know which formulae appear in the formula booklet.

Mr Tom Williams, Ysgol y Creuddyn
May 2023 www.mathemateg.com