Goodness of Fit Tests: Unknown Parameters

Mathematics 47: Lecture 33

Dan Sloughter

Furman University

May 8, 2006

Fitting a family of distributions

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution F.

Fitting a family of distributions

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution F.
- Let $\mathcal{F}$ be a family of distributions depending on r parameters $\theta_1, \theta_2, \ldots, \theta_r$.

Fitting a family of distributions

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution F.
- Let $\mathcal{F}$ be a family of distributions depending on r parameters $\theta_1, \theta_2, \ldots, \theta_r$.
- Suppose we wish to test

$$H_0 : F \in \mathcal{F}$$
$$H_A : F \notin \mathcal{F}.$$

Fitting a family of distributions

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution F.
- Let $\mathcal{F}$ be a family of distributions depending on r parameters $\theta_1, \theta_2, \ldots, \theta_r$.
- Suppose we wish to test

$$H_0 : F \in \mathcal{F}$$
$$H_A : F \notin \mathcal{F}.$$

- We divide the range of the random variables into k disjoint cells and, for $i = 1, 2, \ldots, k$, let

$$p_i = \text{ probability that an observation lies in the } i\text{th cell}$$

and

$$Y_i = \text{ number of observations in the } i\text{th cell.}$$

Fitting a family of distributions

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution F.
- Let $\mathcal{F}$ be a family of distributions depending on r parameters $\theta_1, \theta_2, \ldots, \theta_r$.
- Suppose we wish to test

$$H_0 : F \in \mathcal{F}$$
$$H_A : F \notin \mathcal{F}.$$

- We divide the range of the random variables into k disjoint cells and, for $i = 1, 2, \ldots, k$, let

$$p_i = \text{ probability that an observation lies in the } i\text{th cell}$$

and

$$Y_i = \text{ number of observations in the } i\text{th cell.}$$

- Under the null hypothesis, the probabilities $p_1, p_2, \ldots, p_k$ are functions of the parameters $\theta_1, \theta_2, \ldots, \theta_r$.

Fitting a family of distributions (cont'd)

- That is, for $i = 1, 2, \ldots, k$, we may write $p_i = p_i(\theta_1, \theta_2, \ldots, \theta_r)$.

Fitting a family of distributions (cont'd)

- That is, for $i = 1, 2, \ldots, k$, we may write $p_i = p_i(\theta_1, \theta_2, \ldots, \theta_r)$.
- For $i = 1, 2, \ldots, r$, let $\hat{\theta}_i$ be the maximum likelihood estimator for θ_i and let

$$\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r).$$

Fitting a family of distributions (cont'd)

- That is, for $i = 1, 2, \ldots, k$, we may write $p_i = p_i(\theta_1, \theta_2, \ldots, \theta_r)$.
- For $i = 1, 2, \ldots, r$, let $\hat{\theta}_i$ be the maximum likelihood estimator for θ_i and let

$$\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r).$$

- It may be shown that, when H_0 is true, both

$$-2\log(\Lambda) = 2\sum_{i=1}^{k} Y_i \log\left(\frac{Y_i}{n\hat{p}_i}\right)$$

and

$$Q = \sum_{i=1}^{n} \frac{(Y_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

are asymptotically $\chi^2(k - r - 1)$.

Fitting a family of distributions (cont'd)

- That is, for $i = 1, 2, \ldots, k$, we may write $p_i = p_i(\theta_1, \theta_2, \ldots, \theta_r)$.
- For $i = 1, 2, \ldots, r$, let $\hat{\theta}_i$ be the maximum likelihood estimator for θ_i and let

$$\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r).$$

- It may be shown that, when H_0 is true, both

$$-2\log(\Lambda) = 2\sum_{i=1}^{k} Y_i \log\left(\frac{Y_i}{n\hat{p}_i}\right)$$

and

$$Q = \sum_{i=1}^{n} \frac{(Y_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

are asymptotically $\chi^2(k - r - 1)$.

- Note: we lose one degree of freedom for every parameter that we estimate from the data.

Example

Example

- During World War II it was asked whether German bomb strikes on South London were random or not.

Example

- During World War II it was asked whether German bomb strikes on South London were random or not.
- To test this hypothesis, the city was divided into 576 regions, each $\frac{1}{4}$ of a square kilometer in area.

Example

- During World War II it was asked whether German bomb strikes on South London were random or not.
- To test this hypothesis, the city was divided into 576 regions, each $\frac{1}{4}$ of a square kilometer in area.
- The data were as follows:

Number of Hits	Frequency	Expected Frequency
0	229	227.53
1	211	211.34
2	93	98.15
3	35	30.39
4	7	7.06
5+	1	1.54
Total	576	576.01

Example (cont'd)

Example (cont'd)

- We want to test the hypotheses

$$H_0: \text{ Data are Poisson}$$
$$H_A: \text{ Data are not Poisson.}$$

Example (cont'd)

- We want to test the hypotheses

$$H_0: \text{ Data are Poisson}$$
$$H_A: \text{ Data are not Poisson.}$$

- If λ is the mean of the hypothesized Poisson distribution, then the maximum likelihood estimator of λ is

$$\hat{\lambda} = \frac{0 \cdot 229 + 1 \cdot 211 + 2 \cdot 93 + 3 \cdot 35 + 4 \cdot 7 + 5 \cdot 1}{576} = \frac{535}{576} = 0.9288.$$

Example (cont'd)

Example (cont'd)

- If $p_i(\lambda)$, $i = 0, 1, 2, 3, 4, 5$, is the probability of an outcome in the ith cell, then

$$p_0(\hat{\lambda}) = e^{-0.9288} = 0.3950,$$
$$p_1(\hat{\lambda}) = 0.9288e^{-0.9288} = 0.3669,$$
$$p_2(\hat{\lambda}) = \frac{(0.9288)^2 e^{-0.9288}}{2} = 0.1704,$$
$$p_3(\hat{\lambda}) = \frac{(0.9288)^3 e^{-0.9288}}{6} = 0.0528,$$
$$p_4(\hat{\lambda}) = \frac{(0.9288)^4 e^{-0.9288}}{24} = 0.0122,$$
$$p_5(\hat{\lambda}) = 1 - p_0(\hat{\lambda}) - p_1(\hat{\lambda}) - p_2(\hat{\lambda}) - p_3(\hat{\lambda}) - p_4(\hat{\lambda}) = 0.0027.$$

Example (cont'd)

- If $p_i(\lambda)$, $i = 0, 1, 2, 3, 4, 5$, is the probability of an outcome in the ith cell, then

$$p_0(\hat{\lambda}) = e^{-0.9288} = 0.3950,$$
$$p_1(\hat{\lambda}) = 0.9288e^{-0.9288} = 0.3669,$$
$$p_2(\hat{\lambda}) = \frac{(0.9288)^2 e^{-0.9288}}{2} = 0.1704,$$
$$p_3(\hat{\lambda}) = \frac{(0.9288)^3 e^{-0.9288}}{6} = 0.0528,$$
$$p_4(\hat{\lambda}) = \frac{(0.9288)^4 e^{-0.9288}}{24} = 0.0122,$$
$$p_5(\hat{\lambda}) = 1 - p_0(\hat{\lambda}) - p_1(\hat{\lambda}) - p_2(\hat{\lambda}) - p_3(\hat{\lambda}) - p_4(\hat{\lambda}) = 0.0027.$$

- Note: Since the range of a Poisson random variable is the set of nonnegative integers, the last cell is really the set $\{5, 6, 7, \ldots\}$.

Example (cont'd)

Example (cont'd)

- Multiplying each of these probabilities by 576 yields the expected frequencies shown in the table.

Example (cont'd)

- Multiplying each of these probabilities by 576 yields the expected frequencies shown in the table.
- Note: The fifth cell has an expected frequency of only 1.54; since this is less than 5, we combine the fourth and fifth cells into a single cell with observed frequency 8 and expected frequency 8.60.

Example (cont'd)

- Multiplying each of these probabilities by 576 yields the expected frequencies shown in the table.
- Note: The fifth cell has an expected frequency of only 1.54; since this is less than 5, we combine the fourth and fifth cells into a single cell with observed frequency 8 and expected frequency 8.60.
- Plugging into the respective formulas, we now find that $q = 1.021441$ and $-2\log(\lambda) = 0.9743838$.

Example (cont'd)

- Multiplying each of these probabilities by 576 yields the expected frequencies shown in the table.
- Note: The fifth cell has an expected frequency of only 1.54; since this is less than 5, we combine the fourth and fifth cells into a single cell with observed frequency 8 and expected frequency 8.60.
- Plugging into the respective formulas, we now find that $q = 1.021441$ and $-2\log(\lambda) = 0.9743838$.
- If U is $\chi^2(3)$, the corresponding p-values are $P(U \geq 1.021441) = 0.796064$ and $P(U \geq 0.9743838) = 0.80745$.

Example (cont'd)

- Multiplying each of these probabilities by 576 yields the expected frequencies shown in the table.
- Note: The fifth cell has an expected frequency of only 1.54; since this is less than 5, we combine the fourth and fifth cells into a single cell with observed frequency 8 and expected frequency 8.60.
- Plugging into the respective formulas, we now find that $q = 1.021441$ and $-2\log(\lambda) = 0.9743838$.
- If U is $\chi^2(3)$, the corresponding p-values are $P(U \geq 1.021441) = 0.796064$ and $P(U \geq 0.9743838) = 0.80745$.
- Hence we have no evidence to reject H_0, and it appears that the Poisson model provides a good description of the observed data.

Example

Example

- A test of the lifetimes of 100 lightbulbs yielded the following data:

366.34	150.63	216.19	1870.05	204.58	2234.09	115.51
398.96	2810.98	993.60	2226.60	531.07	1140.28	1701.23
155.72	143.71	817.35	1150.37	488.04	698.01	1276.04
1210.49	1173.72	206.91	233.15	1535.54	4157.71	216.48
593.00	3316.39	49.69	283.51	2723.97	95.42	1036.12
2899.19	1651.60	276.94	365.56	106.38	1763.62	2452.65
238.31	1421.59	62.52	837.63	1995.29	965.04	422.81
96.10	911.44	790.79	1283.83	1730.51	480.00	822.78
3435.11	355.56	3838.57	215.17	1656.36	289.00	804.78
1517.13	855.55	185.74	1238.47	1499.39	2155.75	493.45
62.46	483.64	328.92	438.24	747.85	429.91	582.57
161.34	41.06	590.45	35.34	33.52	179.37	283.00
1218.80	602.75	425.45	688.08	117.92	372.43	582.45
509.95	87.77	512.72	229.54	784.20	288.06	1370.94
228.57	2375.15					

Example (cont'd)

Example (cont'd)

- We wish to test the hypothesis that these data are from an exponential distribution:

$$H_0: \text{ Data are exponential}$$
$$H_A: \text{ Data are not exponential.}$$

Example (cont'd)

- We wish to test the hypothesis that these data are from an exponential distribution:

$$H_0: \text{ Data are exponential}$$
$$H_A: \text{ Data are not exponential.}$$

- We group the data into cells of length 500, count the observed frequencies of each cell, and compute the expected frequencies for each cell.

Example (cont'd)

- We wish to test the hypothesis that these data are from an exponential distribution:

$$H_0: \text{ Data are exponential}$$
$$H_A: \text{ Data are not exponential.}$$

- We group the data into cells of length 500, count the observed frequencies of each cell, and compute the expected frequencies for each cell.
- Since $\bar{x} = 914.29$, we will compute expected frequencies using an exponential distribution with mean 914.29.

Example (cont'd)

- We wish to test the hypothesis that these data are from an exponential distribution:

$$H_0: \text{ Data are exponential}$$
$$H_A: \text{ Data are not exponential.}$$

- We group the data into cells of length 500, count the observed frequencies of each cell, and compute the expected frequencies for each cell.
- Since $\bar{x} = 914.29$, we will compute expected frequencies using an exponential distribution with mean 914.29.
- That is, for example, the expected frequency for the first cell is

$$100 \int_0^{500} \frac{1}{914.29} e^{-\frac{x}{914.29}} = 100 \left(1 - e^{-\frac{500}{914.29}}\right) = 42.12.$$

Example (cont'd)

Example (cont'd)

- We then have the following table:

Interval	Observed frequency	Expected frequency
$[0, 500]$	46	42.12
$(500, 1000]$	21	24.38
$(1000, 1500]$	12	14.11
$(1500, 2000]$	9	8.17
$(2000, 2500]$	5	4.73
$(2500, 3000]$	3	2.74
$(3000, 3500]$	2	1.58
$(3500, 4000]$	1	0.92
$(4000, \infty)$	1	1.26
Total	100	100.01

Example (cont'd)

Example (cont'd)

- We will group the final four cells together because of their low expected frequencies:

Interval	Observed frequency	Expected frequency
$[0, 500]$	46	42.12
$(500, 1000]$	21	24.38
$(1000, 1500]$	12	14.11
$(1500, 2000]$	9	8.17
$(2000, 2500]$	5	4.73
$(2500, \infty)$	7	6.50
Total	100	100.01

Example (cont'd)

Example (cont'd)

- Our test statistics are $q = 1.279737$ and $-2\log(\lambda) = 1.285602$.

Example (cont'd)

- Our test statistics are $q = 1.279737$ and $-2\log(\lambda) = 1.285602$.
- If U is $\chi^2(4)$, then our p-values are $P(U \geq 1.279737) = 0.864804$ and $P(U \geq 1.285602) = 0.8638135$.

Example (cont'd)

- Our test statistics are $q = 1.279737$ and $-2\log(\lambda) = 1.285602$.
- If U is $\chi^2(4)$, then our p-values are $P(U \geq 1.279737) = 0.864804$ and $P(U \geq 1.285602) = 0.8638135$.
- Hence we have no evidence to reject H_0, and it appears that the exponential distribution provides a good description of the observed data.