13: GOODNESS-OF-FIT TESTS

Larsen & Marx 10

Prof Amanda Luby

Up until now, we've learned how to *estimate* parameters and how to draw *inferences* about possible parameter values given a set of data. In all of these scenarios, we've assumed that the form of p_x or f_x is known. In many scenarios, we're instead interested in making inferences about the form of p_x or f_x instead of the value of the parameters.

In general, statistical procedures that seek to determine whether a set of data could reasonably have originated from some probability distribution (or family of probability distributions) is called a **goodness-of-fit** test.

1 The multinomial distribution

Multinomial Distribution

Let X_i denote the number of times that the outcome r_i occurs, for $i = 1, ..., t$ in a series of n independent trials where $p_i = P(r_i)$. Then the *vector* $(X_1, X_2, ..., X_t)$ has a **multinomial** distribution and

$$p_{X_1,...,X_t}(k_1, ..., k_t) = \frac{n!}{k_1!k_2!...k_t!} p_1^{k_1} p_2^{k_2} ....p_t^{k_t}$$

Example: Five observations are drawn at random from a continuous Uniform(0,5) distribution. What is the probability that one observation lies in the interval $[0, 1)$, none in the interval $[1, 2)$, three in the interval $[2, 3)$, one in the interval $[3, 4)$, and none in the interval $[4, 5)$?

2 Goodness of Fit Test: All parameters known

The simplest goodness of fit test arises when we're able to *completely* specify the model that we believe our data came from. For example, whether our observed y_i's came from a $Exp(6.3)$ distribution, or a $N(2.2, 5.4)$ distribution.

Pearson's χ^2 test statistic

Let $r_1, \ldots, r_t$ be the set of outcomes associated with n independent trials. Let X_i be the number of times r_i occurs. Then,

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$

Proof (t=2):

Example: From the uniform example earlier, test $H_0 : p_1 = 1/5, p_2 = 1/5, p_3 = 1/5, p_4 = 1/5, p_5 = 1/5$ against H_1 : at least one different.

3 Goodness of fit tests: parameters unknown

The above test statistic assumes that we know p_i for each class i. Since p_i does not have a hat on it, it's the true population parameter for a data point falling into class i. It's rare that we would know θ for a pdf $f_y(\theta)$, but not be sure about the form of f. A more common scenario is to *estimate* all unknown parameters first, and then use a modified version of Pearson's D Statistic:

Approximate χ^2 test statistic

Suppose that a random sample of n observations is taken from $f_x(x; \theta)$ or $f_x(x; \theta)$, a probability distribution having s unknown parameters. Let $r_1, \ldots, r_t$ be the set of outcomes associated with n independent trials. Let X_i be the number of times r_i occurs, and let $\hat{p}_i$ be the *estimated* probability of r_i, replacing θ in $p_x(x; \theta)$ or $f_x(x; \theta)$ with $\hat{\theta}$. Then,

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

Example: The Poisson probability distribution often models rare events that occur over a period of time. Listed below are the daily numbers of death notices for women over the age of 80 that appeared in the London Times over a 3 year period. Are these fatalities occuring in a pattern consistent with the Poisson pdf?

```
tibble(
  n_deaths = 0:10,
  observed = c(162, 267, 271, 185, 111, 61, 27, 8, 3, 1, 0),
  expected = dpois(n_deaths,
                   lambda = sum(n_deaths*observed)/(365*3))*1096
) %>%
  knitr::kable(digits = 2)
```

n_deaths	observed	expected
0	162	126.53
1	267	273.18
2	271	294.88
3	185	212.21
4	111	114.53
5	61	49.45
6	27	17.79
7	8	5.49
8	3	1.48
9	1	0.36
10	0	0.08