Lecture 9: Asymptotic Tests, Pearson's χ^2 Test

STA261 – Probability & Statistics II

Ofir Harari

Department of Statistical Sciences

University of Toronto

UNIVERSITY OF TORONTO

Outline

UNIVERSITY OF TORONTO

Wilks' Theorem

- Recall that we defined the Generalized Likelihood Ratio Test (GLRT) at level α for testing $\mathcal{H}_0 : \theta \in \Theta_0$ vs. $\mathcal{H}_1 : \theta \in \Theta_1$ (such that the entire parameter space is $\Theta = \Theta_0 \cup \Theta_1$) to be the test corresponding to the rejection region

$$\mathcal{C} = \left\{ \Lambda(\underline{X}) := \frac{\max_{\theta \in \Theta} \mathcal{L}(\theta)}{\max_{\theta \in \Theta_0} \mathcal{L}(\theta)} \geq c \right\},$$

  with the critical value c satisfying $\mathbb{P}\left(\Lambda(\underline{X}) \geq c \middle| \theta \in \Theta_0 \right) = \alpha$.

- This lecture will be dedicated to testing hypotheses of a specific form: suppose that we observe a random sample $X_1, \ldots, X_n \sim f_{\underline{\theta}}$, where $\underline{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$ is a vector of parameters, and we wish to test the null hypothesis

$$\mathcal{H}_0 : \theta_1 = \theta_1^0,\ \theta_2 = \theta_2^0, \ldots, \theta_r = \theta_r^0 \quad (1 \leq r \leq p)$$

  against the unrestricted alternative.

- We shall present and prove a general result, before proceeding to discuss several famous special cases.

UNIVERSITY OF TORONTO

Wilks' Theorem (cont.)

Theorem

Suppose the data and $\mathcal{H}_0$ are as in the previous slide, and let $\Lambda(\underline{X})$ be the resultant GLR statistic. Then, under similar regularity conditions to those required for the asymptotic Normality of the MLE,

$$2\log\Lambda(\underline{X}) \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_r,$$

where r is the number of parameters constrained by $\mathcal{H}_0$.

Proof for $p = r = 1$: EX. mean and variance for normal

Here we prove the result for $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_1 : \theta \neq \theta_0$. Proving the general result requires knowledge of multivariate analysis.

Let $\widehat{\theta}_n$ be the MLE of θ, and write the Taylor expansion of the log-likelihood at θ_0 about $\widehat{\theta}_n$: including a Lagrange reminder

$$\ell(\theta_0) = \ell(\hat{\theta}_n) + \ell'(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{\ell''(\theta_n^*)}{2}(\theta_0 - \hat{\theta})^2,$$

where θ_n^* lies somewhere between θ_0 and $\widehat{\theta}_n$ (this is the Lagrange/mean value form of the remainder).

Wilks' Theorem (cont.)

Proof (cont):

$$\ell(\theta_0) = \ell(\hat{\theta}_n) + \ell'(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{\ell''(\theta_n^*)}{2}(\theta_0 - \hat{\theta})^2 = \ell(\hat{\theta}_n) + \frac{\ell''(\theta_n^*)}{2}(\theta_0 - \hat{\theta})^2,$$

derivative of log likelihood vanishes

since $\ell'(\hat{\theta}_n) = 0$. Suppose that $\mathcal{H}_0 : \theta = \theta_0$ is true, then –

- We have already proven (in Lecture 3) that $-\frac{\ell''(\theta_0)}{n} \xrightarrow{P} \mathcal{I}^*(\theta_0)$

consistency of MLE since theta^* even closer for one observation

- We also know that $\widehat{\theta}_n \xrightarrow{P} \theta_0$, thus $\theta_n^* \xrightarrow{P} \theta_0$ too, and if $\ell(\theta)$ is twice continuously differentiable then $\ell''(\hat{\theta}_n) \xrightarrow{P} \ell''(\theta_0)$ and $\ell''(\theta_n^*) \xrightarrow{P} \ell''(\theta_0)$.

invariance of MLE over cont. transformation

- Recall that $\sqrt{n}(\theta_0 - \hat{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{I}^{*-1}(\theta_0)\right) = \{\mathcal{I}^*(\theta_0)\}^{-1/2}\mathcal{N}(0,1)$, hence

by asymptotic normality of MLE

$$\left[\sqrt{n}(\theta_0 - \hat{\theta})\right]^2 \xrightarrow{\mathcal{D}} \mathcal{I}^{*-1}(\theta_0) \cdot \chi_1^2.$$

take out variance & square

- Now,

$$2\log\Lambda(\underline{X}) = 2\log\frac{\mathcal{L}(\hat{\theta}_n)}{\mathcal{L}(\theta_0)} = 2\left\{\ell(\hat{\theta}_n) - \ell(\theta_0)\right\} = \ell''(\theta_n^*)(\theta_0 - \hat{\theta}_n)^2$$

by 2nd order Taylor expansion

$$= -\frac{\ell''(\theta_0)}{n} \cdot \frac{\ell''(\theta_n^*)}{\ell''(\theta_0)}\left[\sqrt{n}(\theta_0 - \hat{\theta}_n)\right]^2.$$

Wilks' Theorem (cont.)

Proof (cont):

Finally,

$$2\log\Lambda(\underline{X}) = \underbrace{-\frac{\ell''(\theta_0)}{n}}_{\xrightarrow{P}\mathcal{I}^*(\theta_0)} \cdot \underbrace{\frac{\ell''(\theta_n^*)}{\ell''(\theta_0)}}_{\xrightarrow{P}1} \cdot \underbrace{\left[\sqrt{n}(\theta_0 - \hat{\theta}_n)\right]^2}_{\xrightarrow{\mathcal{D}}\mathcal{I}^{*-1}(\theta_0)\cdot\chi_1^2} \xrightarrow{\mathcal{D}} \chi_1^2,$$

by a simple application of Slutsky's Theorem.

Samuel S. Wilks, 1906-1964

Source: genealogy.cornerfamily.com

Example: test of equality of Poisson means

- Let $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_X)$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda_Y)$ be two independent samples, and suppose that we want to test $\mathcal{H}_0 : \lambda_X = \lambda_Y$ vs. $\mathcal{H}_1 : \lambda_X \neq \lambda_Y$.

- We can reparameterize: $\theta_1 = \lambda_X / \lambda_Y$ and $\theta_2 = \lambda_Y$ – then it is easy to see that the null hypothesis can be rewritten as $\mathcal{H}_0 : \theta_1 = 1$, and we are in the realm of Wilks' Theorem (with $r = 1$ restricted parameters under $\mathcal{H}_0$).

- Here the (unrestricted) likelihood is given by

$$\mathcal{L}(\lambda_X, \lambda_Y) = \mathrm{e}^{-m\lambda_X} \frac{\lambda_X^{\sum_{i=1}^m X_i}}{\prod_{i=1}^m X_i!} \times \mathrm{e}^{-n\lambda_Y} \frac{\lambda_Y^{\sum_{i=1}^n Y_i}}{\prod_{i=1}^n Y_i!},$$

same as finding MLE for each of X and Y

and the MLEs are $\widehat{\lambda}_X = \overline{X}$ and $\widehat{\lambda}_Y = \overline{Y}$, leading to

$$\mathcal{L}(\widehat{\lambda}_X, \widehat{\lambda}_Y) = \mathrm{e}^{-m\overline{X}} \frac{\left(\overline{X}\right)^{m\overline{X}}}{\prod_{i=1}^m X_i!} \times \mathrm{e}^{-n\overline{Y}} \frac{\left(\overline{Y}\right)^{n\overline{Y}}}{\prod_{i=1}^n Y_i!}.$$

product of two independent poisson

Test of equality of Poisson means (cont.)

under null, lambda same

- The restricted likelihood under $\mathcal{H}_0 : \lambda_X = \lambda_Y = \lambda_0$ is

$$\mathcal{L}(\lambda) = \mathrm{e}^{-(m+n)\lambda} \frac{\lambda^{\{\sum_{i=1}^m X_i + \sum_{i=1}^n Y_i\}}}{\prod_{i=1}^m X_i! \prod_{i=1}^n Y_i!},$$

in essence, sum of observation over sample size

and the restricted MLE is $\widehat{\lambda}_0 = \dfrac{m\overline{X} + n\overline{Y}}{m+n}$, leading to

$$\mathcal{L}(\widehat{\lambda}_0) = \mathrm{e}^{-m\overline{X} - n\overline{Y}} \frac{\left(\dfrac{m\overline{X} + n\overline{Y}}{m+n}\right)^{m\overline{X}+n\overline{Y}}}{\prod_{i=1}^m X_i! \prod_{i=1}^n Y_i!}.$$

- The GLR is then

$$\Lambda = \frac{\mathcal{L}(\widehat{\lambda}_X, \widehat{\lambda}_Y)}{\mathcal{L}(\widehat{\lambda}_0)} = \frac{\left(\overline{X}\right)^{m\overline{X}} \left(\overline{Y}\right)^{n\overline{Y}}}{\left(\dfrac{m\overline{X} + n\overline{Y}}{m+n}\right)^{m\overline{X}+n\overline{Y}}}.$$

do not know distribution of sum of two independent Poisson means

Test of equality of Poisson means (cont.)

take log

$$\Lambda = \left(\overline{X}\right)^{m\overline{X}} \left(\overline{Y}\right)^{n\overline{Y}} \left(\frac{m\overline{X} + n\overline{Y}}{m+n}\right)^{-m\overline{X}-n\overline{Y}}$$

- From Wilks' Theorem, $2\log\Lambda \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_1$
- A rejection region of an asymptotic test at level α is then given by

$$\mathcal{C} = \left\{ 2\left[m\overline{X}\log\overline{X} + n\overline{Y}\log\overline{Y} - (m\overline{X} + n\overline{Y})\log\left(\frac{m\overline{X} + n\overline{Y}}{m+n}\right)\right] \geq \chi^2_{1,1-\alpha} \right\}$$

- For example, for $m = 30$, $n = 50$, $\overline{X} = 2.2$ and $\overline{Y} = 3$, we have

$$2\log\Lambda = 4.58 > 3.84 = \chi^2_{1,0.95},$$

hence we reject $\mathcal{H}_0 : \lambda_X = \lambda_Y$ at the 5% level.

- This can also be verified by calculating

reject null

$$\text{p-value} = \mathbb{P}\left(2\log\Lambda \geq 4.58 \middle| \lambda_X = \lambda_Y\right) = 0.0323 \quad \text{(calculating for } \chi^2_1 \text{ using R)}$$

UNIVERSITY OF TORONTO

GLRT by Parametric Bootstrap

- If, for whatever reason, one does not trust asymptotics, "exact" testing is also an option, via simulation:
  1. Sample data under $\mathcal{H}_0$: two independent random Poisson samples with the same λ of sizes 30 and 50.
  2. Evaluate the test statistic $2\log\Lambda$ at the simulated data.
  3. Repeat for a very large number of times: N
  4. The empirical p-value is $\dfrac{\#\text{ samples for which } 2\log\Lambda \geq 4.58}{N}$.
- The above technique is famously known as *parametric bootstrap*, and is gaining popularity by the day, along with the increase of computing power.
- Note that the bootstrap p-value converges in probability to the true p-value as $N \to \infty$ regardless of the size of the data, making it a terrific option for testing under small sample sizes.

UNIVERSITY OF TORONTO

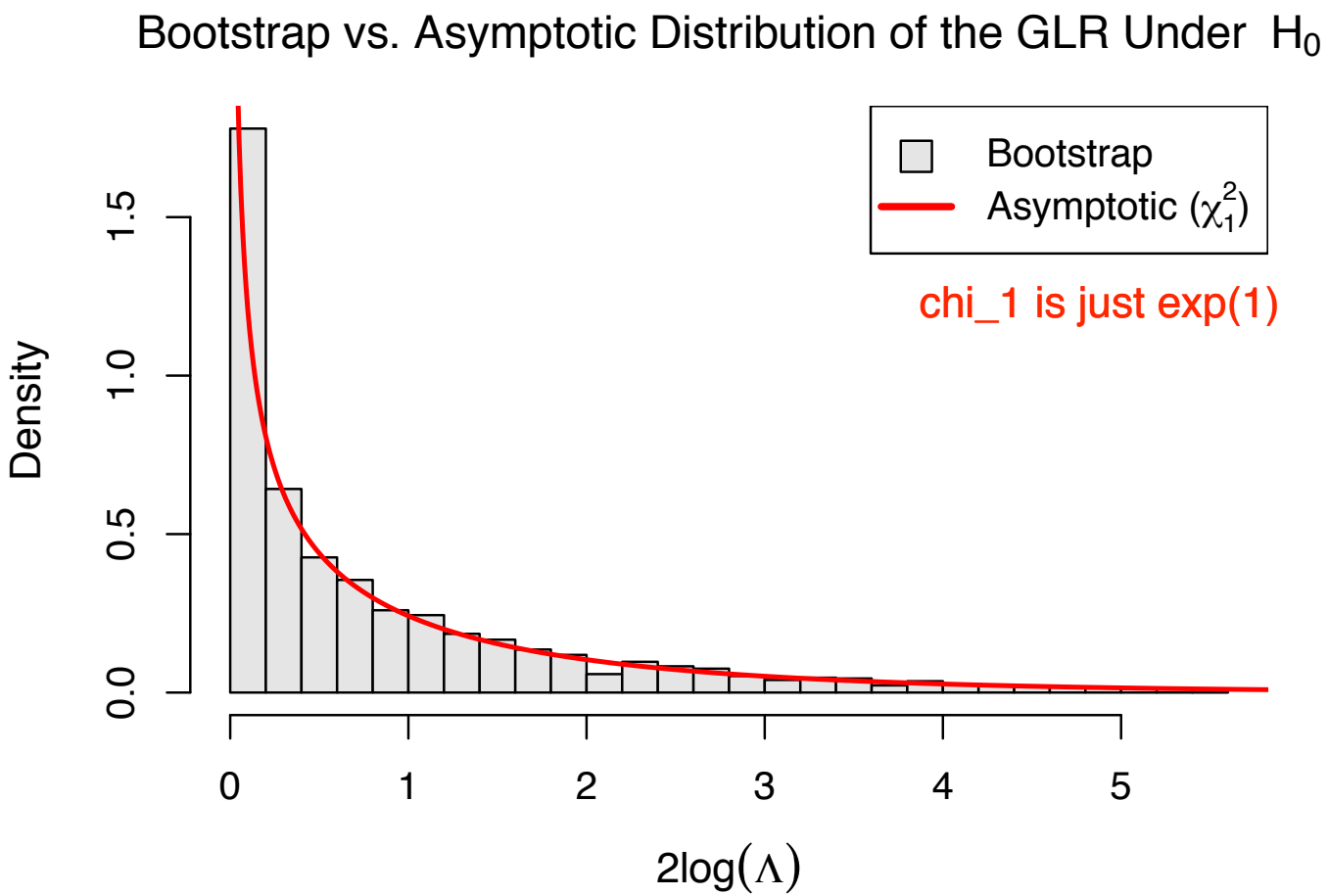
GLRT by Parametric Bootstrap (cont.)

```
> x <- matrix(rpois(m*N, lambda=1), ncol=m)
> y <- matrix(rpois(n*N, lambda=1), ncol=n)
>
> xBar <- apply(x, 1, mean)
> yBar <- apply(y, 1, mean)
>
> T <- m*xBar*log(xBar) + n*yBar*log(yBar)
> T <- T - (m*xBar+n*yBar)*log((m*xBar+n*yBar)/(m+n))
> T <- 2*T #a random sample of the GLR test statistic
>
> (p_value <- sum(T > 4.58)/N) #Bootstrap p-value

[1] 0.032758

> 1-pchisq(4.58, df=1) #asymptotic p-value

[1] 0.03234721
```

asymptotics works very well in this case

GLRT by Parametric Bootstrap (cont.)

```
> hist(T, freq=FALSE, xlab=expression(2*log(Lambda)), col='grey90')
> t <- seq(0, 20, by=.01)
> lines(t, dchisq(t, df=1), lwd=2, col=2)
```



UNIVERSITY OF TORONTO

A Goodness of fit test (cont.)

- Consider the following problem: you play a game of Backgammon with your friend, who insists on using his own pair of dice. The whole thing seems a bit fishy, and you suspect that his dice could be biased. You roll one of his dice 996 times, and record the following counts:

Outcome	1	2	3	4	5	6	Total
Count	140	150	164	170	180	190	996

166

- If the die is fair, then $X \sim \mathrm{U}\{1, \ldots, 6\}$, where X is the outcome of rolling the die once.

- Denoting $p_j = \mathbb{P}(X = j)$, $j = 1, \ldots, 6$, we wish to test

$$\begin{cases} \mathcal{H}_0 : p_1 = \cdots p_6 = \frac{1}{6} \\ \mathcal{H}_1 : \text{ otherwise} \end{cases}$$

at level α.

- Let us consider a more general case.

UNIVERSITY OF TORONTO

A Goodness of fit test (cont.)

- Suppose now that $X_1, \ldots, X_n$ is a random sample from a discrete distribution with k possible values $s_1, \ldots, s_k$, with corresponding probabilities $p_1, \ldots, p_k$ – i.e. $p_j = \mathbb{P}(X = s_j)$. For each X_i
- Denote $O_j = \#\{i : X_i = s_j\}$, then the likelihood/log-likelihood can be expressed as number of times we observe outcome as s_j

$$\mathcal{L}(p_1, \ldots, p_k) = \prod_{i=1}^{k} p_i^{O_i} \quad ; \quad \ell(p_1, \ldots, p_k) = \sum_{i=1}^{n} O_i \log p_i$$

- What are the MLEs of $p_1, \ldots, p_k$?
- Recall that they must add up to 1, hence we face a constrained optimization problem –

$$\max_{p_1,\ldots,p_k} \left\{ \sum_{i=1}^{n} O_i \log p_i \quad \text{s.t.} \quad \sum_{i=1}^{k} p_i = 1 \right\}$$

A Goodness of fit test (cont.)

$$\max_{p_1,\ldots,p_k} \left\{ \sum_{i=1}^{n} O_i \log p_i \quad \text{s.t.} \quad \sum_{i=1}^{k} p_i = 1 \right\}$$

- Solution by the method of Lagrange Multipliers:
  1. Write the Lagrangian:

$$\mathscr{L}(p_1, \ldots, p_k, \lambda) = \sum_{i=1}^{n} O_i \log p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

  2. Solve

$$\frac{\partial \mathscr{L}}{\partial p_j} = \frac{O_j}{p_j} + \lambda = 0 \Longrightarrow \widehat{p}_j = -\frac{O_j}{\lambda}, \; j = 1, \ldots, k$$

  3. Solve

$$\frac{\partial \mathscr{L}}{\partial \lambda} = \sum_{i=1}^{k} p_i - 1 = 0 \Longrightarrow \sum_{j=1}^{k} \widehat{p}_j = -\frac{1}{\lambda} \underbrace{\sum_{j=1}^{k} O_j}_{n} = 1$$

$$\Longrightarrow \lambda = -n \Longrightarrow \widehat{p}_j = \frac{O_j}{n}.$$

now we try to find \lambda

A Goodness of fit test (cont.)

- Suppose now that we wish to test

$$\mathcal{H}_0 : p_1 = p_1^0,\ p_2 = p_2^0,\ \ldots,\ p_k = p_k^0$$

  vs. the unrestricted alternative.

- Recall that the unrestricted likelihood is given by $\mathcal{L}(p_1, \ldots, p_k) = \prod_{i=1}^k p_i^{O_i}$, and we just showed that it is maximized by $\widehat{p}_j = \dfrac{O_j}{n},\ j = 1, \ldots, k,$ hence

$$\mathcal{L}(\widehat{p}_1, \ldots, \widehat{p}_k) = \prod_{i=1}^k \left(\frac{O_i}{n}\right)^{O_i}.$$

- The restricted likelihood (under $\mathcal{H}_0$) is $\mathcal{L}(p_1^0, \ldots, p_k^0) = \prod_{i=1}^k \left(p_i^0\right)^{O_i}$, and the GLR statistic is thus

$$\Lambda = \frac{\mathcal{L}(\widehat{p}_1, \ldots, \widehat{p}_k)}{\mathcal{L}(p_1^0, \ldots, p_k^0)} = \prod_{i=1}^k \left(\frac{O_i}{np_i^0}\right)^{O_i}.$$

UNIVERSITY OF TORONTO

A Goodness of fit test (cont.)

$$\Lambda = \prod_{i=1}^{k} \left(\frac{O_i}{np_i^0} \right)^{O_i}$$

- Recall that we denoted $O_j := \#\{i : X_i = s_j\}$ – the *observed* j^{th} cell count
- We may also denote $E_j := np_j^0$ – the *expected* j^{th} cell count under $\mathcal{H}_0$ – then the GLR can be rewritten as

$$\Lambda = \prod_{i=1}^{k} \left(\frac{O_i}{E_i} \right)^{O_i}.$$

- Note that the number of restrictions in $\mathcal{H}_0 : p_1 = p_1^0,\ p_2 = p_2^0,\ \ldots,\ p_k = p_k^0$ is in fact $k-1$ – since the p_j's sum to 1 – thus, from Wilks' Theorem

$$2\log\Lambda = 2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_{k-1}.$$

the deviance; Large deviance implies rejection of null

Pearson's $\mathcal{X}^2$ Statistic

$$2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_{k-1}$$

- Writing the 2^{nd} order Taylor expansion of $\log O_i$ about E_i –

$$\log O_i \approx \log E_i + \frac{O_i - E_i}{E_i} - \frac{(O_i - E_i)^2}{2E_i^2},$$

we have

$$\begin{aligned}
2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) &= 2\sum_{i=1}^{k} O_i \left(\log O_i - \log E_i\right) \\
&\approx 2\sum_{i=1}^{k} O_i \left\{\frac{O_i - E_i}{E_i} - \frac{(O_i - E_i)^2}{2E_i^2}\right\} \\
&= 2\sum_{i=1}^{k} \left[(O_i - E_i) + E_i\right] \left\{\frac{O_i - E_i}{E_i} - \frac{(O_i - E_i)^2}{2E_i^2}\right\}
\end{aligned}$$

UNIVERSITY OF TORONTO

Pearson's $\mathcal{X}^2$ Statistic (cont.)

$$2\sum_{i=1}^{k} O_i \log\left(\frac{O_i}{E_i}\right) \approx 2\sum_{i=1}^{k} [(O_i - E_i) + E_i] \left\{ \frac{O_i - E_i}{E_i} - \frac{(O_i - E_i)^2}{2E_i^2} \right\}$$

$$= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} - \sum_{i=1}^{k} \frac{(O_i - E_i)^3}{E_i^2} + 2\underbrace{\sum_{i=1}^{k} O_i}_{n} - 2\underbrace{\sum_{i=1}^{k} E_i}_{n}$$

total number of counts

$$= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} - \sum_{i=1}^{k} \frac{(O_i - E_i)^3}{E_i^2}$$

- The second term can be shown to become negligible (in probability) as the count of each cell goes to infinity
- As a result,

$$\mathcal{X}^2 := \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_{k-1}$$

- *Pearson's* χ^2 *test* at level α is the one rejecting $\mathcal{H}_0$ for $\mathcal{X}^2 \geq \chi^2_{k-1,1-\alpha}$

Pearson's χ^2 test of goodness of fit

Karl Pearson, 1857-1936

Source: wikimedia.org

UNIVERSITY OF TORONTO

Example

Back to the fair(?) die example –

- the "O" table is

Outcome	1	2	3	4	5	6	Total
Count	140	150	166	170	180	190	996

- The "E" table is calculated such that the j^{th} cell count is $E_j = np_j$ – here it will be

Outcome	1	2	3	4	5	6	Total
Count	166	166	166	166	166	166	996

- To test at the 5% level, the rejection region is $\mathcal{C} = \left\{\mathcal{X}^2 \geq \chi^2_{5,0.95} = 11.07\right\}$

df = 5 = 6 - 1

- Here

$$\mathcal{X}^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(140-166)^2}{166} + \frac{(150-166)^2}{166} + \frac{(166-166)^2}{166}$$
$$+ \frac{(170-166)^2}{166} + \frac{(180-166)^2}{166} + \frac{(190-166)^2}{166} = 10.36$$

UNIVERSITY OF TORONTO

Example (cont.)

$$\mathcal{C} = \left\{ \mathcal{X}^2 \geq \chi^2_{5,0.95} = 11.07 \right\} \;\; ; \;\; \mathcal{X}^2 = 10.36$$

- As it stands, at the 5% level the evidence against $\mathcal{H}_0$ (fair die) is not strong enough.

- Calculating the p-value –

$$\text{p-value} = \mathbb{P}\left(\mathcal{X}^2 \geq 10.36 \middle| \text{fair die} \right) = \mathbb{P}\left(\chi^2_5 \geq 10.36 \right)$$

  ⋆ Using the χ^2 table:

$$0.9 < \mathbb{P}\left(\chi^2_5 \leq 10.39 \right) < 0.95 \Longrightarrow 0.05 < \text{p-value} < 0.1$$

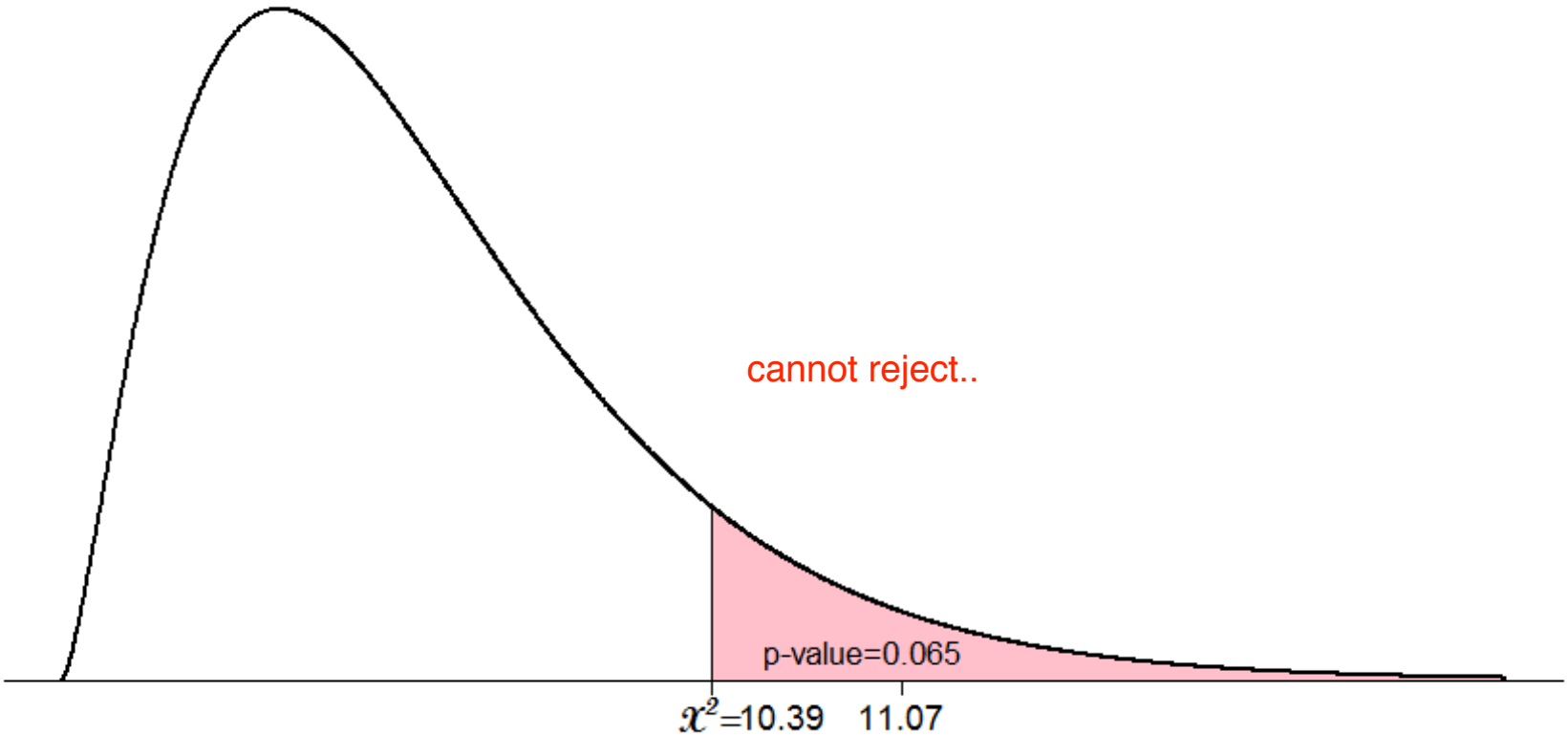
  ⋆ Using R: p-value $= 0.066$

- Here the original GLR statistic would yield an almost identical result -

$$2 \log \Lambda = 2 \sum_{i=1}^{k} O_i \log \left(\frac{O_i}{E_i} \right) = 10.46, \quad \text{p-value} = 0.063.$$

the one not simplified by pearson…

Example (cont.)



UNIVERSITY OF TORONTO

Degrees of freedom in the Goodness of Fit Test

d.f. = dim Theta - dim Theta_0

- One way to count the number of parameters restricted by $\mathcal{H}_0$:

$$\begin{array}{ccccc} \text{No. of params.} & & \text{No. of "free"} & & \text{No. of "free"} \\ \text{restricted} & = & \text{parameters in} & - & \text{parameters} \\ \text{by } \mathcal{H}_0 & & \text{the problem} & & \text{under } \mathcal{H}_0 \end{array}$$

- Suppose now that we wish to test the hypothesis that the data follows a Poisson distribution, and λ is unknown.

need to estimate lambda parameter first

- To make the asymptotics "work", we group the counts into k relatively large cells (0-1, 2-4, 5-8, etc.).

impose on last cell probability such that summation of p is 1

- Under $\mathcal{H}_1$, there are $k-1$ unrestricted cell probabilities.

- Under $\mathcal{H}_0$, there is one unrestricted parameter (λ), and the cell probabilities are determined by its value. λ will be estimated from the data.

- From the above, the test has $k-1-1 = k-2$ degrees of freedom.

once know lambda, we can calculate expected p under the grouped cells

UNIVERSITY OF TORONTO

Degrees of freedom in the Goodness of Fit Test

- In general, the number of degrees of freedom in a goodness of fit test will be –

$$\text{df} = \begin{array}{c}\text{Number}\\ \text{of cells}\end{array} - \begin{array}{c}\text{Number of}\\ \text{estimated}\\ \text{parameters}\end{array} - 1$$

lambda — the last cell restricted by the sum

- For example, if we wish to test the hypothesis that some data follows the Normal distribution, based on observations grouped into k cells (without any particular parameter values in mind), the corresponding χ^2 goodness of fit test will have $k - 3$ degrees of freedom (2 parameters to estimate).

UNIVERSITY OF TORONTO

Categorical data

A test of independence

Consider the following problem: The owner of a laboratory wants to keep sick leave low, by preventing his employees from contracting pneumonia. There is a vaccine for pneumococcal pneumonia, but due to a production problem at the company that produces the vaccine, only half of the employees receive the vaccine. The company sent a nurse to every employee who contracted pneumonia to take a sputum sample for culture to determine the causative agent. They kept track of the number of employees who contracted pneumonia and which type of pneumonia each had. They want to know if providing the vaccine makes a difference.

		Vaccination Status		
		Unvaccinated	**Vaccinated**	**Total**
Health Outcome	**Pneumoccal pneumonia**	38	18	56
	Other pneumonia	16	20	36
	No pneumonia	130	146	276
	Total	184	184	368

UNIVERSITY OF TORONTO

A test of independence (cont.)

- In this example we have two *categorical* (or *qualitative*) variables:
  - "Vaccination status" – takes values in {"Unvaccinated", "Vaccinated"}
  - "Health Outcome" – takes values in {"Pneumoccal pneumonia", "Other pneumonia", "No pneumonia"}
- We would like to answer the question "is there a difference in incidence of pneumonia between the two vaccination groups?" (in the entire population)
- Equivalently, we may test

$$\begin{cases} \mathcal{H}_0: & \text{"Vaccination status" and "Health Outcome" are independent} \\ \mathcal{H}_1: & \text{otherwise} \end{cases}$$

- The derivation of the test is almost identical to that of the goodness of fit test. I will spare you the technical details and illustrate it with the example from the previous slide.

UNIVERSITY OF TORONTO

A test of independence (cont.)

- As in the case of testing for goodness of fit, we denote the observed *contingency table* by 'O'.
- Our goal is now to calculate the 'E' table: the table comprising the expected cell counts under $\mathcal{H}_0$.
- Recall that (statistical) independence implies

$$\mathbb{P}\left(\begin{array}{c}\text{Vaccination}\\ \text{Status}\end{array} = a\,,\ \begin{array}{c}\text{Health}\\ \text{Outcome}\end{array} = b\right)$$

$$= \mathbb{P}\left(\begin{array}{c}\text{Vaccination}\\ \text{Status}\end{array} = a\right)\mathbb{P}\left(\begin{array}{c}\text{Health}\\ \text{Outcome}\end{array} = b\right)$$

- Next we will estimate the marginal distributions of the two variables, using the table margins.

margins of O table is fixed

UNIVERSITY OF TORONTO

A test of independence (cont.)

		Vaccination Status		
		Unvaccinated	**Vaccinated**	**Total**
Health Outcome	**Pneumoccal pneumonia**	E_{11}	E_{12}	56
	Other pneumonia	E_{21}	E_{22}	36
	No pneumonia	E_{31}	E_{32}	276
	Total	184	184	368

- How can we estimate $\mathbb{P}\left(\begin{array}{c}\text{Vaccination}\\ \text{Status}\end{array} = \text{“Unvaccinated”}\right)$?

  Very naturally, by $\dfrac{\#\text{unvaccinated employees}}{\text{total \# of employees}} = \dfrac{184}{368}$

- Similarly, we estimate $\mathbb{P}\left(\begin{array}{c}\text{Health}\\ \text{Outcome}\end{array} = \begin{array}{c}\text{“Pneumoccal}\\ \text{pneumonia”}\end{array}\right)$ by

  $$\frac{\#\text{employees with pneumoccal pneumonia}}{\text{total \# of employees}} = \frac{56}{368}$$

A test of independence (cont.)

Under $\mathcal{H}_0$ (independence), we estimate

$$\mathbb{P}\left(\begin{array}{c}\text{Vaccination}\\ \text{Status}\end{array} = \text{“Unvaccinated”}, \begin{array}{c}\text{Health}\\ \text{Outcome}\end{array} = \begin{array}{c}\text{“Pneumoccal}\\ \text{pneumonia”}\end{array}\right)$$

$$= \mathbb{P}\left(\begin{array}{c}\text{Vaccination}\\ \text{Status}\end{array} = \text{“Unvaccinated”}\right)\mathbb{P}\left(\begin{array}{c}\text{Health}\\ \text{Outcome}\end{array} = \begin{array}{c}\text{“Pneumoccal}\\ \text{pneumonia”}\end{array}\right) \stackrel{\frown}{=} \frac{184}{368}\cdot\frac{56}{368},$$

and the expected cell count is thus

$$E_{11} = 368 \cdot \frac{184}{368}\cdot\frac{56}{368} = \frac{184\cdot 56}{368} = 28.$$

		Vaccination Status		
		Unvaccinated	**Vaccinated**	**Total**
Health Outcome	**Pneumoccal pneumonia**	$E_{11} = 28$	E_{12}	56
	Other pneumonia	E_{21}	E_{22}	36
	No pneumonia	E_{31}	E_{32}	276
	Total	184	184	368

UNIVERSITY OF TORONTO

A test of independence (cont.)

	Unvaccinated	Vaccinated	Total
Pneumoccal pneumonia	$E_{11} = 28$	E_{12}	56
Other pneumonia	E_{21}	E_{22}	36
No pneumonia	E_{31}	E_{32}	276
Total	184	184	368

- Similarly, we will have

$$E_{21} = \frac{184 \cdot 36}{368} = 18$$

	Unvaccinated	Vaccinated	Total
Pneumoccal pneumonia	$E_{11} = 28$	E_{12}	56
Other pneumonia	$E_{21} = 18$	E_{22}	36
No pneumonia	E_{31}	E_{32}	276
Total	184	184	368

- At this point, all other cell counts are determined by the constraints imposed by the table margins!

A test of independence (cont.)

O table:

	Unvaccinated	Vaccinated
Pneumoccal pneumonia	38	18
Other pneumonia	16	20
No pneumonia	130	146

E table:

	Unvaccinated	Vaccinated
Pneumoccal pneumonia	28	28
Other pneumonia	18	18
No pneumonia	138	138

- The test statistics this time will be

$$\mathcal{X}^2 = \sum_{i=1}^{\#\text{rows}} \sum_{j=1}^{\#\text{cols}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

A Test of independence (cont.)

$$\mathcal{X}^2 = \sum_{i=1}^{\#\text{rows}} \sum_{j=1}^{\#\text{cols}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(38-28)^2}{28} + \frac{(18-28)^2}{28} + \frac{(16-18)^2}{18} + \frac{(20-18)^2}{18}$$

$$+ \frac{(130-138)^2}{138} + \frac{(146-138)^2}{138} = 8.51$$

- What is the distribution of $\mathcal{X}^2$ under $\mathcal{H}_0$?
- Wilks' Theorem: $\mathcal{X}^2 \xrightarrow[\mathcal{H}_0]{\mathcal{D}} \chi^2_d$, where d is the number of parameters "restricted" by $\mathcal{H}_0$.
- One way to count the number of parameters restricted by $\mathcal{H}_0$:

$$\begin{array}{ccccc} \text{No. of params.} & & \text{No. of "free"} & & \text{No. of "free"} \\ \text{restricted} & = & \text{parameters in} & - & \text{parameters} \\ \text{by } \mathcal{H}_0 & & \text{the problem} & & \text{under } \mathcal{H}_0 \end{array}$$

A Test of independence (cont.)

- What is the total number of "free" parameters in the problem?
  - If the table has r rows and c columns, there are $r \times c$ unknown cell probabilities
  - However, fixing any $r \cdot c - 1$ of them forces the remaining one to complete all the others to 1. Hence there are $r \cdot c - 1$ "free" parameters.

under alternative, the final cell count has to complement everything else

- How many "free" parameters are there under $\mathcal{H}_0$?
  - Once the *marginal* probabilities

under null, every cell count is restricted

$$\mathbb{P}\left(\begin{array}{c}\text{Rows}\\ \text{variable}\end{array} = a_i\right), \quad i = 1, \ldots, r \quad \text{and} \quad \mathbb{P}\left(\begin{array}{c}\text{Columns}\\ \text{variable}\end{array} = b_j\right), \quad j = 1, \ldots, c$$

are fixed, the independence assumption leaves no choice to the cell probabilities but to satisfy

$$p_{ij} = \mathbb{P}(\text{Rows variable} = a_i)\,\mathbb{P}(\text{Columns variable} = b_j).$$

  - We are thus only "free" to choose $(r-1) + (c-1)$ marginal probabilities.

the last marginal count is fixed to satisfy constraint (sum=1)

r*c - 1 - (r - 1 + c - 1) = r(c-1) - (c-1) = (r-1)(c-1)

A Test of independence (cont.)

- The number of parameters restricted by $\mathcal{H}_0$ (for an $r \times c$ table) is thus

$$d = rc - 1 - (r-1) - (c-1) = rc - r - c + 1 = (r-1)(c-1),$$

  and we reject $\mathcal{H}_0$ at level α if $\mathcal{X}^2 \geq \chi_{(r-1)(c-1),1-\alpha}$.

- In the pneumonia example we had a 3×2 table, so the number of degrees of freedom is $(3-1)(2-1) = 2$. This is also the number of expected cell counts we had to calculate before the margins dictated the rest.

- The rejection region (at the 5% level) is thus $\mathcal{C} = \left\{\mathcal{X}^2 \geq \chi^2_{2,0.95} = 5.99\right\}$

- We calculated $\mathcal{X}^2 = 8.51$, so at the 5% we reject $\mathcal{H}_0$ (and conclude that there is dependence between the vaccination status and the health outcome).

- Can also calculate (using R)

$$\text{p-value} = \mathbb{P}\left(\mathcal{X}^2 \geq 8.51 \middle| \mathcal{H}_0\right) = \mathbb{P}\left(\chi^2_2 \geq 8.51\right) = 0.014.$$