

# ST2132 Goodness-of-fit

Semester 1 2022/23

- ▶ Assuming a statistical model for data, we can estimate parameters, SEs, and construct CI's. But these do not indicate how well the model fits the data.
- ▶ Goodness-of-fit of models is often check using a hypothesis test. We will look at:
  1. Pearson's  $X^2$ , for multinomial data
  2. Likelihood ratio (LR)

# Possibilities for a die

Roll a die  $n$  times independently and in the same way:

$$(X_1, \dots, X_6) \sim \text{Multinomial}(n, (p_1, \dots, p_6))$$

1. It could be that  $p_1 = p_2 = p_3$ ,  $p_4 = p_5$ , i.e., there are at most 3 different probabilities.
2. Or  $p_1 = p_2 = p_3$ ,  $p_4 = p_5 = p_6$ .
3. Or  $p_1 = \dots = p_6$ , i.e., die is fair.

The three models are nested in the general multinomial model.  
Which might be preferred?

A general die can be represented as

$$\Omega = \{(p_1, \dots, p_6) : p_i > 0, \sum_{i=1}^6 p_i = 1\}$$

1.  $p_1 = p_2 = p_3, p_4 = p_5$  correspond to a subset of  $\Omega$ :

$$\{(\theta_1, \theta_1, \theta_1, \theta_2, \theta_2, \theta_3) : \theta_i > 0, 3\theta_1 + 2\theta_2 + \theta_3 = 1\}$$

2.  $p_1 = p_2 = p_3, p_4 = p_5 = p_6$ :

$$\{(\theta_1, \theta_1, \theta_1, \theta_2, \theta_2, \theta_2) : \theta_i > 0, \theta_1 + \theta_2 = 1/3\}$$

# Multinomial goodness-of-fit

Let  $(X_1, \dots, X_r) \sim \text{Multinomial}(n, \mathbf{p})$ , with  $n, r$  fixed. Set of all distributions:

$$\Omega = \{(p_1, \dots, p_r) : p_i > 0, \sum_{i=1}^r p_i = 1\}$$

Consider a subset where  $\mathbf{p}$  depends on  $\theta \in \Theta \subset \mathbb{R}^k$ ,  $k < r - 1$ :

$$\Omega_0 = \{(p_1(\theta), \dots, p_r(\theta)) : \theta \in \Theta\}$$

Given realisations  $(x_1, \dots, x_r)$ , to judge whether  $\mathbf{p} \in \Omega_0$ .

Assume  $(342, 500, 187)$  is a realisation of  $\mathbf{X} \sim \text{Trinomial}(1029, \mathbf{p})$ .

$$\Omega = \{(p_1, p_2, p_3) : p_i > 0, \sum_{i=1}^3 p_i = 1\}$$

HWE says  $\mathbf{p}$  is in

$$\Omega_0 = \{(1 - \theta)^2, 2\theta(1 - \theta), \theta^2) : 0 < \theta < 1\}$$

Goodness-of-fit test:

1. Null hypothesis  $H_0 : \mathbf{p} \in \Omega_0$ .
2. Alternative hypothesis  $H_1 : \mathbf{p} \in \Omega_1 = \Omega - \Omega_0$ .
3. Calculate test statistic and  $P$ -value, then conclude.

# Observed vs expected counts

If  $H_0$  is true, we expect  $X_1$  to be

$$1029 \times (1 - \theta)^2 \approx 1029 \times (1 - 0.42)^2 \approx 340.6$$

<i>Observed</i>	<i>Expected</i>	$(O - E)^2/E$
342	340.6	0.006
500	502.8	0.016
187	185.6	0.011
1029	1029	0.033

The  $P$ -value is roughly  $\Pr(\chi_1^2 \geq 0.033) \approx 0.86$ .

0.033 measures a distance between observed and expected frequencies. If  $H_0$  is true, the chance is quite high to get data that are more extreme than (342,500,187). HWE seems to fit well.

## Assumption.

$(X_1, \dots, X_r) \sim \text{Multinomial}(n, \mathbf{p}(\theta)), \theta \in \Theta \subset \mathbb{R}^k, k < r - 1.$

Definitions.

$\hat{\theta}$ : ML estimator of  $\theta$ .  $n\mathbf{p}(\hat{\theta})$ : random expected counts.

Chi-square statistic:

$$\chi^2 = \sum_{i=1}^r \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

**Theorem.** As  $n \rightarrow \infty$ , the distribution of  $\chi^2$  converges to  $\chi_{r-1-k}^2$ .

Strictly speaking, the dimension of  $\Theta$  has to be  $k$ .



# $\chi^2$ goodness-of-fit test

$(X_1, \dots, X_r) \sim \text{Multinomial}(n, \mathbf{p}), \mathbf{p} \in \Omega$ , with  $n$  large.

$\Omega_0 = \{\mathbf{p}(\theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ .  $\Omega_1 = \Omega - \Omega_0$ .

1.  $H_0 : \mathbf{p} \in \Omega_0$  (assumption in previous slide).
2.  $H_1 : \mathbf{p} \in \Omega_1$ .
3. Substituting each  $X_i$  by  $x_i$ , and  $\hat{\theta}$  by the ML estimate, we get a realisation  $x^2$  of  $X^2$ .
4. The  $P$ -value:

$$\Pr(X^2 \geq x^2) \approx \Pr(\chi_{r-1-k}^2 \geq x^2)$$

The smaller it is, the more suspicious we are of  $H_0$ .

- ▶ HWE:  $r = 3$ .  $k = 1$ , since the probabilities are modelled as functions of  $\theta$ .  $r - 1 - k = 1$ .
- ▶  $k$  can be 0, in which case  $\Omega_0$  is a single point. Then the expected counts are exact, not estimated.

For a die, if  $H_0$  says it is fair, i.e.,  $p_i = 1/6$ , then there is no parameter to estimate.  $r - 1 - k = 5$ .

Tutorial 9 Question 3:  $x^2 = 14.2$ ,  $P(X^2 \geq 14.2) \approx \Pr(\chi_5^2 \geq 14.2) \approx 0.01$ . The die is likely unfair.

- Define

$$G = 2 \sum_{i=1}^r X_i \log \left( \frac{X_i}{np_i(\hat{\theta})} \right)$$

- In the following,  $g$  is very close to  $x^2$ .

<i>Example</i>	<i>n</i>	$x^2$	$g$
HWE	1209	0.033	0.032
T8Q3	60	14.20	14.15
T8Q4	3839	2.02	2.02

- $G$  is a **likelihood ratio** statistics.

# Likelihood ratio and $G$

$(X_1, \dots, X_r) \sim \text{Multinomial}(n, \mathbf{p})$ .

- ▶ Maximum of likelihood  $L(\mathbf{p}) = \prod_{i=1}^r p_i^{x_i}$  over  $\Omega$ :

$$L_1 =$$

- ▶ Maximum of likelihood  $L(\theta) = \prod_{i=1}^r p_i(\theta)^{x_i}$  over  $\Omega_0$ :

$$L_0 =$$

- ▶ Always,  $L_1/L_0 \geq 1$ . The larger the ratio, the more we doubt  $H_0$ , which says that  $\mathbf{p} \in \Omega_0$ .



$$2 \log \left( \frac{L_1}{L_0} \right) =$$

# $\Omega$ and parameter space

In applications, we often identify the big model  $\Omega$  with a natural parameter space  $\Theta$ . Then the small model  $\Omega_0$  is a subset of  $\Theta$ .

- ▶ For the general trinomial distribution, we can define

$$\Omega = \{(p_1, p_2) : p_i > 0, p_1 + p_2 < 1\}$$

- ▶ The HWE model is the subset  $\Omega_0$  with

$$p_1 = (1 - t)^2, p_2 = 2t(1 - t), \quad 0 < t < 1$$

- ▶ Let  $\Omega_1 = \Omega - \Omega_0$  and  $\theta = (p_1, p_2)$ . The goodness-of-fit test of the HWE model can be stated as follows:

$$H_0 : \theta \in \Omega_0 \quad H_1 : \theta \in \Omega_1$$

## Assumption.

$n$  IID RV's density defined by  $\theta \in \Omega$  with  $k_1$  independent parameters.

$L_1$ : maximum likelihood value over  $\Omega$ .

Nested in  $\Omega$  is  $\Omega_0$  with  $k_0 < k_1$  independent parameters.

$L_0$ : maximum likelihood value over  $\Omega_0$ .

**Theorem.** Suppose  $\theta \in \Omega_0$ . As  $n \rightarrow \infty$ , the distribution of

$$G = 2 \log \left( \frac{L_1}{L_0} \right)$$

converges to  $\chi^2_{k_1 - k_0}$ .

Letting  $\ell_0 = \log L_0$ ,  $\ell_1 = \log L_1$ ,  $G = 2(\ell_1 - \ell_0)$ .

# LR goodness-of-fit test

Assumption as in previous slide.  $\Omega_1 = \Omega - \Omega_0$ .

1.  $H_0 : \theta \in \Omega_0$ .
2.  $H_1 : \theta \in \Omega_1$ .
3.  $L_0$  and  $L_1$  are the maximum likelihood values under  $\Omega_0$  and  $\Omega$ .

$$g = 2 \log \left( \frac{L_1}{L_0} \right)$$

is a realisation of  $G$ .

4. The  $P$ -value is calculated with distribution of  $G$  under  $H_0$ :

$$\Pr(G \geq g) \approx \Pr(\chi_{k_1 - k_0}^2 \geq g)$$

# Multinomial goodness-of-fit

- ▶ To judge whether multinomial data  $(x_1, \dots, x_r)$  might have come from a simpler model with  $k < r - 1$  parameters, either

$$G = 2 \sum_{i=1}^r X_i \log \left( \frac{X_i}{np_i(\hat{\theta})} \right), \quad \chi^2 = \sum_{i=1}^r \frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

can be used. For large  $n = \sum_{i=1}^r x_i$ , under  $H_0$ , both are approximately  $\chi^2_{r-1-k}$ .

- ▶ If the simpler model consists of a distribution specified by  $\theta_0$ , then  $k = 0$  and  $\hat{\theta}$  should be replaced by  $\theta_0$ .
- ▶ For multinomial data, there is no need to evaluate  $L_0$  and  $L_1$  to compute  $G$ . Other examples may also yield such shortcuts.



## Bacterial clumps in milk (pg 344)

- Assume 400 counts of bacterial clumps are realisations of IID  $\text{Poisson}(\lambda)$  RV's. By ML estimate of  $\lambda$  is the mean number: 2.44.

<i>Number</i>	0	1	2	3	4	5	6	$\geq 7$
<i>Frequency</i>	56	104	80	62	42	27	9	20
<i>Expected</i>	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0

- The model looks bad. We will assess how bad it is relative to a larger model:

For  $i = 1, \dots, 400$ ,  $x_i$  is a realisation of  $X_i \sim \text{Poisson}(\lambda_i)$ , which are independent.

# Poisson likelihood ratio

1.  $\Omega$ : For  $i = 1, \dots, n$ ,  $X_i \sim \text{Poisson}(\lambda_i)$  are independent.

$$\ell(\lambda_1, \dots, \lambda_n) =$$

Maximum loglikelihood under  $\Omega$  :  $\ell_1 =$

2.  $\Omega_0$ : Every  $\lambda_i = \lambda$ .

$$\ell(\lambda) =$$

Maximum likelihood under  $\Omega_0$  :  $\ell_0 =$

$$G = 2 \sum_{i=1}^n X_i \log \left( \frac{X_i}{\bar{X}} \right) \approx \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\bar{X}}$$

Suppose every  $\lambda_i = \lambda$ . For large  $n$ ,  $G \sim \chi_{n-1}^2$  approximately.

# Poisson dispersion test

1.  $H_0$ : rates are all equal.
2.  $H_1$ : rates are not all equal.
3. Sample mean and variance are 2.44 and 4.59.

$$g \approx \frac{\sum_{i=1}^{400} (x_i - \bar{x})^2}{\bar{x}} = \frac{399 \times 4.59}{2.44} \approx 751$$

4. The  $P$ -value is approximately

$$\Pr(\chi_{399}^2 \geq 751) \approx 0$$

The rates are likely different.

# Normal data

Based on  $N(\mu, \sigma^2)$  realisations  $x_1, \dots, x_n$ , might we conclude that  $\mu = 0$ ?

- ▶ Idea: if  $\bar{X}$  is far from 0, reject  $H_0 : \mu = 0$ .  $H_1 : \mu \neq 0$ .
- ▶ Suppose  $\sigma$  is known. Under  $H_0$ ,

$$\frac{\sqrt{n}\bar{X}}{\sigma} \sim N(0, 1)$$

- ▶ The  $P$ -value

$$\Pr\left(|Z| \geq \frac{\sqrt{n}|\bar{x}|}{\sigma}\right) = \Pr\left(Z^2 \geq \frac{n\bar{x}^2}{\sigma^2}\right)$$

The  $P$ -value is called two-tailed. One-tailed  $P$ -values correspond to  $H_1$  saying  $\mu > 0$  or  $\mu < 0$ .

# LR test on normal data

$$H_0 : \mu = 0. \quad H_1 : \mu \neq 0.$$

$\Omega = \mathbb{R}$ ,  $\Omega_0 = \{0\}$ . LR test can be applied.

1.  $\sigma$  is known.

$$G = \frac{n\bar{X}^2}{\sigma^2}$$

Under  $H_0$ ,  $G \sim \chi_1^2$  exactly. Identical to previous slide.

2.  $\sigma$  is unknown. For large  $n$ ,

$$G \approx \frac{n\bar{X}^2}{\hat{\sigma}^2}$$

Under  $H_0$ ,  $G \sim \chi_1^2$  approximately.

# Normal data: known $\sigma$

1.  $\Omega$ :

$$\ell(\mu) = -\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}$$
$$\ell_1 = -\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2} = -\frac{n\hat{\sigma}^2}{2\sigma^2}$$

2.  $\Omega_0$ :

$$\ell_0 = -\frac{\sum_{i=1}^n X_i^2}{2\sigma^2} = -\frac{n\hat{\mu}_2}{2\sigma^2}$$

3.

$$G = \frac{n(-\hat{\sigma}^2 + \hat{\mu}_2)}{\sigma^2} = \frac{n\bar{X}^2}{\sigma^2}$$

Suppose  $\mu = 0$ .  $G \sim \chi_1^2$ .

# Normal data: unknown $\sigma$

1.  $\Omega$ :

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}$$

$$\ell_1 = -\frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2}$$

2.  $\Omega_0$ :

$$\ell(\sigma) = -n \log \sigma - \frac{\sum_{i=1}^n X_i^2}{2\sigma^2}$$

$$\ell_0 = -\frac{n}{2} \log \hat{\mu}_2 - \frac{n}{2}$$

3.

$$G = n \log \left( \frac{\hat{\mu}_2}{\hat{\sigma}^2} \right) \approx \frac{n \bar{X}^2}{\hat{\sigma}^2}$$

Suppose  $\mu = 0$ . For large  $n$ ,  $G \sim \chi_1^2$  approximately.

# Conclusion (1)

- ▶ The LR test applies in many situations where the investigator wants to know the goodness-of-fit of a model *relative* to a larger model. If  $n$  is large, the  $P$ -value can be computed using a  $\chi^2$  distribution.
- ▶ The test assumes the larger model is valid, and does not assess *its* goodness-of-fit.
- ▶ A  $P$ -value is not a probability that  $H_0$  is true.  $H_0$  is either true or false.  $P$ -value is computed assuming  $H_0$  is true.
- ▶ Even if a model seems to fit the data, it may not mean the data were generated randomly according to the model. Plotting data in sequence is an important diagnosis. But for prediction, this point may not be so important.



## Conclusion (2)

- ▶ Statistical inference consists of two main areas: parameter estimation and hypothesis testing.
- ▶ MOM and ML are general estimation methods. To apply to data, a statistical model is needed. The procedures are blind to the goodness-of-fit of the model.
- ▶ We only test hypotheses relating to goodness-of-fit. But the general framework applies similarly when testing other hypotheses.
- ▶ Statistical inference can be quite procedural. Watch out for pitfalls in other aspects of data analysis, such as the choice of a model.