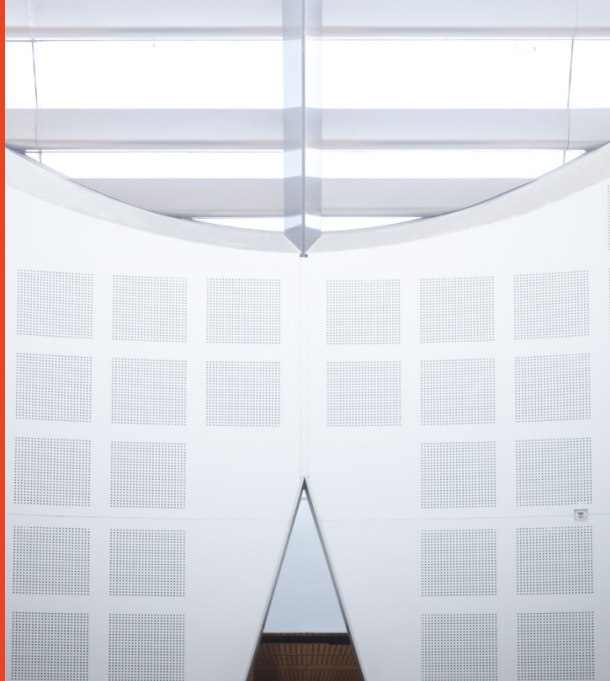
Test for Goodness of Fit (Chi-squared Test)



THE UNIVERSITY OF SYDNEY

Interesting Facts about the Chi-Squared Distribution

Link

For the Chi-squared test, we are going to use the Chi-Squared Distribution. Recall from Lecture 3:

Definition (Chi-Squared distribution)

The Chi-Squared distribution is the sum of squared independent Standard Normal random variables $Z_i \sim N(0,1)$ $i = 1, 2, \ldots, n$. It can only take positive values and is typically right skewed.

We say the variable $X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$ with n degrees of freedom, and mean $E(X) = n$ and variance $Var(X) = 2n$.

The pdf is:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \qquad \text{for } x \in (0, \infty)$$

Gamma distribution

- $X \sim \Gamma(\alpha, \beta)$. The probability density function is given by
- $f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}$ for $x > 0$ where $\alpha > 0$ and $\beta > 0$.

- $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$

- Chi-square distribution is a special case of the gamma distribution when $\alpha = \frac{k}{2}$ and $\beta = 2$; i.e., $\Gamma\left(\alpha = \frac{k}{2}, \beta = 2\right) = \chi^2_k$

- If $Q \sim \chi^2_k$, $aQ \sim \Gamma\left(\alpha = \frac{k}{2}, \beta = 2a\right)$ for $a > 0$.

Chi-square test for independence

- χ^2-test is a very flexible test because it can be applied for a range of situations.
  - Fitting discrete probability to categories of counts (Mendel genetics)
  - Fitting models for continuous data based using discrete intervals (Random number generator)
- We will introduce a final example, which looks at how to establish independence between two discrete variables. i.e. We test on the null hypothesis H_0 : variables are independent vs. H_A : variables are not independent.
- Let's begin with an example: imagine we have two variables in which we can tabulate counts of smoking and rate of lung cancer.

Table: Contingency table of smoking status and lung cancer.

	No smoking	Smoking	Total
No lung cancer	200	1400	1600
Lung cancer	100	1300	1400
Total	300	2700	3000

Deriving the probabilities under independence

Table: We then calculate the proportions of the margins of the contingency table.

	No smoking	Smoking	Total
No lung cancer	200	1400	$1600/3000 = 0.53$
Lung cancer	100	1300	$1400/3000 = 0.47$
Total	$300/3000 = 0.1$	$2700/3000 = 0.9$	1

- Recall that for two events to be independent, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Table: We then complete the table of probability assuming independence

	No smoking	Smoking	Total
No lung cancer	0.53*0.1	0.53*0.9	0.53
Lung cancer	0.47*0.1	0.47*0.9	0.47
Total	0.1	0.9	1

Deriving the expected counts

	No smoking	Smoking	Total
No lung cancer	3000*0.53*0.1 = 159	3000*0.53*0.9 = 1431	1600
Lung cancer	3000*0.47*0.1 = 141	3000*0.47*0.9 = 1269	1400
Total	300	2700	3000

- Then $\tau = \frac{200^2}{159} + \frac{1400^2}{1431} + \frac{100^2}{141} + \frac{1300^2}{1269} - 3000 = 23.92.$
- $\tau \sim_{H_0} \chi^2_{(r-1)\times(c-1)}$, with r being the number of rows and c being the number of colums in the contingency table.
- Thus, the p-value can be calculated as:

$$\mathbb{P}(\chi_1^2 \geq \tau) = (\chi_1^2 \geq 23.92) < 0.05$$

- Hence, we reject the null hypothesis that smoking status and lung cancer are independent.

Table: Contingency table of smoking status and lung cancer.

	No smoking	Smoking	Total
No lung cancer	200	1400	1600
Lung cancer	100	1300	1400
Total	300	2700	3000

```
> tabel1 <- matrix(c(200,100,1400,1300), ncol=2)
> colnames(table1) <- c("No smoking", "Smoking")
> rownames(table1) <- c("No lung cancer", "Lung cancer")
> table1
               No smoking Smoking
No lung cancer        200    1400
Lung cancer           100    1300
> chisq <- chisq.test(table1, correct=FALSE); chisq

        Pearson's Chi-squared test

data:  table1
X-squared = 23.81, df = 1, p-value = 1.064e-06

> chisq$expected     # expected frequencies
               No smoking Smoking
No lung cancer        160    1440
Lung cancer           140    1260
```

Steps for the Chi-Square Goodness-of-Fit Test

Context Consider a set of categorical data with g categories in which fall observed counts O_i, for $i = 1, 2, \ldots, g$. A probability model is proposed for the categories, and we want to test whether it is adequate.

Preparation Construct the following table:

Class	1	2	3	...	g	Totals
Observed Counts	O_1	O_2	O_3	...	O_g	$\sum_{i=1}^{g} O_i = n$
Expected Counts	E_1	E_2	E_3	...	E_g	$\sum_{i=1}^{g} E_i = n$

Notes:

- O_i are given, and E_i need to be worked out from the hypothesised model H_0, so $E_i = nP(category\ \ i)$.
- Sometimes we need to estimate k parameter(s) of the model first, before we can work out E_i.

H H_0: Model fits. vs H_1: Model doesn't fit.

A Cochran's Rule: Check that $E_i \geq 1$ and no more than 20% of E_i are less than 5. If some of the E_i are too small, then we combine categories together.

T

- Definition Formula: $\tau = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{g-k-1}$ (under H_0)
- Calculation Formula: $\tau = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n \sim \chi^2_{g-k-1}$ (under H_0)
- Large values of τ will argue against H_0 for H_1.
  (This indicates a difference between O_i and E_i.)
- The observed value is τ_0.

P P-value $= P(\chi^2_{g-k-1} \geq \tau_0)$.

C Weigh up the P-value.

Example 1: Mendel's Early Genetics Model

Mendel did much work in early genetics in the 19th Century, but it wasn't appreciated until later. He conducted experiments on the distributions of traits in pea plants. In one experiment, he classified 556 peas according to shape (Round or Angular) and colour (Yellow or Green). He predicted that the 4 different 'offspring' (RY, RG, AY, AG) would occur in the ratio 9:3:3:1. He observed counts of 315, 108, 101 and 32.

Does Mendel's theory fit the data?

Example 1: Mendel's Early Genetics Model

Preparation Construct the following table:

Class	RY	RG	AY	AG	Totals
Observed Counts	315	108	101	32	556
Expected Counts	312.75	104.25	104.25	34.75	556

where:

$E_1 = \frac{9}{9+3+3+1} \times 556 = \frac{9}{16} \times 556 = 312.75$

$E_2 = E_3 = \frac{3}{16} \times 556 = 104.25.$

$E_4 = \frac{1}{16} \times 556 = 34.75.$

So the parameters are: $g = 4, k = 0.$

Example 1: Mendel's Early Genetics Model

Preparation Construct the following table:

Class	RY	RG	AY	AG	Totals
Observed Counts	315	108	101	32	556
Expected Counts	312.75	104.25	104.25	34.75	556

where:

$E_1 = \frac{9}{9+3+3+1} \times 556 = \frac{9}{16} \times 556 = 312.75$

$E_2 = E_3 = \frac{3}{16} \times 556 = 104.25.$

$E_4 = \frac{1}{16} \times 556 = 34.75.$

So the parameters are: $g = 4, k = 0.$

H H_0: Model 9:3:3:1 fits. vs H_1: Model doesn't fit.

A Cochran's Rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.

T

- Calculation Formula: $\tau = \sum_{i=1}^{4} \frac{O_i^2}{E_i} - 556 \sim \chi_3^2$ (under H_0)
- Large values of τ will argue against H_0 for H_1, as this indicates a difference between O_i and E_i.)
- The observed value is $\tau_0 = \frac{315^2}{312.75} + \frac{108^2}{104.25} + \frac{101^2}{104.25} + \frac{32^2}{34.75} - 556 \approx 0.47$.

```
> o=c(315,108,101,32)
> e=c(312.75,104.25,104.25,34.75)
> sum((o-e)^2/e)

[1] 0.470024

> sum(o^2/e) - 556

[1] 0.470024
```

H H_0: Model 9:3:3:1 fits. vs H_1: Model doesn't fit.

A Cochran's Rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.

T

- Calculation Formula: $\tau = \sum_{i=1}^{4} \frac{O_i^2}{E_i} - 556 \sim \chi_3^2$ (under H_0)
- Large values of τ will argue against H_0 for H_1, as this indicates a difference between O_i and E_i.)
- The observed value is $\tau_0 = \frac{315^2}{312.75} + \frac{108^2}{104.25} + \frac{101^2}{104.25} + \frac{32^2}{34.75} - 556 \approx 0.47$.

```
> o=c(315,108,101,32)
> e=c(312.75,104.25,104.25,34.75)
> sum((o-e)^2/e)

[1] 0.470024

> sum(o^2/e) - 556

[1] 0.470024
```

P P-value $= P(\chi^2_3 \geq 0.47) > 0.25$ using tables.

```
> 1-pchisq(0.47,3)
[1] 0.9254311
```

C As the P-value is so large, the data is consistent with Mendel's model.

```
> offspring <- c(315, 108, 101, 32)
> chisq <- chisq.test(offspring, p = c(9/16, 3/16, 3/16, 1/16))
> chisq

        Chi-squared test for given probabilities

data:  offspring
X-squared = 0.47002, df = 3, p-value = 0.9254
```

Example 2: Random Number Generator

Suppose a Random Number Generator (producing numbers in $[0, 1]$ with equal likelihood) is being tested. 1,000 values yield the following results.

Cat.	[0,0.1)	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,0.6)	[0.6,0.7)	[0.7,0.8)
O_i	136	105	107	89	97	84	76	84

Cat.	[0.8,0.9)	[0.9,1]	Total
O_i	105	117	1000

Test the goodness of fit of the U[0,1] model to these counts.

Example 2: Random Number Generator

Preparation Construct the following table:

Cat.	[0,0.1)	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,0.6)	[0.6,0.7)	[0.7,0.8)
O_i	136	105	107	89	97	84	76	84
E_i	100	100	100	100	100	100	100	100

Cat.	[0.8,0.9)	[0.9,1]	Total
O_i	105	117	1000
E_i	100	100	1000

as $E_i = 1000/10 = 100$ for $i = 1, 2, \ldots, 10$.

So the parameters are: $g = 10, k = 0$.

H H_0: U[0,1] Model fits. vs H_1: Model doesn't fit.

A Cochran's Rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.

T

- Calculation Formula: $\tau = \sum_{i=1}^{10} \frac{O_i^2}{E_i} - 1000 \sim \chi_9^2$ (under H_0)
- Large values of τ will argue against H_0 for H_1, as this indicates a difference between O_i and E_i.)
- The observed value is $\tau_0 = \frac{136^2}{100} + \frac{105^2}{100} + \ldots \frac{117^2}{100} - 1000 = 29.02$.

```
> o=c(136,105,107,89,97,84,76,84,105,117)
> e=c(100,100,100,100,100,100,100,100,100,100)
> sum((o-e)^2/e)

[1] 29.02

> sum(o^2/e) - 1000

[1] 29.02
```

P P-value $= P(\chi^2_9 \geq 29.02) < 0.01$ using tables.

```
> 1-pchisq(29.02,9)

[1] 0.0006430267
```

C As the P-value is so small, the data is not consistent with random number generator.

Example 3: Testing Data fits a Binomial Model

Data results in the following frequency table and plot.

Category	0	1	2	3	Total
O_i	19	34	27	20	100

Test whether the data could be modelled by $Bin(n, p)$ for some p.

Example 3: Testing Data fits a Binomial Model

Preparation

(1) Fit parameters

In order to fit a Binomial model, we need the 2 parameters n and p. Given the outcomes $0, 1, 2, 3$, we have $n = 3$, but p is not given, so we need to estimate it from the data using the formula

$$\hat{p} = \frac{0 \times 19 + 1 \times 34 + 2 \times 27 + 3 \times 20}{3 \times 100} \approx 0.493$$

The formula arises because 100 Bin(3,p) trials is equivalent to 300 Bernoulli(p) trials.

```
> x=c(0,1,2,3)
> o=c(19,34,27,20)
> sum(x*o)/(3*sum(o))
[1] 0.4933333
```

(2) Construct the following table:

Category	0	1	2	3	Total
O_i	19	34	27	20	100
E_i	13.03	38.02	36.97	11.98	100

as $E_i = \binom{3}{i}(0.493)^i(1-0.493)^{3-i} \times 100$ for $i = 0, 1, 2, 3$.

```
> round(dbinom(x,3,0.493)*100, 2)
[1] 13.03 38.02 36.97 11.98
```

So the parameters are: $g = 4, k = 1$.

H H_0: Bin(3,p) Model fits. vs H_1: Model doesn't fit.

A Cochran's Rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.

T

- Calculation Formula: $\tau = \sum_{i=0}^{3} \frac{O_i^2}{E_i} - 100 \sim \chi^2_{4-1-1} = \chi^2_2$ (under H_0)
- Large values of τ will argue against H_0 for H_1, as this indicates a difference between O_i and E_i.
- The observed value is $\tau_0 = \frac{19^2}{13.03} + \ldots \frac{20^2}{11.98} - 100 \approx 11.2$.

```
> o=c(19,34,27,20)
> e=c(13,38,37,12)
> sum((o-e)^2/e)

[1] 11.22632

> sum(o^2/e) - 100

[1] 11.22632
```

P P-value $= P(\chi^2_2 \geq 11.2) < 0.01$ using tables.

```
> 1-pchisq(11.2,2)
[1] 0.003697864
```

C As the P-value is so small, the data is not consistent with a Bin(3,p) model.

Example 4

Turner considered 1000 bags each of 10 oranges. He counted the number of rotten oranges in each bag and obtained the numbers below. Are these results consistent with a binomial distribution for the number of rotten oranges per bag?

Number rotten	0	1	2	3	4	5	6
Observed	334	369	191	63	22	12	9

$$P = \frac{\text{number rotten}}{\text{total oranges}} = \frac{0*334 + 1*369 + 2*191 + \ldots + 6*9}{10000} = 0.1142$$

Example 4

P = 0.1142

Number rotten	Observed		Probability	Expected
0	334	P(X = 0)	0.2974	297.4108
1	369	P(X = 1)	0.3834	383.4310
2	191	P(X = 2)	0.2224	222.4488
3	63	P(X = 3)	0.0765	76.4767
4	22	P(X = 4)	0.0173	17.2543
5	12	P(X = 5)	0.0027	2.6694
6	9	P(X ≥ 6)	0.0003	0.3090
	1000			

Cochran's rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.

$\frac{2}{7}$ or 28.57% of E_i are less than 5. So, we combine the last 2 rows.

Example 4

H_0: The sample comes from a binomial distribution.

H_1: The sample does not come from a binomial distribution.

Number rotten	Observed		Probability	Expected	$\frac{(o - e)^2}{e}$
0	334	P(X = 0)	0.2974	297.4108	4.5014
1	369	P(X = 1)	0.3834	383.4310	0.5431
2	191	P(X = 2)	0.2224	222.4488	4.4461
3	63	P(X = 3)	0.0765	76.4767	2.3749
4	22	P(X = 4)	0.0173	17.2543	1.3053
5 or more	21	P(X ≥ 5)	0.0030	2.9783	109.0474
	1000			1000	
				Test statistic	122.2182
				p-value	0.000

Reject H_0 and conclude that the counts of rotten oranges do not follow a binomial distribution.

Chi-squared test for normality

- Since a normal distribution has numerical data, we must begin by subdividing the range of the normal distribution into a set of intervals, or categories, in order to obtain nominal data.
- In practice, you can use any intervals you like to facilitate the calculation of the normal probabilities.
- The number of intervals chosen should comply with the rule that all expected values be ≥ 5. Because the number of degrees of freedom is $g - 3$, the minimum number of intervals is $g = 4$.

Choosing class intervals

- If the sample size is ≤ 80, the intervals are:

Interval	Probability
$Z \leq 1$	0.1587
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$Z > 1$	0.1587

- when the sample size is < 32, at least one expected value will be less than 5.

- If the sample size is ≤ 220 and > 80, the intervals are:

Interval	Probability
$Z \leq -1.5$	0.0668
$-1.5 < Z \leq -0.5$	0.2417
$-0.5 < Z \leq 0.5$	0.3829
$0.5 < Z \leq 1.5$	0.2417
$Z > 1.5$	0.0668

- If the sample size is more than 220, the intervals are:

Interval	Probability
$Z \leq -2$	0.0228
$-2 < Z \leq -1$	0.1359
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$1 < Z \leq 2$	0.1359
$Z > 2$	0.0228

Example 5

Test the hypothesis that the following sample of data is drawn from a normal population. The sample mean and the standard deviation are 34 and 12, respectively (Use $\alpha = 0.01$)

Class	Frequency
10 up to 20	15
20 up to 30	24
30 up to 40	30
40 up to 50	18
50 up to 60	13

Example 5

$\overline{X}$ = 34 and s = 12

H_0: The data are normally distributed.

H_1: The data are not normally distributed.

Class		Probability	Observed	Expected	$\frac{(o-e)^2}{e}$
10 up to 20	P(X < 20)	0.1217	15	12.1673	0.6595
20 up to 30	P(20 ≤ X < 30)	0.2478	24	24.7769	0.0244
30 up to 40	P(30 ≤ X < 40)	0.3220	30	32.2021	0.1506
40 up to 50	P(40 ≤ X < 50)	0.2173	18	21.7326	0.6411
50 up to 60	P(50 ≤ X < 60)	0.0912	13	9.1211	1.6495
			100		
				Test statistic	3.1251
				P-value	0.2096

Retain H_0 and cannot conclude that the data are not normally distributed.

Example 6

The following observations were drawn from a large population.

22 18 25 28 19 20 24 26 19 26 27 22 23 25 25 18 20 26 18 26

27 24 20 19 18 17 13 14 16 10 15 16 14 13 18

Test if the above sample of data is drawn from a normal population.

Example 6

$\overline{X}$ = 20.31429 and s = 4.806683

H_0: The data are normally distributed.

H_1: The data are not normally distributed.

	Probability	Expected	Observed	$\frac{(o - e)^2}{e}$
P(Z ≤ -1)	0.1587	5.5529	6	0.0360
P(-1 < Z ≤ 0)	0.3413	11.9471	14	0.3528
P(0 < Z ≤ 1)	0.3413	11.9471	8	1.3040
P(Z > 1)	0.1587	5.5529	7	0.3371
			35	
			Test statistic	2.0669
			P-value	0.1502

Retain H_0 and cannot conclude that the data are not normally distributed.