Topic11: Test for Goodness of Fit (Chi-squared Test)

Outline

Topic11: Test for Goodness of Fit (Chi-squared Test)

- Example1: Mendel's Early Genetics Model
- Example2: Random Number Generator
- Example3: Testing Data fits a Binomial Model
- Example4: Testing Data fits a Normal Model
- Interesting Facts about the Chi-Squared Distribution
- Steps for the Chi-Squared Test
- Worked Examples

Example1: 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?

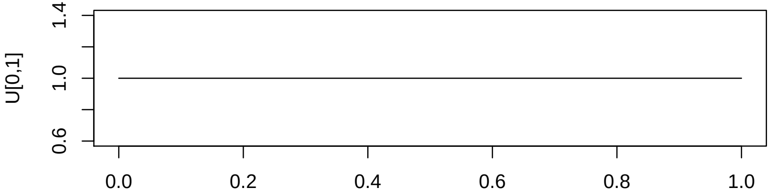
Example2: Random Number Generator

Suppose a Random Number Generator is being tested. 1000 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

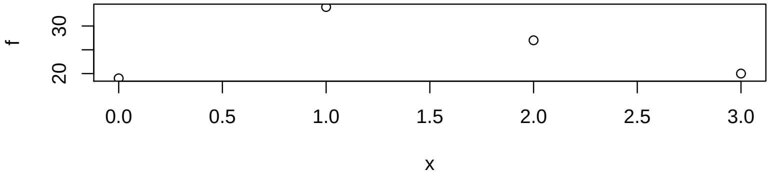
Test the goodness of fit of the U[0,1] model to these counts.



Example3: 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

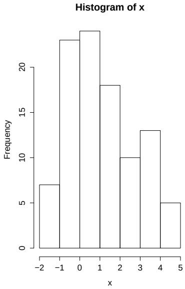


Test whether the data could be modelled by $Bin(n, p)$ for some p.

Putting this example into context, imagine a sports journalist claims that Michael Jordan's free throws follow a Binomial distribution with probability 80%. ▸ Sports Example ▸ Larry Bird highlights

Example4: Testing Data fits a Normal Model

Data results in the following histogram.



Test whether the following histogram could be modelled by a Normal distribution $N(\mu, \sigma^2)$?

Interesting Facts about the Chi-Squared Distribution

For the Goodness of Fit test, we are going to use the Chi-Squared Distribution. Recall from Chapter 5:

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.

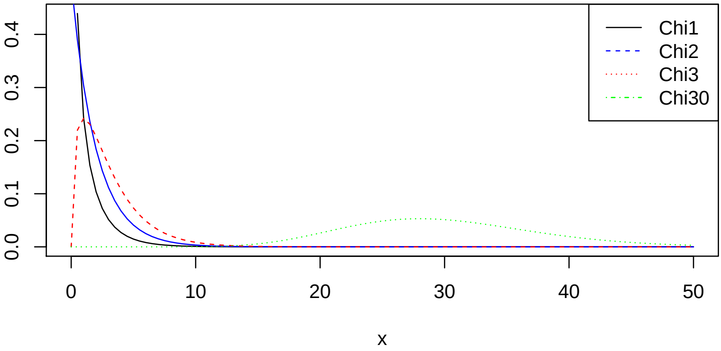
We say the variable $X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$ with n degrees of freedom, and mean $E(X) = n$ and variance $Var(X) = 2n$.

The pdf is:

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

▸ Link

11



Steps for the Chi-Squared Test

$\boxed{\text{Context}}$ 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.

$\boxed{\text{Preparation}}$ 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: 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:

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

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

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

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

$\boxed{\text{H}}$ H_0: Model 9:3:3:1 fits. vs H_1: Model doesn't fit.

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

$\boxed{\text{T}}$

- 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
```

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

```
1-pchisq(0.47,3)

## [1] 0.9254311
```

$\boxed{\text{C}}$ As the P-value is so large, the data is consistent with Mendel's model.

Example2: 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)	[0.8,0.9)	[0.9,1]	Total
O_i	136	105	107	89	97	84	76	84	105	117	1000
E_i	100	100	100	100	100	100	100	100	100	100	1000

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

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

$\boxed{\text{H}}$ H_0: U[0,1] Model fits. vs H_1: Model doesn't fit.

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

$\boxed{\text{T}}$

- 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
```

11

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

```
1-pchisq(29.02,9)

## [1] 0.0006430267
```

$\boxed{\text{C}}$ As the P-value is so small, the data is not consistent with random number generator.

Example3: 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 $i = 0, 1, 2, 3$.

```
dbinom(x,3,0.493)*100

## [1] 13.03238 38.01755 36.96775 11.98232
```

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

$\boxed{\text{H}}$ H_0: Bin(3,p) Model fits. vs H_1: Model doesn't fit.

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

$\boxed{\text{T}}$

- 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
```

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

```r
1-pchisq(11.2,2)

## [1] 0.003697864
```

$\boxed{\text{C}}$ As the P-value is so small, the data is not consistent with a Bin(3,p) model.

Extension: $X \sim \chi^2_2 = \exp(-\frac{x}{2})$, where

$$P(X \geq x) = e^{-\frac{x}{2}}$$

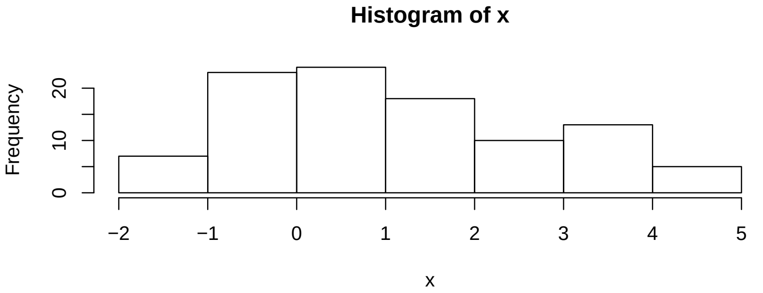
so the exact P-value is

$$P(\chi^2_2 \geq 11.2) = e^{-\frac{11.2}{2}} = 0.003697864$$

Example4: Testing Data fits a Normal Model

This is a harder problem, as to fit a Normal distribution we need to estimate the 2 parameters: μ and σ^2.

We present a summarised solution, and then show how it could be done in R (Extension).



Preparation

(1) To fit the 2 parameters, we divide the continuous data into categories, by finding out how many observations fit in each class of the histogram.

Class	[-2,-1)	[-1,0)	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	Total
O_i	7	23	24	18	10	13	5	100

To work out the mean, we use the midpoint of each class as an estimate of the obervations in that class. So the 1st class has 7 observations in [-2,-1) each approximated by -1.5.

Hence the mean estimate is

$$\hat{\mu} = \frac{7 \times (-1.5) + 23 \times (-0.5) + 24 \times (0.5) + 18 \times 1.5 + 10 \times 2.5 + 13 \times 3.5 + 5 \times 4.5}{100} = 1.1$$

Similarly, we can work out the variance

$$\hat{\sigma^2} = \frac{7 \times (-1.5)^2 + 23 \times (-0.5)^2 + 24 \times (0.5)^2 + \ldots + 5 \times 4.5^2 - 100 \times (1.1)^2}{99} = 2.727273$$

So

$$\hat{\sigma} = \sqrt{2.727273} \approx 1.65$$

Therefore, for fitting the Normal we use the 2 parameters: $\mu = 1.1$ and $\sigma = 1.65$.

Note that we based these estimates on the grouped data, not the original ungrouped data. If we used the ungrouped data we would get a larger statistic and so a misleading small P-value.

(2) Work out the Expected Values for $N(1.1, 1.65^2)$.

Note: We now change the first interval to $(-\infty, -1)$ and last to $[4, \infty)$, to be consistent with the Normal which spans $(-\infty, \infty)$.

We construct the following table:

Class	(-∞,-1)	[-1,0)	[0,1)	[1,2)	[2,3)	[3,4)	[4,∞)	Total
O_i	7	23	24	18	10	13	5	100
E_i	10.16	15.09	22.33	23.14	16.80	8.54	3.94	100

where the Expected Counts in (-∞,-1) $=$ 10.16

```
pnorm(-1,1.1,1.65)*100

## [1] 10.15574
```

and the Expected Counts in [-1,0) = 15.09

```
(pnorm(0,1.1,1.65)-pnorm(-1,1.1,1.65))*100

## [1] 15.09351
```

Other values follow:

```
(pnorm(1,1.1,1.65)-pnorm(0,1.1,1.65))*100

## [1] 22.33439

(pnorm(2,1.1,1.65)-pnorm(1,1.1,1.65))*100

## [1] 23.14431

(pnorm(3,1.1,1.65)-pnorm(2,1.1,1.65))*100

## [1] 16.79603
```

11

```
(pnorm(4,1.1,1.65)-pnorm(3,1.1,1.65))*100

## [1] 8.535032

(1-pnorm(4,1.1,1.65))*100

## [1] 3.940986
```

So the parameters are: $g = 7, k = 2$.

$\boxed{\text{H}}$ H_0: $N(1.1, 1.65^2)$ Model fits. vs H_1: Model doesn't fit.
$\boxed{\text{A}}$ Cochran's Rule: All $E_i \geq 1$ and no more than 20% of E_i are less than 5.
$\boxed{\text{T}}$

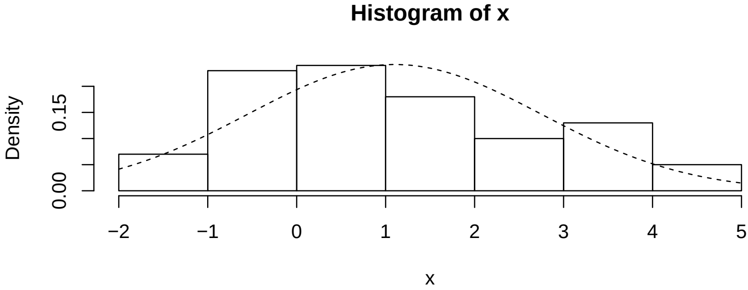
- Calculation Formula: $\tau = \sum_{i=1}^{7} \frac{O_i^2}{E_i} - 100 \sim \chi^2_{7-1-2} = \chi^2_4$ (under H_0)
- Large values of τ will argue against H_0 for H_1, as this indicates a difference between O_i and E_i.)
- The observed value is $\tau_0 = \frac{7^2}{10.16} + \ldots \frac{5^2}{3.94} - 100 \approx 11.76$.

$\boxed{\text{P}}$ P-value $= P(\chi^2_4 \geq 11.76) \in (0.01, 0.025)$ using tables.

```
1-pchisq(11.76,4)
```

```
## [1] 0.01922812
```

$\boxed{\text{C}}$ As the P-value is small, the data are not consistent with a Normal model, as evident from the Normal curve on the histogram.

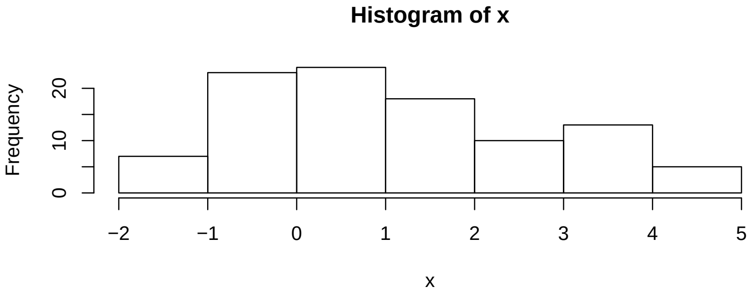


Solving in R (Extension)

Note: These results differ a tiny bit, as previously we rounded the estimate of standard deviation to 2dp.

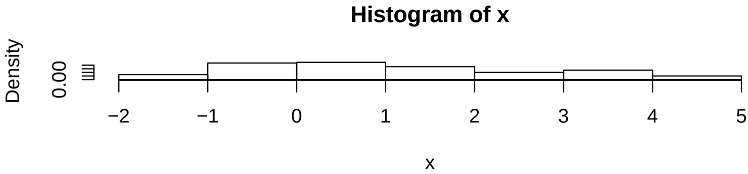
(1) Scan data and produce histogram

```
x <- read.table("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/w13.txt")
x=unlist(x) #Changes data frame to vector
hist(x)
```



(2) Find the Observed Counts, from a histogram of the data.

```
#Lists the class breaks (-2,-1], (-1,0] etc
hist(x,pr=T)$breaks

## [1] -2 -1  0  1  2  3  4  5

freq=hist(x,pr=T)$counts  #Observed Counts O_i
```



```
freq

## [1]  7 23 24 18 10 13  5
```

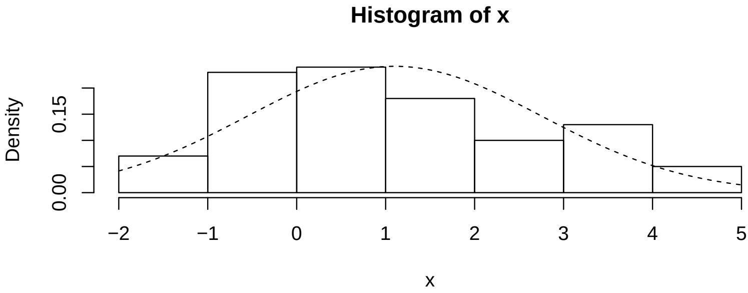
(3) Estimate the 2 parameters of Normal: mean and sd.

```
mids=(-2:4)+.5     #Midpoints of each class
mids

## [1] -1.5 -0.5  0.5  1.5  2.5  3.5  4.5

gr.sum=sum(freq*mids)
gr.sum

## [1] 110

gr.sumsq=sum(freq*mids^2)
gr.sumsq

## [1] 391
```

```
gr.mean=gr.sum/100   #Estimate of mean
gr.mean

## [1] 1.1

gr.var=1/99* (gr.sumsq - 1/100* gr.sum^2)
gr.var

## [1] 2.727273

gr.sd=sqrt(gr.var)   #Estimate of sd
gr.sd

## [1] 1.651446
```

(4) Find the Expected Counts, from fitting a $N(1.1, 1.651446^2)$ model.

```
#Add Normal PDF to the histogram
hist(x, pr=T)
curve(dnorm(x,m=gr.mean,s=gr.sd),lty=2,add=T)
```



```
#Normal probability at lower threshold of each interval
lower.probs=pnorm(-1:4,m=gr.mean,s=gr.sd)
lower.probs

## [1] 0.1017553 0.2526790 0.4758576 0.7071154 0.8750325 0

#The probability in each interval (upper-lower)
exp.probs=diff(c(0,lower.probs,1))
exp.probs

## [1] 0.10175530 0.15092370 0.22317860 0.23125775 0.16791
## [7] 0.03954103

#Find Expected frequencies/counts E_i
exp.freq= 100* exp.probs
exp.freq

## [1] 10.175530 15.092370 22.317860 23.125775 16.791712
```

(5) Calculate the chi-squared test statistic, and P-value.

```
#Find Chi-squared contributions
contrib = ((exp.freq-freq)^2)/exp.freq
contrib

## [1] 0.9910041 4.1431939 0.1267861 1.1361164 2.7470308 2

# Put O-i, E_i, Chi-Sq in table
cbind(freq,exp.freq,contrib)

##      freq  exp.freq   contrib
## [1,]    7 10.175530 0.9910041
## [2,]   23 15.092370 4.1431939
## [3,]   24 22.317860 0.1267861
## [4,]   18 23.125775 1.1361164
## [5,]   10 16.791712 2.7470308
## [6,]   13  8.542650 2.3257383
## [7,]    5  3.954103 0.2766496
```

11

```
#Calculate Chi-squared test statistic
tau.obs=sum(((exp.freq-freq)^2)/exp.freq)
tau.obs

## [1] 11.74652

#Calculate P-value
1-pchisq(tau.obs, df=length(freq)-2-1)

## [1] 0.0193392
```