Lecture 9. Tests of goodness-of-fit and independence

Goodness-of-fit of a fully-specified null distribution

Suppose the observation space $\mathcal{X}$ is partitioned into k sets, and let p_i be the probability that an observation is in set i, $i = 1, \ldots, k$.

Consider testing H_0 : the p_i's arise from a fully specified model against H_1: the p_i's are unrestricted (but we must still have $p_i \geq 0$, $\sum p_i = 1$).

This is a **goodness-of-fit** test.

Example 9.1

Birth month of admissions to Oxford and Cambridge in 2012

Month	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
n_i	470	515	470	457	473	381	466	457	437	396	384	394

Are these compatible with a uniform distribution over the year? □

- Out of n independent observations let N_i be the number of observations in the ith set.
- So $(N_1, \ldots, N_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$.
- For a generalised likelihood ratio test of H_0, we need to find the maximised likelihood under H_0 and H_1.
- **Under $\mathbf{H_1}$**: $\text{like}(p_1, \ldots, p_k) \propto p_1^{n_1} \ldots p_k^{n_k}$ so the loglikelihood is $l = \text{constant} + \sum n_i \log p_i$.

  We want to maximise this subject to $\sum p_i = 1$.

  By considering the Lagrangian $\mathcal{L} = \sum n_i \log p_i - \lambda(\sum p_i - 1)$, we find mle's $\hat{p}_i = n_i/n$. Also $|\Theta_1| = k - 1$.
- **Under $\mathbf{H_0}$**: H_0 specifies the values of the p_i's completely, $p_i = \tilde{p}_i$ say, so $|\Theta_0| = 0$.
- Putting these two together, we find

$$2 \log \Lambda = 2 \log \left(\frac{\hat{p}_1^{n_1} \ldots \hat{p}_k^{n_k}}{\tilde{p}_1^{n_1} \ldots \tilde{p}_k^{n_k}} \right) = 2 \sum n_i \log \left(\frac{n_i}{n\tilde{p}_i} \right). \tag{1}$$

- Here $|\Theta_1| - |\Theta_0| = k - 1$, so we reject H_0 if $2 \log \Lambda > \chi^2_{k-1}(\alpha)$ for an approximate size α test.

Example 9.1 continued:

Under H_0 (no effect of month of birth), $\tilde{p}_i$ is the proportion of births in month i in 1993/1994 - this is *not* simply proportional to number of days in month, as there is for example an excess of September births (the 'Christmas effect').

Month	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
n_i	470	515	470	457	473	381	466	457	437	396	384	394
$100\tilde{p}_i$	8.8	8.5	7.9	8.3	8.3	7.6	8.6	8.3	8.6	8.5	8.5	8.3
$n\tilde{p}_i$	466.4	448.2	416.3	439.2	436.9	402.3	456.3	437.6	457.2	450.0	451.3	438.2

- $2 \log \Lambda = 2 \sum n_i \log \left(\frac{n_i}{n\tilde{p}_i} \right) = 44.9$
- $\mathbb{P}(\chi^2_{11} > 44.86) = 3x10^{-9}$, which is our p-value.
- Since this is certainly less than 0.001, we can reject H_0 at the 0.1% level, or can say 'significant at the 0.1% level'.
- NB The traditional levels for comparison are $\alpha = 0.05, 0.01, 0.001$, roughly corresponding to 'evidence', 'strong evidence', and 'very strong evidence'.

Likelihood ratio tests

A similar common situation has $H_0 : p_i = p_i(\theta)$ for some parameter θ and H_1 as before. Now $|\Theta_0|$ is the number of independent parameters to be estimated under H_0.

Under $\mathbf{H_0}$: we find mle $\hat{\theta}$ by maximising $\sum n_i \log p_i(\theta)$, and then

$$2 \log \Lambda = 2 \log \left(\frac{\hat{p}_1^{n_1} \dots \hat{p}_k^{n_k}}{p_1(\hat{\theta})^{n_1} \dots p_k(\hat{\theta})^{n_k}} \right) = 2 \sum n_i \log \left(\frac{n_i}{n p_i(\hat{\theta})} \right). \tag{2}$$

Now the degrees of freedom are $k - 1 - |\Theta_0|$, and we reject H_0 if $2 \log \Lambda > \chi^2_{k-1-|\Theta_0|}(\alpha)$.

Pearson's Chi-squared tests

Notice that (1) and (2) are of the same form.

Let $o_i = n_i$ (the observed number in ith set) and let e_i be $n\tilde{p}_i$ in (1) or $np_i(\hat{\theta})$ in (2). Let $\delta_i = o_i - e_i$. Then

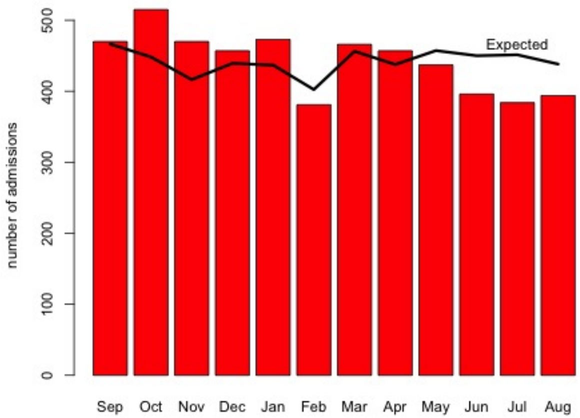
$$\begin{aligned} 2\log\Lambda &= 2\sum o_i \log\left(\frac{o_i}{e_i}\right) \\ &= 2\sum (e_i + \delta_i)\log\left(1 + \frac{\delta_i}{e_i}\right) \\ &\approx 2\sum\left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i}\right) \\ &= \sum \frac{\delta_i^2}{e_i} = \sum \frac{(o_i - e_i)^2}{e_i}, \end{aligned}$$

where we have assumed $\log\left(1 + \frac{\delta_i}{e_i}\right) \approx \frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2}$, ignored terms in δ_i^3, and used that $\sum \delta_i = 0$ (check).

This is **Pearson's chi-squared statistic**; we refer it to $\chi^2_{k-1-|\Theta_0|}$.

Example 9.1 continued using R:

```
chisq.test(n,p=ptilde)
data:  n
X-squared = 44.6912, df = 11, p-value = 5.498e-06
```



Example 9.2

Mendel crossed 556 smooth yellow male peas with wrinkled green female peas. From the progeny let

- N_1 be the number of smooth yellow peas,
- N_2 be the number of smooth green peas,
- N_3 be the number of wrinkled yellow peas,
- N_4 be the number of wrinkled green peas.

We wish to test the goodness of fit of the model $H_0 : (p_1, p_2, p_3, p_4) = (9/16, 3/16, 3/16, 1/16)$, the proportions predicted by Mendel's theory.

Suppose we observe $(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$.

We find $(e_1, e_2, e_3, e_4) = (312.75, 104.25, 104.25, 34.75)$, $2 \log \Lambda = 0.618$ and $\sum \frac{(o_i - e_i)^2}{e_i} = 0.604$.

Here $|\Theta_0| = 0$ and $|\Theta_1| = 4 - 1 = 3$, so we refer our test statistics to χ_3^2.

Since $\chi_3^2(0.05) = 7.815$ we see that neither value is significant at 5% level, so there is no evidence against Mendel's theory.

In fact the p-value is approximately $\mathbb{P}(\chi_3^2 > 0.6) \approx 0.96$. $\square$

NB So in fact could be considered as a suspiciously good fit

Example 9.3

In a genetics problem, each individual has one of three possible genotypes, with probabilities p_1, p_2, p_3. Suppose that we wish to test $H_0 : p_i = p_i(\theta)\ i = 1, 2, 3$, where $p_1(\theta) = \theta^2$, $p_2(\theta) = 2\theta(1-\theta)$, $p_3(\theta) = (1-\theta)^2$, for some $\theta \in (0, 1)$.

We observe $N_i = n_i,\ i = 1, 2, 3$ ($\sum N_i = n$).

Under H_0, the mle $\hat{\theta}$ is found by maximising

$$\sum n_i \log p_i(\theta) = 2n_1 \log\theta + n_2 \log(2\theta(1-\theta)) + 2n_3 \log(1-\theta).$$

We find that $\hat{\theta} = (2n_1 + n_2)/(2n)$ (check).

Also $|\Theta_0| = 1$ and $|\Theta_1| = 2$.

Now substitute $p_i(\hat{\theta})$ into (2), or find the corresponding Pearson's chi-squared statistic, and refer to χ_1^2. $\square$

Testing independence in contingency tables

A table in which observations or individuals are classified according to two or more criteria is called a **contingency table**.

Example 9.4

500 people with recent car changes were asked about their previous and new cars.

		New car		
		Large	Medium	Small
Previous	Large	56	52	42
car	Medium	50	83	67
	Small	18	51	81

This is a two-way contingency table: each person is classified according to previous car size and new car size. □

- Consider a two-way contingency table with r rows and c columns.
- For $i = 1, \ldots, r$ and $j = 1, \ldots, c$ let p_{ij} be the probability that an individual selected at random from the population under consideration is classified in row i and column j (ie in the (i, j) cell of the table).
- Let $p_{i+} = \sum_j p_{ij} = \mathbb{P}(\text{in row } i)$, and $p_{+j} = \sum_i p_{ij} = \mathbb{P}(\text{in column } j)$.
- We must have $p_{++} = \sum_i \sum_j p_{ij} = 1$, ie $\sum_i p_{i+} = \sum_j p_{+j} = 1$.
- Suppose a random sample of n individuals is taken, and let n_{ij} be the number of these classified in the (i, j) cell of the table.
- Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$, so $n_{++} = n$.
- We have

$$(N_{11}, N_{12}, \ldots, N_{1c}, N_{21}, \ldots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, p_{12}, \ldots, p_{1c}, p_{21}, \ldots, p_{rc})$$

- We may be interested in testing the null hypothesis that the two classifications are independent, so test
  - $H_0 : p_{ij} = p_{i+}p_{+j}$, $i = 1, \ldots, r$, $j = 1, \ldots, c$ (with $\sum_i p_{i+} = 1 = \sum_j p_{+j}$, $p_{i+}, p_{+j} \geq 0$),
  - $H_1 : p_{ij}$'s unrestricted (but as usual need $p_{++} = 1$, $p_{ij} \geq 0$).
- Under H_1 the mle's are $\hat{p}_{ij} = n_{ij}/n$.
- Under H_0, using Lagrangian methods, the mle's are $\hat{p}_{i+} = n_{i+}/n$ and $\hat{p}_{+j} = n_{+j}/n$.
- Write o_{ij} for n_{ij} and let $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n_{i+}n_{+j}/n$.
- Then

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} \log\left(\frac{o_{ij}}{e_{ij}}\right) \approx \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  using the same approximating steps as for Pearson's Chi-squared test.
- We have $|\Theta_1| = rc - 1$, because under H_1 the p_{ij}'s sum to one.
- Further, $|\Theta_0| = (r-1) + (c-1)$, because $p_{1+}, \ldots, p_{r+}$ must satisfy $\sum_i p_{i+} = 1$ and $p_{+1}, \ldots, p_{+c}$ must satisfy $\sum_j p_{+j} = 1$.
- So $|\Theta_1| - |\Theta_0| = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$.

Example 9.5

In Example 9.4, suppose we wish to test H_0 : the new and previous car sizes are independent.

We obtain:

	o_{ij}	New car Large	New car Medium	New car Small	
Previous car	Large	56	52	42	150
	Medium	50	83	67	200
	Small	18	51	81	150
		124	186	190	500

	e_{ij}	New car Large	New car Medium	New car Small	
Previous car	Large	37.2	55.8	57.0	150
	Medium	49.6	74.4	76.0	200
	Small	37.2	55.8	57.0	150
		124	186	190	500

Note the margins are the same.

Then $\sum\sum \frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 36.20$, and $\text{df} = (3-1)(3-1) = 4$.

From tables, $\chi_4^2(0.05) = 9.488$ and $\chi_4^2(0.01) = 13.28$.

So our observed value of 36.20 is significant at the 1% level, ie there is strong evidence against H_0, so we conclude that the new and present car sizes are not independent.

It may be informative to look at the contributions of each cell to Pearson's chi-squared:

		New car		
		Large	Medium	Small
Previous	Large	9.50	0.26	3.95
car	Medium	0.00	0.99	1.07
	Small	9.91	0.41	10.11

It seems that more owners of large cars than expected under H_0 bought another large car, and more owners of small cars than expected under H_0 bought another small car.

Fewer than expected changed from a small to a large car. □