THE UNIVERSITY OF MELBOURNE
FACULTY OF BUSINESS & ECONOMICS

Helpsheet

Giblin Eunson Library

PROBABILITY DISTRIBUTION

Use this sheet to help you:

- identify which distribution to select for a given situation
- with the method of calculation for probability, expectation or mean and variance of each distribution
- calculation of confidence interval estimators of μ

Author: Carter, D.
Design and layout: Shaw, J.

Discrete Probability Distributions

1. **The Binomial distribution** – is used when there are two outcomes

i) success: p

ii) failure: q (1 – p)

The trials are independent and the random variable, X, equals the number of successes in the n trials.

Set the number of trials, n, number of successful trials, x, then

$$\Pr(X = x) = {}^{n}C_{r}p^{x}(1 - p)^{n-x}, \qquad x = 0, 1, 2, \ldots\ldots.n \qquad (X \qquad X)$$

$$E(X) = \mu = np$$

$$V(X) = \sigma^2 = npq$$

$$SD(X) = \sigma = \sqrt{npq}$$

Example: In an instant lottery with 20% winning tickets, if X is equal to the number of winning tickets among the 8 that are purchased, the probability of purchasing 2 winning tickets is:

$$\Pr(X = 2) = {}^{8}C_{2}(0.2)^{2}(0.8)^{6}$$

$$= 0.2936$$

2. **The Poisson distribution** – is used when there are two outcomes

i) success: μ = average number of successes in an interval of time or specific region of space.

ii) failure

Examples of a couple of situations where the Poisson distribution may be used are: the number of cars arriving at a service station in one hour or the number of accidents in one day on a particular stretch of highway.

$$\Pr(X = x) = p(x) = \frac{e^{-\mu}\mu^{x}}{x!}, \qquad x = 0, 1, 2$$

μ = average number of successes occurring in the given time interval.
There are tables for selected values of μ.

Probability Density Functions (pdf)

There are a number of continuous distributions (probability density functions, pdf's). These include the normal distribution, student t-distribution, chi-squared distribution and the F distribution.

1. Normal Distribution – its curve is symmetric about the mean and its random variable has a range of -• to •. It has a random variable, X; mean, μ; and variance, σ^2. ($N(\mu, \sigma^2)$) To calculate normal probabilities it is necessary to use a given table, calculator or computer program such as Excel.

To use a table (and some calculators) we need to standardise using the rule $\frac{x - \mu}{\sigma}$, which gives us the standard normal random variable, z.

The standard normal distribution:

$\mu_Z = 0$, $\sigma^2_Z = 1$, $\sigma_Z = 1$.

Example: The amount of time to assemble a computer is normally distributed with a mean of 50 minutes and a standard deviation of 6 minutes. What is the probability that a computer is assembled between 50 and 59 minutes?

N(50, 36)

$$\Pr(50 < X < 59) = \Pr\left(\frac{50 - 50}{6} < \frac{x - \mu}{\sigma} < \frac{59 - 50}{6}\right)$$

$= \Pr(0 < Z < 1.5)$

= 0.4332 (from a table of values for z)

The Normal distribution can generally be used as an approximation to the binomial distribution and works best if $np \geq 5$ and $nq \geq 5$.

In order to approximate the discrete binomial distribution to the continuous normal distribution we need to add or subtract 0.5 to the values of X.

The 0.5 is called the continuity correction factor.

Example: If X is a binomial random variable with n = 100 and p = 0.6, use the normal distribution to find the approximate probability of $\Pr(X \leq 70)$.

From the binomial distribution:

$$\mu = np \qquad \sigma = \sqrt{npq}$$

$$\therefore \mu = 0.6 \times 100 \qquad \therefore \sigma = \sqrt{100 \times 0.6 \times 0.4}$$

$$= 60 \qquad = 4.899$$

Then the normal approximation is:

$$\Pr(X \leq 70) \approx \Pr(-.05 \leq Y \leq 70.5)$$

$$= \Pr\left(\frac{-0.5 - 60}{4.899} \leq \frac{Y - \mu}{\sigma} \leq \frac{70.5 - 60}{4.899}\right)$$

$$= \Pr(-12.3495 \leq z \leq 2.1433)$$

$$= 0.984$$

If a population is normally distributed or n ≥ 30, the confidence interval estimator of μ is:

$$\overline{X} - z_{\sigma/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{X} + z_{\sigma/2}\frac{\sigma}{\sqrt{n}}$$

The probability (1 – α) is called the confidence level.
Three commonly used confidence levels and $z\alpha/2$ values are:

Confidence level	α	α/2	$z_{\alpha/2}$
0.90	**0.10**	**0.05**	**1.645**
0.95	**0.05**	**0.025**	**1.96**
0.99	**0.01**	**0.005**	**2.575**

The 95% confidence interval estimator of μ implies that 95% of the values of $\overline{X}$ (sample mean) will create intervals that contain the true value of the population mean (μ). The other 5% of sample means will create intervals that do not include the population mean. i.e. 95% of the interval estimates will be correct and 5% will be incorrect.

2. **Student t Distribution** – is used for estimating the population mean, μ, when the variance, σ^2, is unknown and the population is normally distributed.

t-statistic = , $\dfrac{(\overline{x} - \mu)}{s/\sqrt{n}}$,

E(t) = 0, V(t) = $\dfrac{n-1}{n-3}$ for $n > 3$

(V(t) is always greater than 1.)

Student t is more widely dispersed than the standard normal distribution.

The extent to which the t-distribution is more spread out than the standard normal distribution is determined by a function of the sample size called the degrees of freedom(d.f. denoted by v) which varies by the t-statistic.

Degrees of freedom, v = n – 1.

As with the normal distribution there are tables of values or Excel can be used for calculations for $t_{\alpha,v}$.

The confidence interval estimator of μ, with σ^2 unknown, and population normally distributed is:

$$\overline{X} \pm t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$$

If the population variance, σ^2, is unknown, the population is normally distributed and n > 200, then the confidence interval estimator of μ is:

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Student t distribution is based on using the sample variance to estimate population variance. note – the denominator is the same as the degrees of freedom.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

3. Chi-squared distribution (χ^2)- is used when the population variance and sample variance are known. It can also be used for a multinomial experiment or contingency table. It is positively skewed from 0 - •. Its shape depends on the number of degrees of freedom.

Test statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ or

$\chi^2 = \sum \frac{(o - e)^2}{e}$, where o = observed value, e = expected value for multinomial distributions or contingency tables.

The degrees of freedom for a multinomial distribution is the number of possible outcomes minus 1.i.e. n – 1

The degrees of freedom for a contingency table is the number rows minus 1 multiplied by the number of columns minus 1. i.e. (r – 1)(c – 1).

Significance level: α

There are tables of values for $\chi^2_{\alpha,n-1}$ and $\chi^2_{1-\alpha,n-1}$ or the values can be calculated in Excel.

$$\text{LCL} = \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \text{ and } \text{UCL} = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Example: Management believes that the daily demand for a product is normally distributed with a variance level equal to 250. In an experiment to test this belief about σ^2, the daily demand was recorded for 25 days. The data have been summarised as follows:

$\sum x_i = 1265$ $\sum x_i^2 = 76,010$

i) Do the data provide sufficient evidence to show the management's belief about the variance is untrue for $\alpha = 0.01$?

$H_0: \sigma^2 = 250$

$H_A: \sigma^2 \neq 250$

Test statistic: $$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Level of significance: $\alpha = 0.01$

Decision rule: reject

$$\chi^2_{0.995,24} = 9.886, \quad \chi^2_{0.005,24} = 45.5585$$

$$H_0 \text{ if } \chi^2 < \chi^2_{1-\alpha/2,n-1} \text{ or } > \chi^2_{\alpha/2,n-1}$$

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \frac{1}{24}\left[76010 - \frac{(1265)^2}{25}\right] = 500$$

$$\chi^2 = \frac{24(500)}{250} = 48$$

$$\chi^2_{0.995,24} = 9.886, \quad \chi^2_{0.005,24} = 45.5585$$

As 48 > 45.5585 reject null hypothesis. The data does not provide sufficient evidence to support the management's belief about the variance at $\alpha = 0.01$.

ii) Determine the 95% confidence level estimate of σ^2.

$$\text{LCL} = \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} = \frac{24(500)}{\chi^2_{0.025,24}} = \frac{24(500)}{39.3641} = 304.85$$

$$\text{UCL} = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{24(500)}{\chi^2_{0.975,24}} = \frac{24(500)}{12.401} = 967.66$$

The 95% confidence interval estimate of σ^2 (304.85,967.66)

4. F distribution – is used to draw inferences about $\frac{\sigma_1^2}{\sigma_2^2}$, using the ratio of two independent sample variances $\frac{s_1^2}{s_2^2}$.

The F distribution has two numbers of degrees of freedom, $\nu_1 = n_1 - 1$, numerator degrees of freedom and $\nu_2 = n_2 - 1$, denominator degrees of freedom. (ν = nu).

There are tables or Excel can be used to calculate F_{α,ν_1,ν_2}.

$$F_{1-\alpha/2,\nu_1,\nu_2} = \frac{1}{F_{\alpha/2,\nu_2,\nu_1}}$$

Confidence interval estimator of $\frac{\sigma_1^2}{\sigma_2^2}$,

$$LCL = \frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2,\nu_1,\nu_2}} \quad \text{and} \quad UCL = \frac{s_1^2}{s_2^2}F_{\alpha/2,\nu_2,\nu_1}$$

Example: Variance is often the measurement used to assess the risk of an investment. In a risk comparison, given the following information of two investments; investment 1 $s^2 = 82.23$ and investment 2 $s^2 = 530.2143$:

$$LCL = \frac{s_1^2}{s_2^2}\frac{1}{F_{0.01,9,7}} \quad \text{and} \quad UCL = \frac{s_1^2}{s_2^2}F_{0.01,7,9}$$

$$LCL = \left(\frac{82.23}{530.2143}\right)\frac{1}{6.72} = 0.0231$$

$$UCL = \left(\frac{82.23}{530.2143}\right)5.61 = 0.870$$

$$UCL = \left(\frac{82.23}{530.2143}\right)5.61 = 0.870$$

ii) Do these data provide sufficient evidence to indicate that investment 2 is riskier than investment 1 (use $\alpha = 0.05$)?

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$$

Test Statistic: $F = \frac{s_1^2}{s_2^2} = \frac{82.23}{530.2143} = 0.1551$

$\alpha = 0.05$, decision rule:

reject H_0 if $F < \frac{1}{3.29} = 0.304$

since $F = 0.155 < 0.304$, reject H_0

Yes, there is sufficient evidence to conclude that investment 2 is riskier than investment 1.

References:

Selvanathan, Anthony, *Australian Business Statistics* Edition 4 2006, Cengage Learning Australia

Hogg, Robert V., Tanis, Elliot A., P*robability and Statistical Inference* 7th Edition 2006, Pearson Education Inc. Pearson Prentice Hall, NJ USA.