Goodness of Fit Test

Context	A total of n observed frequencies over g classes and a proposed probability model needing k estimated parameters
Hypothesis	H_0 : Model fits data
Test Statistic	$\tau = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - n \overset{H_0}{\sim} \chi^2_{g-k-1}$

1. Plant Genotypes (no parameters estimated)

A sample of 100 plants have genotypes A, B, and C occuring with the frequencies 18, 55 and 27 respectively. We are interested in the null hypothesis that A, B, and C are in the ratio of 1:2:1.

(a) Preparation: Fill out the following table

Genotype	A	B	C	Total
Observed frequency, O_i	18	55	27	100
Expected frequency, E_i				100

(b) Hypothesis: State H_0 and H_1.

(c) Assumptions: What are the assumptions for a χ^2 test and are they valid here?

(d) Test statistic: What is the test statistic and its distribution under H_0?

(e) P-value: Calculate the p-value using the χ^2 tables.

```
1-pchisq(2.62,2)

## [1] 0.2698201
```

(f) Conclusion: Draw your conclusion based on the p-value.

Solution

(a) $\boxed{\text{Preparation}}$

Genotype	A	B	C	Total
Observed frequency, O_i	18	55	27	100
Expected frequency, E_i	25	50	25	100

(b) $\boxed{\text{H}}$ H_0 : The Genotypes A, B and C occur in the ratio $1:2:1$ vs H_1 : Not H_0.

(c) $\boxed{\text{A}}$ We need $E_i \geq 1$ (true here), and no more than 20% of $E_i < 5$ (Cochran's Rule - also true here).

(d) $\boxed{\text{T}}$ $\tau = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - 100 \overset{H_0}{\sim} \chi^2_{g-k-1} = \chi^2_{3-0-1} = \chi^2_2$

Large values of τ argue against H_0 for H_1 (as τ gets large if any large gaps between O_i and E_i).

Observed value: $\tau_0 = \frac{18^2}{25} + \frac{55^2}{50} + \frac{27^2}{25} - 100 = 2.62.$

(e) $\boxed{\text{P}}$ P-value $= P(\chi_2^2 > 2.62) > 0.25.$

```
1-pchisq(2.62,2)
```

(f) $\boxed{\text{C}}$ Given $P - value$ is large, data supports H_0 and we conclude that the sample is consistent with Genotypes in the ratio 1:2:1.

2. Model Fit (no parameters estimated)

100 observations are made on a random variable only taking values 0, 1, 2 and 3. The frequencies are shown below:

Value	0	1	2	3
Frequency	22	38	32	8

A goodness of fit test is applied to see if these frequencies are well-described by $\mathcal{B}(3, 0.5)$ probabilities.

(a) Show that the χ^2 goodness-of-fit statistic is 9.653.

Value	0	1	2	3
O_i	22	38	32	8
E_i	12.5			

where $E_i = 100\binom{3}{i}0.5^i 0.5^{3-i}$.

(b) Show that the corresponding p-value is somewhere in the interval (0.01, 0.025).

(c) What is your conclusion?

Solution

$\boxed{\text{Preparation}}$

Value	0	1	2	3	Total
Frequency O_i	22	38	32	8	100
E_i	12.5	37.5	37.5	12.5	100

where $E_i = 100 \times p_i$, where p_i is probability from $Bin(3, 0.5)$.

So $E_i = 100 \times \binom{3}{i}(0.5)^i(0.5)^{3-i}$.

Eg $E_1 = 100 \times \binom{3}{0}(0.5)^0(0.5)^3 = 12.5 = E_3$.

$\boxed{\text{H}}$ H_0 : These frequencies are well-described by $\mathcal{B}(3, 0.5)$ probabilities. vs H_1 : Not H_0.

$\boxed{\text{A}}$ We need $E_i \geq 1$ (true here), and no more than 20% of $E_i < 5$ (Cochran's Rule - also true here).

(a)

$\boxed{\text{T}}$ $\tau = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - 100 \overset{H_0}{\sim} = \chi^2_{4-0-1} = \chi^2_3$

Large values of τ argue against H_0 for H_1 (as τ gets large if any large gaps between O_i and E_i).

Observed value: $\tau_0 = \frac{22^2}{12.5} + \frac{38^2}{37.5} + \ldots \frac{8^2}{12.5} - 100 = 9.653.$

(b $\boxed{\text{P}}$ Using Table: P-value $= P(\chi^2_3 > 9.653) \in (0.01, 0.025)$.

```
1-pchisq(9.653,3)

## [1] 0.02175819
```

(c) $\boxed{\text{C}}$ Given $P - value$ is small, we reject H_0 and conclude that the $\mathcal{B}(3, 0.5)$ is not a good fit for data.

3. Model Fit (1 parameter estimated)

For the previous question, we want to see if another Binomial distribution might explain the frequencies better.

(a) Preparation: Estimate p using

$$\hat{p} = \text{overall proportion of successes} = \frac{(0)(22) + (1)(38) + (2)(32) + (3)(8)}{(3)(100)}$$

(b) Preparation: Fill out the frequencies

Value	0	1	2	3	Total
Observed frequency, O_i	22	38	32	8	100
Expected frequency, E_i	19.51				100

where $E_i = 100\binom{3}{i}(\hat{p})^i(1-\hat{p})^{3-i}$.

(c) Test the hypothesis that the frequencies are well-described by $\mathcal{B}(3, \hat{p})$ probabilities.

```
1-pchisq(0.88,2)
```

Solution

(a) $\boxed{\text{Preparation}}$

Estimate p using

$$\hat{p} = \text{overall proportion of successes} = \frac{(0)(22) + (1)(38) + (2)(32) + (3)(8)}{(3)(100)} = 0.42$$

Hence: $k = 1$.

(b) $\boxed{\text{Preparation}}$

Value	0	1	2	3	Total
Observed frequency, O_i	22	38	32	8	100
Expected frequency, E_i	19.51	42.39	30.69	7.41	100

where $E_i = 100\binom{3}{i}(0.42)^i(1-0.42)^{3-i}$.

Eg $E_0 = 100\binom{3}{0}(0.42)^0(1-0.42)^3 = 19.5112$

```r
100*dbinom(0,3,0.42)

## [1] 19.5112

100*dbinom(1,3,0.42)

## [1] 42.3864
```

(c) $\boxed{\text{H}}$ H_0 : These frequencies are well-described by $\mathcal{B}(3, 0.42)$ probabilities. vs H_1 : Not H_0.

$\boxed{\text{A}}$ We need $E_i \geq 1$ (true here), and no more than 20% of $E_i < 5$ (Cochran's Rule - also true here).

$\boxed{\text{T}}$ $\tau = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - 100 \overset{H_0}{\sim} = \chi^2_{4-1-1} = \chi^2_2$

Note: $k = 1$ as we had to estimate 1 parameter $\hat{p}$.

Large values of τ argue against H_0 for H_1.
Observed value: $\tau_0 = \frac{22^2}{19.51} + \frac{38^2}{42.39} + \ldots \frac{8^2}{7.41} - 360 = 0.8753231$.

$\boxed{\text{P}}$

Using Table, P-value $= P(\chi^2_2 > 0.88) > 0.25$.

```r
1-pchisq(0.88,2)

## [1] 0.6440364
```

$\boxed{\text{C}}$ Given $P - value$ is so large, data strongly supports H_0 and we conclude that the $\mathcal{B}(3, 0.42)$ seems a good fit for thedata.

Extra Questions

4. NSW Fatal Accidents 1993 (no parameters estimated)

The number of fatal accidents on NSW roads in months with 31 days in 1993 were:

Jan	Mar	May	July	Aug	Oct	Dec
44	56	37	42	59	59	63

Test the claim that the accident rate is the same for all months.

Hint: Show that the $\tau = 12.09$ and p-value is close to 0.05.

Solution

fboxPreparation

Month	Jan	Mar	May	July	Aug	Oct	Dec	Total
O_i	44	56	37	42	59	59	63	360
E_i	$\frac{360}{7}$	$\frac{360}{7}$	$\frac{360}{7}$	$\frac{360}{7}$	$\frac{360}{7}$	$\frac{360}{7}$	$\frac{360}{7}$	360

$\boxed{\text{H}}$ H_0 : The months have the same number of fatal accidents vs H_1 : Not H_0.

$\boxed{\text{A}}$ We need $E_i \geq 1$ (true here), and no more than 20% of $E_i < 5$ (Cochran's Rule - also true here).

$\boxed{\text{T}}$ $\tau = \sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \frac{O_i^2}{E_i} - 100 \overset{H_0}{\sim} \chi^2_{g-k-1} = \chi^2_{7-0-1} = \chi^2_6$

Large values of τ argue against H_0 for H_1.
Observed value: $\tau_0 = \frac{44^2}{360/7} + \frac{56^2}{360/7} + \dots \frac{63^2}{360/7} - 360 = 12.09.$

$\boxed{\text{P}}$

Using Table:
P-value $= P(\chi^2_6 > 12.09) \in (0.05, 0.1)$.

$\boxed{\text{C}}$ Given $P - value > 0.05$ (just), we retain H_0 and conclude that the months seem to have similar number of fatal accidents.

```r
x=c(44,56,37,42,59,59,63)
t = sum(x^2/(360/7))-360
t

## [1] 12.08889

1-pchisq(t,6)

## [1] 0.06001493
```

5. (Extension: Assessing the goodness of fit of a Normal distribution - estimating 2 parameters)

We can use the Chi-squared Test to assess the fit of a Normal distribution to grouped data.
For the Normal distribution, 2 parameters need to be estimated (both mean and sd), hence k=2.
To perform a goodness of fit test based on grouped data, we need to estimate the parameters (mean and sd) from the grouped data.

```r
x=scan(file=url("http://www.maths.usyd.edu.au/math1005/r/w13.txt"))    # Scan in data
```

```r
hist(x,pr=T)                  # Produce a histogram of the data

hist(x,pr=T)$breaks           # Display the breaks in the histogram

freq=hist(x,pr=T)$counts      # Find the counts in each interval
freq


mids=(-2:4)+.5                  # Find the midpoints of the intervals.
mids

gr.sum=sum(freq*mids)         # Find the sum of grouped data
gr.sum


gr.sumsq=sum(freq*mids^2)     # Find the sum of squares of grouped data
gr.sumsq

gr.mean=gr.sum/100            # Find the mean of grouped data
gr.mean

gr.var=1/99* (gr.sumsq - 1/100* gr.sum^2)  # Find the variance of grouped data
gr.var

gr.sd=sqrt(gr.var)      # Find the mean of grouped data
gr.sd

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

lower.probs=pnorm(-1:4,m=gr.mean,s=gr.sd)      # Finding expected probabilites
lower.probs

exp.probs=diff(c(0,lower.probs,1))
exp.probs

exp.freq= 100* exp.probs                        # Expected frequencies
exp.freq

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

cbind(freq,exp.freq,contrib)

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

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

Solution

```
> x=scan(file=url("http://www.maths.usyd.edu.au/math1005/r/w13.txt"))
Read 100 items
> hist(x,pr=T)

> hist(x,pr=T)$breaks
[1] -2 -1  0  1  2  3  4  5

> freq=hist(x,pr=T)$counts
```

```
> freq
[1]  7 23 24 18 10 13  5

mids=(-2:4)+.5
> 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
> gr.sum
[1] 110

> gr.var=1/99* (gr.sumsq - 1/100* gr.sum^2)
> gr.var
[1] 2.727273

> gr.sd=sqrt(gr.var)
> gr.sd
[1] 1.651446

> lower.probs=pnorm(-1:4,m=gr.mean,s=gr.sd)
> lower.probs
[1] 0.1017553 0.2526790 0.4758576 0.7071154 0.8750325
[6] 0.9604590

> exp.probs=diff(c(0,lower.probs,1))
> exp.probs
[1] 0.10175530 0.15092370 0.22317860 0.23125775
[5] 0.16791712 0.08542650 0.03954103

> exp.freq= 100* exp.probs
> exp.freq
[1] 10.175530 15.092370 22.317860 23.125775 16.791712
[6]  8.542650  3.954103

> contrib = ((exp.freq-freq)^2)/exp.freq
> contrib
[1] 0.9910041 4.1431939 0.1267861 1.1361164 2.7470308
[6] 2.3257383 0.2766496

> cbind(freq,exp.freq,contrib)
     freq  exp.freq   contrib
[1,]    7 10.175530 0.9910041
[2,]   23 15.092370 4.1431939   # High contribution, above Normal curve
[3,]   24 22.317860 0.1267861
[4,]   18 23.125775 1.1361164
[5,]   10 16.791712 2.7470308   # High contribution, below Normal curve
[6,]   13  8.542650 2.3257383   # High contribution, above Normal curve
[7,]    5  3.954103 0.2766496

> tau.obs=sum(((exp.freq-freq)^2)/exp.freq)
> tau.obs
```

```
[1] 11.74652

> 1-pchisq(tau.obs, df=length(freq)-2-1)   # Note we estimate both mean and sd, so k=2
[1] 0.0193392                                # Hence we would reject the fit of Normal.
```

Comment: Perhaps the population that this sample comes from is bimodal, rather than unimodal (Normal). Most of the discrepancy comes from the 2 intervals where the 2 'modes' (peaks) are.