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.

Idea: sample $\{x_1, \dots, x_n\}$. Put x_i's into k groups (k arbitrary)

Assume a particular $f_x(x;\theta)$ and find the expected # of observations in each group under f_x

compare observed counts to expected counts → 'close' → likely that $X \sim f_x$

→ 'far away' → unlikely that $X \sim f_x$

1 The multinomial distribution

t = # of categories

$\vec{X} = (X_1, \dots, X_t)$

K = # of observations

$\sum p_i = 1$

Multinomial Distribution → Extension of binomial for ≥2 outcomes

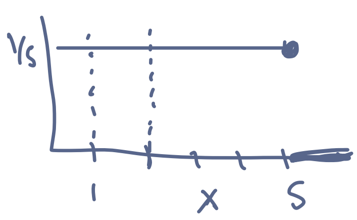
Let X_i denote the number of times that the outcome r_i occurs, for $i = 1, \dots, t$ in a series of n independent trials where $p_i = P(r_i)$. Then the *vector* $(X_1, X_2, \dots, X_t)$ has a **multinomial** distribution and

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

t=2

$$\frac{n!}{k_1!(1-k_1)!} p_1^{k_1}(1-p_1)^{n-k_1}$$

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 $[4, 5)$?



$P(0 \le y \le 1) = 1/5$

$\vdots$

$P(4 \le y \le 5) = 1/5$

$$p_{X_1,X_2,X_3,X_4,X_5}(1,0,3,1,0) = \frac{5!}{1!0!3!1!0!}\left(\frac{1}{5}\right)^1\left(\frac{1}{5}\right)^0\left(\frac{1}{5}\right)^3\left(\frac{1}{5}\right)^1\left(\frac{1}{5}\right)^0$$

$$= \frac{5!}{3!}\left(\frac{1}{5}\right)^5 = .0064$$

In R:

```
dmultinom(x=c(1,0,3,1,0),
  prob=c(.2,.2,.2,.2,.2))
```

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.

$H_0: f_Y(y) = f_0(y)$

$H_1: f_Y(y) \neq f_0(y)$

$\xrightarrow[\text{leave discrete data}]{\text{'bin' continuous data}}$

$H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_t = p_{t0}$

$H_1: p_i \neq p_{i0}$ for at least 1 i

Pearson's χ^2 test statistic

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

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

for the approximation to be adequate, $np_i \geq 5$ for all i

Proof (t=2):

$$D = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

since $t=2$, $\sum p_i = 1$

$p_2 = 1 - p_1$

$X_1 + X_2 = n$

$\downarrow$

lose a degree of freedom since $\sum p_i = 1$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1-p_1))^2}{n(1-p_1)}$$

$$= \frac{(X_1 - np_1)^2(1-p_1) + (-X_1 + np_1)^2 p_1}{np_1(1-p_1)}$$

$$= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \xrightarrow{X_1 \sim Bin(n, p_1)} \left[\frac{X_1 - E[X_1]}{\sqrt{Var[X_1]}}\right]^2 \to Z^2 \sim \chi^2_1$$

why we need $np_i \geq 5$

(skip)

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 $p_x(x;\theta)$, 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,

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

for approximation to be adequate, $n\hat{p}_i \geq 5$ for all 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)
```

sum(observed)

$P(X=k)$ plugging in $\hat{\lambda}$

n

X_i : # of deaths on day i

H_0: $X_i \sim \text{Poisson}(\lambda)$

H_1: $X_i \not\sim \text{Poisson}(\lambda)$

(1) Figure out 'expected' column $n\hat{p}_i$

(1a) Estimate $\lambda \to \hat{\lambda}_{MLE}$

$$\hat{\lambda}_{MLE} = \frac{\text{\# of fatalities}}{\text{\# of days}}$$

~~t=11~~

t=8

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

7+ , 12 , 7.3

$$\hat{\lambda} = \bar{X} = \frac{0 \cdot 162 + 1 \cdot 267 + 2 \cdot 271 + \cdots 9 \cdot 1}{162 + 267 + \cdots + 1} = 2.157$$

$$\text{Expected: } n \cdot \hat{p}_k \to n \cdot P(X = k) \to n \cdot \frac{e^{-2.157}\, 2.157^k}{k!}$$

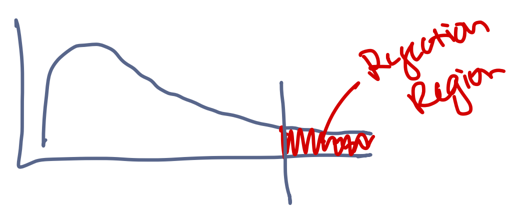
(2) compute test stat:

$$d_1 = \frac{(162 - 126.53)^2}{126.53} + \frac{(267 - 273.18)^2}{273.18} + \cdots + \frac{(12 - 7.3)^2}{7.3} \approx 25.98$$

(3) compute p-value

$$d_1 \sim \chi^2_{8-1-1}$$

↑ categories, ← $\sum p_i = 1$, ← estimated 1 param $\hat{\lambda}$



qchisq(.95, 6)
= 12.592

→ reject H_0 that X_i's follow a Poisson distribution.

Quiz 3)

Sampling distributions

$N, T, \chi^2 \rightarrow$ when are they appropriate?

approximation vs exact ; one vs. two sample ; goodness of fit

Identify CI or H_0/H_1 for a given scenario, and determine test statistic / sampling distribution

Interpreting results

Diff ways of calculate standard error $\begin{cases} \text{"plug in" } s, \hat{p} \\ \text{under the} \\ \text{conservative SE's for } p \end{cases}$

How are CI's / HT's impacted by sample size / α?

Relationship between rejection region R and CI

Finding and using power functions $\rightarrow$ Type I / Type II error

Composite nulls : α, size, etc

Setting up GLRT's

Sunday: HW 9 graded, HW10 solutions posted

Extra office hour : 2-3 pm