Chi-squared tests 6F

1 H_0: $X \sim \text{Geo}(0.6)$ is a suitable model.
H_1: $X \sim \text{Geo}(0.6)$ is not a suitable model.

As $p = 0.6$, calculate the expected frequencies using the equation:
$E_i = \text{P}(X = i) \times 300 = 0.6(0.4)^{i-1} \times 300 = 180(0.4)^{i-1}$

Require values of E to be at least 5, so combine the final two columns. The observed and expected frequencies and the calculations for the goodness of fit are:

k	1	2	3	4	$\geqslant 5$	Total
Observed (O_i)	207	66	13	9	5	300
Expected (E_i)	180	72	28.8	11.52	7.68	300
$\frac{(O_i - E_i)^2}{E_i}$	4.05	0.5	8.668	0.551	0.935	14.705

Degrees of freedom $\nu = 4$ (five data cells with a single constraint on the total)
From the tables: $\chi^2_4(1\%) = 13.277$

As 14.705 is greater than 13.277, H_0 should be rejected at the 1% level. There is evidence to conclude that $X \sim \text{Geo}(0.6)$ is not a suitable model.

2 H_0: $X \sim \text{Geo}(0.4)$is a suitable model.
H_1: $X \sim \text{Geo}(0.4)$is not a suitable model.

As $p = 0.4$, calculate the expected frequencies using the equation:
$E_i = \text{P}(X = i) \times 100 = 0.4(0.6)^{i-1} \times 100 = 40(0.6)^{i-1}$

The observed and expected frequencies and the calculations for the goodness of fit are:

k	1	2	3	4	5	$\geqslant 6$	Total
Observed (O_i)	42	26	10	8	10	4	100
Expected (E_i)	40	24	14.4	8.64	5.184	7.776	100
$\frac{(O_i - E_i)^2}{E_i}$	0.1	0.167	1.344	0.047	4.474	1.834	7.966

Degrees of freedom $\nu = 5$ (six data cells with a single constraint on the total)
From the tables: $\chi^2_5(5\%) = 11.070$

As 7.966 is less than 11.070, there is insufficient evidence to reject H_0 at the 5% level, so conclude that $X \sim \text{Geo}(0.4)$ is a suitable model for the data.

3 a If the data is drawn from $X \sim \text{Geo}(p)$ then the reciprocal of the mean will be an estimate for p.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{61\times1+24\times2+11\times3+1\times4+2\times5+1\times6}{100} = 1.62$$

Estimate $p = 1.62^{-1} \approx 0.617$

b H_0: The geometric distribution $X \sim \text{Geo}(0.617)$ is a suitable model.
H_1: The geometric distribution $X \sim \text{Geo}(0.617)$ is not a suitable model.

As $p = 0.617$, calculate the expected frequencies using the equation:
$E_i = \text{P}(X = i)\times100 = 0.617(0.383)^{i-1}\times100 = 61.7(0.383)^{i-1}$

The observed and expected frequencies are:

k	1	2	3	4	5	$\geqslant 6$	Total
Observed (O_i)	61	24	11	1	2	1	100
Expected (E_i)	61.7	23.63	9.05	3.47	1.33	0.82	

Require expected values to be at least 5, so combine the final three columns:

k	1	2	3	$\geqslant 4$	Total
Observed (O_i)	61	24	11	4	100
Expected (E_i)	61.7	23.63	9.05	5.62	100
$\frac{(O_i - E_i)^2}{E_i}$	0.008	0.006	0.420	0.467	0.901

Degrees of freedom $\nu = 2$ (four data cells with two constraints: the total and the estimate of p)
From the tables: $\chi^2_2(5\%) = 5.991$

As 0.901 is less than 5.991, there is insufficient evidence to reject H_0 at the 5% level, so conclude that a geometric distribution $X \sim \text{Geo}(0.617)$ may be a suitable model for the data.

4 a If the data is drawn from $X \sim \text{Geo}(p)$ then the reciprocal of the mean will be an estimate for p.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{76\times1+17\times2+4\times3+2\times4+1\times5}{100} = 1.35$$

Estimate $p = 1.35^{-1} \approx 0.741$

b H_0: A geometric distribution is a suitable model.
H_1: A geometric distribution is not a suitable model.

As $p = 0.741$, calculate the expected frequencies using the equation:

$$E_i = \text{P}(X = i) \times 100 = 0.741(0.259)^{i-1} \times 100 = 74.1(0.259)^{i-1}$$

The observed and expected frequencies are:

k	1	2	3	4	$\geqslant 5$	Total
Observed (O_i)	76	17	4	2	1	100
Expected (E_i)	74.1	19.2	4.98	1.29	0.42	100

Require expected values to be at least 5, so combine the final three columns:

k	1	2	$\geqslant 3$	Total
Observed (O_i)	76	17	7	100
Expected (E_i)	74.1	19.2	6.71	100
$\frac{(O_i - E_i)^2}{E_i}$	0.049	0.252	0.013	0.313

Degrees of freedom $\nu = 1$ (three data cells with two constraints: the total and the estimate of p)
From the tables: $\chi^2_1(2.5\%) = 5.024$

As 0.313 is less than 5.024, there is insufficient evidence to reject H_0 at the 2.5% significance level and to conclude that a geometric distribution is not a suitable model.

5 a If Michael's theory is correct then at each letter, the probability of it being a vowel would be $\frac{5}{26}$

In that case, the number of letters until the next vowel would follow a $\text{Geo}\left(\frac{5}{26}\right)$ distribution.

b H_0: The geometric distribution $X \sim \text{Geo}\left(\frac{5}{26}\right)$ is a suitable model.

H_1: The geometric distribution $X \sim \text{Geo}\left(\frac{5}{26}\right)$ is not a suitable model.

As $p = \frac{5}{26}$, calculate the expected frequencies using the equation:

$$E_i = \text{P}(X = i) \times 75 = \frac{5}{26}\left(\frac{21}{26}\right)^{i-1} \times 75 = 14.42\left(\frac{21}{26}\right)^{i-1}$$

The observed and expected frequencies are:

k	1	2	3	4	5	6	7	$\geqslant 8$	Total
Observed (O_i)	12	14	11	10	5	9	7	7	75
Expected (E_i)	14.42	11.65	9.41	7.60	6.14	4.96	4.00	16.82	75

Require expected values at least 5.0, so combine the final three columns:

k	1	2	3	4	5	$\geqslant 6$	Total
Observed (O_i)	12	14	11	10	5	23	75
Expected (E_i)	14.42	11.65	9.41	7.60	6.14	25.78	75
$\frac{(O_i - E_i)^2}{E_i}$	0.407	0.474	0.269	0.758	0.211	0.299	2.419

Degrees of freedom $\nu = 5$ (six data cells with a single constraint on the total)
From the tables: $\chi^2_5(5\%) = 11.070$

As 2.42 is less than 11.070, there is insufficient evidence to reject H_0 at the 5% significance level and to suggest that that the $\text{Geo}\left(\frac{5}{26}\right)$ distribution is not a suitable model.

c The test supports the theory that there is a $\frac{5}{26}$ chance that each letter will be a vowel, which is consistent with Michael's theory, but not as specific. The data does not test that individual letters are random (for example, the same result could be obtained by the monkey only hitting O or X, with $\frac{5}{26}$ and $\frac{21}{26}$ probability respectively at each stroke).

Challenge

If each person has an (independent, constant) probability p of buying a cup, then if Y is the number of people Ellen records:

$\mathrm{P}(Y=10)=p^{10}$ (no non-buyers)

$\mathrm{P}(Y=11)=\binom{10}{9}p^{10}(1-p)^{1}$ (nine buyers and one non-buyer then one buyer)

And in general:

$\mathrm{P}(Y=10+k)=\binom{9+k}{9}p^{10}(1-p)^{k}$ (nine buyers and k non-buyers, then one buyer)

This is a negative binomial distribution $Y\sim$ Negative B$(10,p)$, with $\mathrm{E}(Y)=\dfrac{10}{p}$

(Alternatively notice that Y could be seen as a sum of 10 independent geometric distributions, each representing the number of customers until the next sale, each of which therefore having mean $\dfrac{1}{p}$)

From the data: $\bar{y}=\dfrac{\sum fy}{\sum f}=\dfrac{10\times10+25\times11+29\times12+15\times13+15\times14+10\times15}{10+25+29+15+15+10}=12.29$

Then estimate $p=\dfrac{10}{\bar{y}}=0.814$

H_0: Negative B (10, 0.814) is a suitable model.
H_1: Negative B (10, 0.814) is not a suitable model.

As $p = 0.814$, calculate the expected frequencies using the equation:

$$E_i=\mathrm{P}(Y=i)\times104=\binom{i-1}{9}0.814^{10}(0.186)^{i-10}\times104=13.28\binom{i-1}{9}(0.186)^{i-10}$$

The observed and expected frequencies are:

Number of people	10	11	12	13	14	15	⩾16	Total
Observed (O_i)	10	25	29	15	15	10	0	104
Expected (E_i)	13.28	24.71	25.27	18.80	11.37	5.92	4.65	104

Require expected values at least 5.0, so combine the final two columns:

Number of people	10	11	12	13	14	⩾15	Total
Observed (O_i)	10	25	29	15	15	10	104
Expected (E_i)	13.28	24.71	25.27	18.80	11.37	10.57	104
$\dfrac{(O_i-E_i)^2}{E_i}$	0.810	0.003	0.551	0.768	1.159	0.031	3.321

Degrees of freedom $\nu=4$ (six data cells with two constraints: the total and the estimate for p)
From the tables: $\chi^2_4(5\%)=9.488$

As 3.32 is less than 9.488, there is insufficient evidence to reject H_0 at the 5% significance level, and so conclude that a negative binomial distribution may be a suitable model.