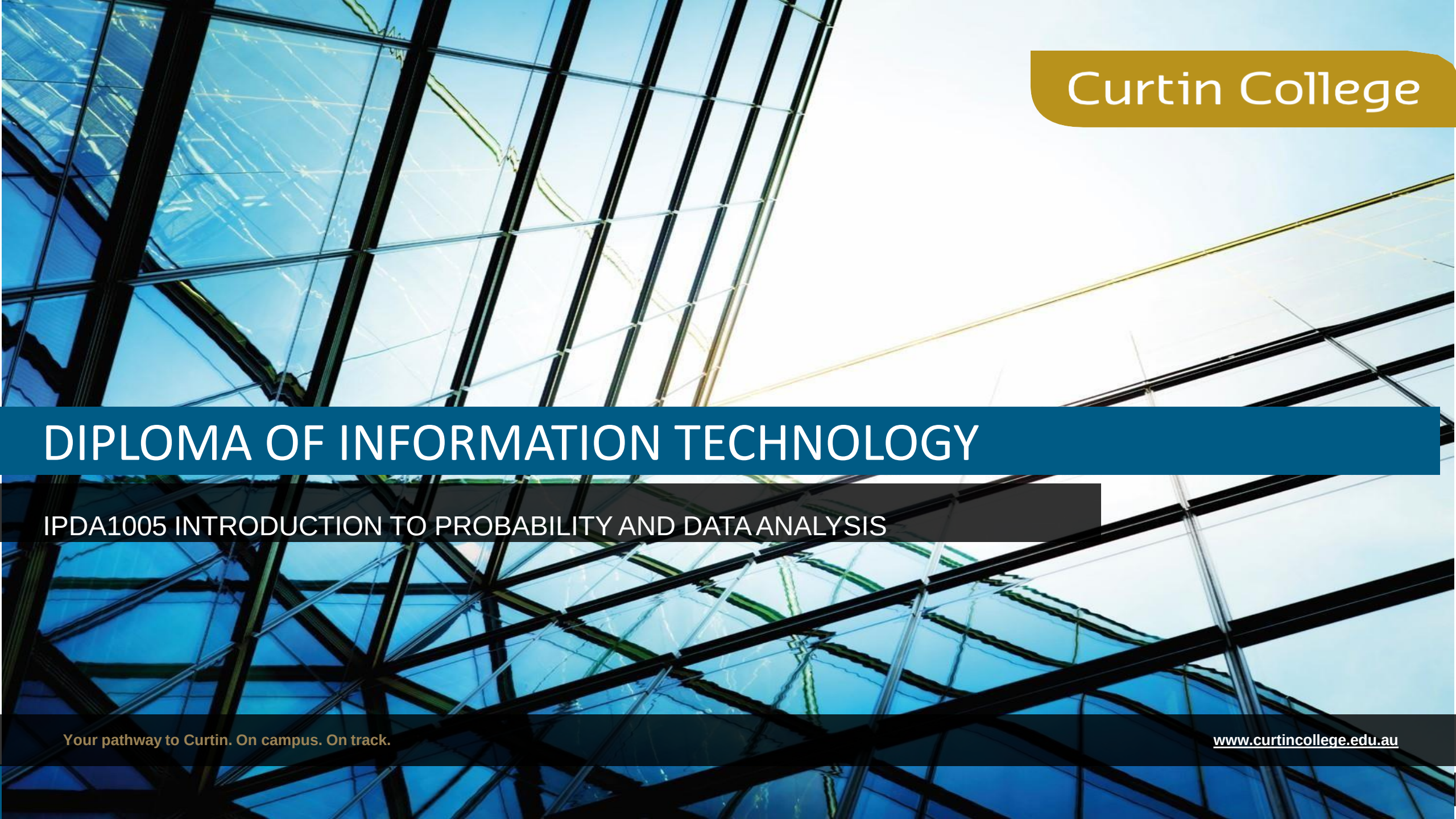

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Outline

1. Goodness-of-Fit Test

2. Two-way Tables and Test of Independence

Goodness of Fit

- Until now, we have justified using certain probability models
  - on theoretical grounds (binomial, hypergeometric, negative binomial, poisson), or
  - by assuming that they were appropriate for the physical situation that was being described
- But how can we *test* whether data we have collected are consistent with a particular probability model?
- The idea of goodness-of-fit testing is to assess whether what we actually observed (the data) is 'close enough' to what we would expect to observe for random observations from a particular probability model.

Goodness of Fit

- If I toss a coin 50 times and observe 28 heads, is it reasonable to think that the coin is fair? We can answer this by testing $H_0 : p = .5 \quad \text{against} \quad H_1 : p \neq .5.$
- A die is tossed 120 times, and the number of times each face turns up is counted and yields the following table:

Outcome (x)	1	2	3	4	5	6
Observed Frequency (O)	20	22	17	18	19	24

- Here, the hypothesis of a fair die is the hypothesis that the distribution of outcomes is the discrete uniform distribution

  $$p(x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6$$

- There are 6 proportions to test, and our earlier proportion test cannot do the job.

Test of Goodness of Fit

- H_0 : Die is fair (all outcomes are equally likely)
  H_1 : Die is not fair (at least one outcome differs in likelihood).
- Under H_0 the expected frequency of each outcome is $120\left(\frac{1}{6}\right) = 20$.

Outcome	1	2	3	4	5	6
Observed Frequency (O)	20	22	17	18	19	24
Expected Frequency (E)	20	20	20	20	20	20

- We need a test statistic that accumulates all deviations between Observed and Expected frequencies without letting positive deviations cancel out negative deviations.
- The statistic is $\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$
- This statistic will be -
  - Small if H_0 is true, since differences between Observed and Expected will be small.
  - Large if H_0 is false, since difference between Observed and Expected will be large.

Test of Goodness of Fit

In general, the test statistic for a goodness-of-fit test for independent data arranged in k cells is

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

which has approximately a chi-squared distribution with $k-1$ degrees of freedom, if all $E_i \geq 5$.

- $df = k - 1$ because the total count is known, and so if $k-1$ expected frequencies are known, the expected frequency for the remaining cell is already determined.
- For the die-rolling example, $k = 6$, and the test statistic is

$$\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \cdots + \frac{(24-20)^2}{20} = 1.7$$

Test of Goodness of Fit

- $df = k - 1 = 5$, and the p-value for the test is $P(\chi^2_5 > 1.7) = 0.889$ This very large p-value provides no evidence at all against H_0. The data are consistent with the die being fair.
- **Assumptions**: As part of the hypothesis test we should state and evaluate the assumptions, which are that the observations must be independent (usually achieved through random sampling) and the expected counts must be at least 5. We didn't get a detailed description of how the die was rolled, but randomness is a reasonable assumption. The expected counts are all 20, which is well over 5, so they are sufficiently large.

An Example

- **Example:** Two different phenotypes of tomato plants are crossed twice, and, according to Mendel's laws of inheritance, the four resulting phenotypes should occur in the ratios 9:3:3:1. The observed distribution of phenotypes was

Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4
926	288	293	104

Is the observed distribution consistent with Mendel's laws?

- H_0 : the phenotypes occur according to Mendel's laws.
  H_1 : the phenotypes follow some other pattern.
- To convert ratios to fractions, divide by the total of the ratios. Thus, 9:3:3:1 becomes 9/16, 3/16, 3/16, 1/16. The total of the observed phenotypes is 1611, and we calculate expected counts.

	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4
Obs	926	288	293	104
Exp	906.19	302.06	302.06	100.69

An Example

- The test requires independent observations and expected counts ≥ 5. Genetic inheritance is inherently random (according to Mendel's laws) and the smallest expected count is over 100, so the assumptions are met.
- We add another line to the table for the test statistic calculation.

	Pheno 1	Pheno 2	Pheno 3	Pheno 4	Total
Obs	926	288	293	104	1611
Exp	906.19	302.06	302.06	100.69	1611
χ^2	0.433	0.654	0.272	0.109	1.468

- $df = 3$ and P-value $= P(\chi^2_3 > 1.468) = 0.6897$
  The large P-value gives us no reason to reject the null hypothesis. The data are consistent with Mendel's laws.

When Parameters are Estimated

- Sometimes the hypothesis may be that the data come from a distribution with unspecified parameters, such as Poisson distribution with unspecified mean.
- Here $df = k - 1 - d$, where d is the number of parameters estimated from the data.
- **Example:** We want to test whether the following data come from a Poisson distribution.

x	0	1	2	3	4	≥ 5
Obs	5	6	11	7	5	0

- As the Poisson mean λ is unspecified, we estimate it from the data by

$$\overline{x} = \frac{1(6) + 2(11) + 3(7) + 4(5)}{5 + 6 + 11 + 7 + 5} = \frac{69}{34} = 2.029$$

- Then expected counts follow the Poisson(2.029) distribution.

x	0	1	2	3	4	≥ 5	Total
Obs	5	6	11	7	5	0	34
Exp	4.47	9.07	9.20	6.22	3.16	1.88	34

When Parameters are Estimated

- Three cells have expected counts < 5, breaching the size condition.
- This can be remedied by combining with adjacent cells.

x	0-1	2	3	≥ 4	Total
Obs	11	11	7	5	34
Exp	13.54	9.20	6.22	5.04	34

- All expected frequencies are now ≥ 5. The test statistic calculation is:

x	0-1	2	3	≥ 4	Total
χ^2	0.476	0.352	0.097	0.000	0.9253

- $df = 4 - 1 - 1 = 2$. Due to low expected frequencies, we have reduced the data from 6 cells to 4, and we deduct 1 df for the estimated Poisson mean.
- p-value $= P(\chi^2_2) > 0.9253 = 0.6296$.
- This is another large p-value that provides no evidence against H_0. The data are consistent with a Poisson distribution.

GoF for Continuous Distributions

- The method outlined above can also be used for testing goodness of fit for continuous distributions.
- Suppose we have data that we suspect come from an exponential(3) distribution.
- If we collect a sample of size n to test this, we can divide the range of the data into intervals and count the frequencies of observations that fall into each of these intervals. We then compute the expected frequencies of these cells by multiplying n by the corresponding probability computed under the assumption of exponential(3) distribution.
- The intervals should be chosen so that the minimum frequency condition is met.
- If the number of intervals is k, then $df = k - 1 - d$, where d is the number of parameters (if any) that are estimated from the data.

GoF for Continuous Distributions

Example: We have a random sample of size 50 that we want to test for normality. The sample mean and standard deviation are $\overline{x} = 3.978$ and $s = 3.024$. Observed frequencies in 6 intervals are as follows:

x	< 1	1-3	3-4	4-5	5-7	> 7
Obs	10	12	3	7	10	8

At the 5% level of significance, test whether the data come from a normal distribution.

- We compute expected counts using a $N(3.978, 3.024^2)$ distribution. For instance, the expected count in cell "1-3" is calculated by the R command `50*(pnorm(3,3.978,3.024)-pnorm(1,3.978,3.024))`.

x	< 1	1-3	3-4	4-5	5-7	> 7	Total
Obs	10	12	3	7	10	8	50
Exp	8.12	10.54	6.49	6.47	10.44	7.94	50
χ^2	0.436	0.202	1.873	0.043	0.019	0.000	2.574

GoF for Continuous Distributions

- The test requires independent observations and expected counts ≥ 5. We were told that the sample was random, and the smallest expected count is 6.47, so the expected counts are sufficiently large. The chi-squared test will be valid.
- The χ^2 test statistic is 2.574. We compare this with a chi-square distribution on $6 - 1 - 2 = 3$ degrees of freedom.
- p-value $= P(\chi^2_3 > 2.574) = 0.462$. This is well above .05 so at a 5% significance level we do not reject the hypothesis that the data came from a normal distribution. The data may well come from a normal distribution.

Two-way Tables

- A common and important problem involves collecting data that can be classified in two ways.
- The data table will be arranged in $r \geq 2$ rows and $c \geq 2$ columns, so that there are $r \times c$ cells containing frequencies or counts.
- There are two commonly encountered situations.
  1. We may be comparing the distribution of a common set of categories between several populations. For example, customers of 3 department store chains may have the same 5 payment categories: cash, Apple pay, store credit card, Visa, and MasterCard, and we might compare the distribution of payment categories between the department stores.
  2. There may be a single population of interest, with each individual in the population categorized by two different factors. For example, customers making a purchase might be classified according to the department in which the purchase was made (6 departments), and according to the method of payment (5 methods).
- Both scenarios are analysed in the same way using a two-way table.

Test of Independence

- The two-way table of counts is known as a *contingency table.*
- We test for independence (or association or relationship) between the two sets of categories by means of a chi-squared test. Once we have the Observed and Expected counts, the test statistic is calculated in the same way as for the Goodness-of-fit test. The difference arises from how the Expected counts are calculated.
- In a test of independence, H_0 is always that the two sets of categories are independent. Under H_0, the expected frequency in cell (i, j) is given by

$$E_{ij} = \frac{i^{\text{th}} \text{ row total} \times j^{\text{th}} \text{ column total}}{\text{Total number of observations}} = \frac{r_i \times c_j}{n}$$

  where r_i is the number of observations in the ith row and c_j is the number of observations in the jth column.

Test of Independence

- As before, test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ summed over all cells. The statistic has a chi-square distribution with $df = (r-1)(c-1)$.

  Example: A study of the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline ("An Analysis of Price Aggressiveness in Gasoline Marketing," J. Market. Res., 1970: 36–42) reports the accompanying data based on a sample of $n = 441$ stations. At level $\alpha = 0.01$, does the data suggest that facility conditions and pricing policy are independent of each other?

Test of Independence

- The observed frequencies are given in the table below:

Status	Aggressive	Neutral	Non-aggressive	Total
Substandard	24	15	17	56
Standard	52	73	80	205
Modern	58	86	36	180
Total	134	174	133	441

- The *expected* frequencies are -

Status	Aggressive	Neutral	Non-aggressive	Total
Substandard	17.02	22.10	16.89	56
Standard	62.29	80.88	61.83	205
Modern	54.69	71.02	54.29	180
Total	134	174	133	441

- The minimum expected count is 16.89 so the expected counts are sufficiently large.

Test of Independence

$$\chi^2 = \frac{(24 - 17.02)^2}{17.02} + \cdots + \frac{(36 - 54.29)^2}{54.29} = 22.47.$$

- $df = 2 \times 2 = 4$, and the p-value is $P(\chi^2_4 > 22.47) = 0.00016$.
- The p-value is far below the significance level so we easily reject H_0 at a 1% significance level.
- There is very strong evidence that the distribution of facility conditions is not the same across gasoline stations with different pricing policies. The observed distributions of facility conditions at stations in each pricing group is shown below.

Status	Aggressive	Neutral	Non-aggressive
Substandard	.179	.086	.128
Standard	.388	.420	.602
Modern	.433	.494	.271
Total	1.00	1.00	1.00