Therefore, $X_i \sim B(n, p_i) \Rightarrow Var(X_i) = np_i(1-p_i)$ — large when $p = 1/2$, small when $p \approx 0$ or 1

observed counts

$$\Lambda = \prod_{i=1}^{m} \left(\frac{np_{i0}}{x_i}\right)^{x_i} \quad \text{and} \quad -2\log\Lambda = 2\sum_{i=1}^{m} x_i \log\left(\frac{x_i}{np_{i0}}\right).$$

reasonable?

Q. Why multipled by x_i?

expected counts

Consider, e.g., $n=1000$ ① $x_i = 2$, $np_{i0} = 1$ ② $x_i = 200$, $np_{i0} = 100$

– H_0 is rejected if

$$-2\log\Lambda(X_1, \ldots, X_m) \geq c.$$

– The constant c can be determined by the fact that under H_0,

$$\text{test statistic} \rightarrow -2\log\Lambda \xrightarrow{D} \chi^2_{m-1} \leftarrow \text{asymptotic null distribution}$$

(i.e., $-2\log\Lambda$ is asymptotically χ^2_{m-1} distributed when $n \to \infty$) because

$$\dim(\Omega) - \dim(\Omega_0) = (m-1) - 0 = m-1.$$

Note. Not $m \to \infty$ (LN,CH8,p.86)

5/6

Question 7.14 (examine distribution assumption in statistical modeling $\Rightarrow$ goodness of fit)

- In statistical modeling, we often see the statement:

  $X_1, \ldots, X_n$ are i.i.d. from a distribution with cdf $F(\cdot|\Theta)$.

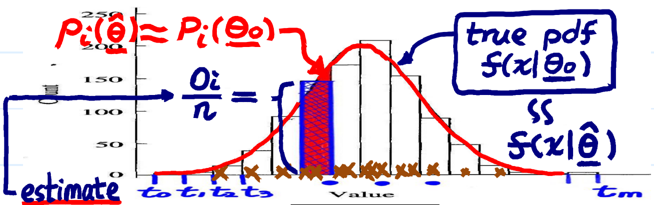
  determine useful & useless information. Recall: sufficient statistics

- Suppose that the independent and identical assumptions are true, can we examine (using data) whether the distribution assumption (i.e., $F(\cdot|\Theta)$) is reasonable?

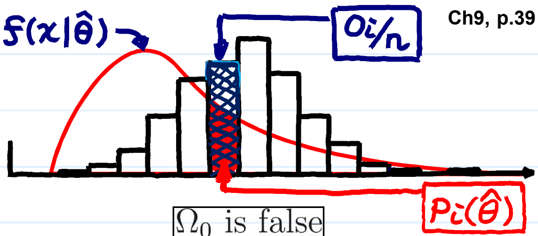
e.g. $\dim(\Omega_A) = \infty$ in LNp.5

original statistical model $\Omega = \{F(\cdot|\Theta)\} \xrightarrow{\text{enlarge}} \Omega = \overbrace{\{F(\cdot|\Theta)\}}^{\Omega_0} \cup \overbrace{\{\text{other distributions}\}}^{\Omega_A}$

Recall. histogram (直方圖) $\updownarrow$ pdf/pmf



Ω_0 is true



Ω_0 is false

- Observe data $X_1, \ldots, X_n$. $\quad (I_{(t_0,t_1]}(X_i), I_{(t_1,t_2]}(X_i), \cdots, I_{(t_{m-1},t_m]}(X_i))$ sum over i
- Let O_i, $i = 1, \ldots, m$, be the number of $X_1, \ldots, X_n$ that fall in $(t_{i-1}, t_i]$.
- Then,

  transformation $\rightarrow$ $$O_1, \ldots, O_m \sim \text{Multinomial}(n, p_1, \ldots, p_m),$$

  where p_i is the "red area" in the graph when Ω_0 is true.

Example 7.17 (GLR tests for goodness-of-fit, TBp.341-342)

- Ω: the vector of cell probabilities $\mathbf{p} = (p_1, \ldots, p_m)$ that are free except for the constraints that $p_i \geq 0$, $\sum_{i=1}^{m} p_i = 1$.

  Assume $m-1 > k$

  – Note that $\dim(\Omega) = m-1$.

- H_0: $\Omega_0 \subset \Omega$,

  Recall. Hardy-Weinberg Equilibrium (LN,CH8,p.24) $P_{AA} = (1-\theta)^2$, $P_{Aa} = 2\theta(1-\theta)$, $P_{aa} = \theta^2$

  $$\Omega_0 = \{\mathbf{p}(\Theta) : \mathbf{p}(\Theta) = (p_1(\Theta), \ldots, p_m(\Theta))\} \text{ e.g.}$$

  is the vector of cell probabilities from $F(\cdot|\Theta)$, where Θ is a k-dimensional unknown parameter, i.e., $p_i(\Theta) = F(t_i|\Theta) - F(t_{i-1}|\Theta)$.

  – Note that $\dim(\Omega_0) = k$.

- Goodness-of-fit tests judge plausibility of the models in H_0 relative to H_A: $\Omega \backslash \Omega_0$ using data $O_1, \ldots, O_m$.
- Likelihood ratio test
  - The LR statistic is

$$\Lambda = \frac{\max_{p\in\Omega_0}\left[\cancel{\frac{n!}{o_1!o_2!\cdots o_m!}}\, p_1(\Theta)^{o_1}\cdots p_m(\Theta)^{o_m}\right]}{\max_{p\in\Omega}\left(\cancel{\frac{n!}{o_1!o_2!\cdots o_m!}}\, p_1^{o_1}\cdots p_m^{o_m}\right)} = \prod_{i=1}^{m}\left(\frac{p_i(\hat\Theta)}{\hat p_i}\right)^{o_i}$$

multinomial joint pmf

where

restricted $\hat{p}_i = p_i(\hat\theta)$

  - $\hat\Theta$ is the MLE of Θ, and — MLE based on $O_1, \cdots, O_m$, rather than $X_1, \cdots, X_n$.
  - $\hat p_1, \hat p_2, \ldots, \hat p_m$ are the unrestricted MLE, i.e.,

$$\hat p_i = o_i/n\,,\ i = 1, 2, \ldots, m.$$

  - Then, since $O_i = n\hat p_i$,

use ratio to compare

$$-2\log\Lambda = 2\sum_{i=1}^{m} n\hat p_i \log\left(\frac{n\hat p_i}{np_i(\hat\Theta)}\right) = 2\sum_{i=1}^{m} O_i \log\left(\frac{O_i}{E_i}\right),$$

where

  - $O_i = n\hat p_i$ is the *observed* counts, and
  - $E_i = np_i(\hat\Theta)$ is the *expected* counts.

cf. the test statistic in Ex.7.16 (LNp.38)

  - H_0 is rejected if $-2\log\Lambda \ge c$.

when H_0 is true

  - To decide the constant c, since

$$\underbrace{\dim(\Omega)}_{m-1} - \dim(\Omega_0) = m - \overbrace{k}^{k} - 1,$$

under H_0, the large sample (i.e., n is large) distribution of $-2\log\Lambda$ is χ^2_{m-k-1}, i.e., $c = \chi^2_{m-k-1}(\alpha)$. ← $(1-\alpha)$-quantile of χ^2_{m-k-1}

Question 7.15

Compare
- the Ω_0 and Ω in item 3, LNp.5, and ← data: $X_1, \cdots, X_n$

  dim $= k$ dim $= \infty$
- the Ω_0 and Ω in Ex.7.17, LNp.39. ← data: $O_1, \cdots, O_m$

  dim $= k$ dim $= m-1$

Are they different? What and Why different?

Example 7.18 (Pearson's Chi-square test, TBp.342-343)

- For the same H_0 vs. H_A and data $O_1, \ldots, O_m$ in Ex.7.17, LNp.39, *Pearson's chi-square test* for goodness of fit:

use difference to compare

  - The test statistic is:

determine what observations "more extreme"

(reasonable?)

$$X^2 = \sum_{i=1}^{m}\frac{(O_i - E_i)^2}{E_i}.$$

Q: Why divided by E_i? Consider the 2 cases
① $O_i = 2$, $E_i = 1$
② $O_i = 999$, $E_i = 1000$
when $n = 5000$

  - H_0 is rejected if $X^2 \ge c$.

when $n \to \infty$

  - Under H_0, $X^2 \xrightarrow{D} \chi^2_{m-k-1}$. ← asymptotic null distribution

- Under H_0, the X^2 statistic is asymptotically equivalent (note: $n \to \infty$) to the LR test statistic given in LNp.40 because:

$X^2 \approx -2\log\Lambda$ under H_0

$$
\begin{aligned}
-2\log\Lambda &= 2n\sum_{i=1}^{m}\underbrace{\hat{p}_i}_{x}\log\left(\frac{\hat{p}_i}{\underbrace{p_i(\hat{\theta})}_{x_0}}\right) \\
&= 2n\sum_{i=1}^{m}\left[\hat{p}_i - p_i(\hat{\theta})\right] + n\sum_{i=1}^{m}\frac{1}{p_i(\hat{\theta})}\left[\hat{p}_i - p_i(\hat{\theta})\right]^2 + \cdots \\
&= 0 + \sum_{i=1}^{m}\frac{(\underbrace{n\,\hat{p}_i}_{=O_i} - n\,p_i(\hat{\theta}))^2}{\underbrace{n\,p_i(\hat{\theta})}_{=E_i}} + \underbrace{\cdots}_{\approx 0} \\
&\approx 0 + X^2
\end{aligned}
$$

$\sum_{i=1}^{m}\hat{p}_i = 1$

$\sum_{i=1}^{m}p_i(\hat{\theta}) = 1$

under H_0

– by Taylor expansion of $x\log(x/x_0)$ about x_0

$$f(x) = x\log(x/x_0) = 0 + (x - x_0) + \frac{1}{2}\frac{1}{x_0}(x - x_0)^2 + \underbrace{\cdots}_{\approx 0 \text{ if } x \approx x_0}$$

– and, the fact that when H_0 is true and n is large, $\hat{p}_i \approx p_i(\hat{\theta})$.

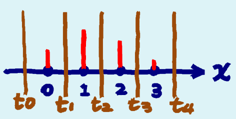
- **Note:** Pearson's chi-square test is more commonly used than the LR test, since it is easier to calculate.

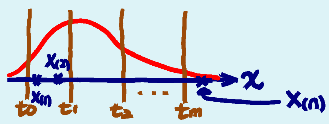
Remarks.

1. There is a distinction between $O_1, \ldots, O_m$ and $X_1, \ldots, X_n$ (especially for continuous case) empirical cdf (textbook, sec. 10.2)
   - $X_1, \ldots, X_n \Rightarrow X_{(1)}, \ldots, X_{(n)} \Rightarrow O_1, \ldots, O_m$
   - $O_1, \ldots, O_m \underset{\not\Rightarrow \text{ continuous}}{\overset{\Rightarrow \text{ discrete (possible)}}{}} X_{(1)}, \ldots, X_{(n)} \underset{\not\Rightarrow \text{ if not i.i.d.}}{\overset{\Rightarrow \text{ if i.i.d.}}{}} X_1, \ldots, X_n$

   If $X_1, \cdots, X_n$ are i.i.d, order statistics $X_{(1)}, \cdots, X_{(n)}$ are sufficient for any distribution

   $X_1, \cdots, X_n$: discrete r.v.'s





   $X_1, \cdots, X_n$: continuous r.v.'s

2. The MLE of θ based on $O_1, \ldots, O_m$ can be different from the MLE of θ based on $X_1, \ldots, X_n$.
3. Different choices of $(t_{i-1}, t_i]$, $i = 1, \ldots, m$, can cause different results. (**Note.** The choice should not depend heavily on observed data.)

   Note In Ex. 7.17, t_i's are not functions of data, i.e., t_i's are not statistics, not r.v.'s. ~~$t_i(X_1, \ldots, X_n)$~~

5/8

4. It is recommended that O_i, $E_i \geq 5$. ← a result guaranteed by large n ← asymptotic property

Example 7.19 (Hardy-Weinberg Equilibrium, TBp.343-344, or Ex.6.15, LN, Ch8, p.24)

- $n = 1029$, the cell probabilities are $\underbrace{(1-\theta)^2}_{AA}$, $\underbrace{2\theta(1-\theta)}_{Aa}$, $\underbrace{\theta^2}_{aa}$ under the Hardy-Weinberg Equilibrium model and the MLE of θ is $\hat{\theta} = 0.4247$.