# STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 19

# Approximate sampling distribution of $\Lambda$

- Consider the general testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

- Suppose the parameter space $\Omega = \Theta_0 \cup \Theta_1$ is a $p$-dimensional parameter space—that is, there are a total of $p$ free parameters.
- Now suppose the null hypothesis $\Theta_0$ is a $q$-dimensional subspace—that is, there are a total of $q$ free parameters in $\Theta_0$.
- Then when the sample size is large, under regularity conditions (those that make Fisher's approximation hold,) we have

$$-2\log\Lambda = 2[\log L(\hat{\theta}) - \log(\hat{\theta}_0)] \sim_{approx} \chi^2_{p-q}.$$

- An approximate level $\alpha$ LR test reject when

$$-2\log\Lambda > F^{-1}_{\chi^2_{p-q}}(1-\alpha)$$

- The approximate $p$-value is

$$p(\mathbf{X}) = 1 - F_{\chi^2_{p-q}}(-2\log\Lambda).$$

# Example: Lot testing

- Consider the lot testing problem with $X_1, X_2, \ldots, X_n$ i.i.d. from Exponential($\lambda$).
- Test
$$H_0 : \lambda = 1.0 \quad \text{vs} \quad H_1 : \lambda \neq 1.0.$$
- What are $p$ and $q$?
- The test statistic (check!)
$$-2\log\Lambda = 2n[\log(1/\bar{X}) - 1 + \bar{X}] \sim_{approx} \chi_1^2.$$
under $H_0$.
- So an approximate level $\alpha = 0.05$ LR test rejects when
$$-2\log\Lambda > F_{\chi_1^2}^{-1}(0.95) \approx 3.84.$$
- One can also compute the approximate p-value
$$p(\mathbf{X}) = 1 - F_{\chi_1^2}(-2\log\Lambda).$$

- For example, if $n = 30$ and $\bar{X} = 0.82$, then

$$-2\log\Lambda = 2\cdot 30\cdot(\log(1/0.82) - 1 + 0.82) = 1.11.$$

  which is smaller than 3.84.
- Thus we do not reject the null at the 5% level.
- In fact, the $p$-value is

$$p(\mathbf{X}) = 1 - F_{\chi^2_1}(1.11) \approx 0.29.$$

  So we would not reject at 10% and even 20% levels. Not much evidence against the null at all!
- Exercise: Repeat the exercise for (i) $n = 30$ and $\bar{X} = 0.42$; (ii) $n = 200$ and $\bar{X} = 0.82$.

# Goodness-of-fit tests

- Next we will look at a particularly important class of applications of the LR test, called the *goodness-of-fit tests*.
- They are for checking whether certain model assumptions fit the data well.
- We will also see how the LR tests in these applications can lead to the so-called *Chi-square tests*.
- Let us begin with some examples.

# Example: Roulette wheel

- A Roulette wheel has 38 slots "1","2",...,"36","0","00".
- The ball will drop into one of the slots on each play. If the wheel is balanced, then the chance of getting each of the slot is $\theta_1 = \theta_2 = \cdots = 1/38$.
- So if we play the game $n$ times, say $n = 1,000$, we will have the counts $X_i$ that the ball has been in the $i$th slot.
- $(X_1, X_2, \ldots, X_{38})$ will again have a Multinomial$(n; \theta_1, \theta_2, \ldots, \theta_{38})$ distribution.
- A hypothesis that one may want to test is whether the wheel is balanced. That is

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = \frac{1}{38}.$$

vs

$$H_1 : \text{Otherwise}.$$

# Roulette wheel

| Roulette wheel outcome | Observed | Expected | Difference |
|---|---|---|---|
| 1 | 30 | 26.3 | 3.7 |
| 2 | 38 | 26.3 | 11.7 |
| 3 | 20 | 26.3 | -6.3 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 35 | 25 | 26.3 | -1.3 |
| 36 | 30 | 26.3 | 3.7 |
| 0 | 15 | 26.3 | -11.3 |
| 00 | 29 | 26.3 | 2.7 |
| Total | 1,000 | 1,000 | 0 |

Is the wheel fair?

# Another example: Weldon's Dice data

- In 1894 biologist Frank Weldon collected the result of 26,306 independent rolls of a set of 12 dice.
- For each roll, his (graduate?) assistants counted the number of dice that show a 5 or 6.
- One question one may ask with these data is whether the dice are fair.

# Weldon's Dice data

| No. of Dice showing 5 or 6 | Observed | Theory | Difference |
|---|---|---|---|
| 0 | 185 | 203 | -18 |
| 1 | 1149 | 1217 | -68 |
| 2 | 3265 | 3345 | -80 |
| 3 | 5475 | 5576 | -101 |
| 4 | 6114 | 6273 | -159 |
| 5 | 5194 | 5018 | 176 |
| 6 | 3067 | 2927 | 140 |
| 7 | 1331 | 1254 | 77 |
| 8 | 403 | 392 | 11 |
| 9 | 105 | 87 | 18 |
| 10 | 14 | 13 | 1 |
| 11 | 4 | 1 | 3 |
| 12 | 0 | 0 | 0 |
| Total | 26,306 | 26,306 | 0 |

What do you think?

- Under the null hypothesis that the dice are all fair, the number of 5's or 6's out of each roll should have a Binomial$(12, \frac{1}{3})$ distribution.
- Let $Y_j$ denote the number of 5 or 6's one gets on the $j$th roll. Then for $k = 0, 1, 2, \ldots, 12$

$$P(Y_j = k) = \binom{12}{k}(1/3)^k \cdot (2/3)^{12-k}.$$

- Now for $k = 0, 1, 2, \ldots 12$, let $X_k$ be the total number of rolls for which $k$ 5 or 6's are observed. That is

$$X_k = \sum_{j=1}^{n} \mathbf{1}(Y_j = k)$$

where $n = 26,306$.

- What is the joint distribution of $(X_0, X_1, X_2, \ldots, X_{12})$?
- It is a Multinomial$(n; \theta_0, \theta_1, \theta_2, \ldots, \theta_{12})$ distribution.
- Testing whether the dice are fair is essentially testing

$$H_0 : \theta_k = \binom{12}{k}(1/3)^k \cdot (2/3)^{12-k} \text{ for } k = 0, 1, 2, \ldots, 12.$$

vs

$$H_1 : \text{Otherwise.}$$

That is, to have at least one $\theta_k$ such that

$$\theta_k \neq \binom{12}{k}(1/3)^k \cdot (2/3)^{12-k}.$$

# Goodness-of-fit test for multinomial data

- Weldon's data and the Roulette wheel fall into the general form of testing a simple null hypothesis on a multinomial distribution against the "complement".
- That is, we want to test

  $$H_0 : \theta_1 = a_1, \theta_2 = a_2, \ldots, \theta_k = a_k$$

  where $a_k$ are known probabilities vs

  $$H_1 : \text{otherwise—that is, at least one } \theta_i \neq a_i.$$

- What test can we use?

# The generalized LR test

- What are the corresponding set of parameter values corresponding to each hypothesis?

$$\Theta_0 = \{(\theta_1, \theta_2, \dots, \theta_k) : \theta_i = a_i \text{ for } i = 1, 2, \dots, k\}$$

and

$$\Theta_1 = \Theta \backslash \Theta_0$$

where

$$\Theta = \left\{ (\theta_1, \theta_2, \dots, \theta_k) : \theta_i \geq 0 \text{ for } i = 1, 2, \dots, k, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

- Reject when

$$\Lambda = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} < K.$$

- The restricted MLE

$$\hat{\theta}_0 = (a_1, a_2, \ldots, a_k).$$

- What is the unrestricted MLE? (This was a homework problem.)

$$\hat{\theta} = \left( \frac{X_1}{n}, \frac{X_2}{n}, \ldots, \frac{X_k}{n} \right)$$

- To see this, recall that

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{X_1!X_2!\cdots X_k!} \theta_1^{X_1} \theta_2^{X_2} \cdots \theta_k^{X_k} \text{ for } \theta_i \geq 0 \text{ and } \sum_{i=1}^{k} \theta_i = 1.$$

  Thus

$$\log L(\theta_1, \ldots, \theta_k) = Constant + \sum_{i=1}^{k} X_i \log \theta_i$$

  and so by noting that $\theta_k = 1 - \theta_1 - \theta_2 - \cdots - \theta_{k-1}$, we have

$$\frac{\partial}{\partial \theta_i} \log L(\theta_1, \theta_2, \ldots, \theta_k) = \frac{X_i}{\theta_i} - \frac{X_k}{\theta_k} = 0.$$

- Back to our LR test,

$$\begin{aligned}
\Lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} &= \frac{L(a_1, a_2, \ldots, a_k)}{L(\frac{X_1}{n}, \frac{X_2}{n}, \ldots, \frac{X_k}{n})} \\
&= \frac{\frac{n!}{X_1!X_2!\cdots X_k!} a_1^{X_1} a_2^{X_2} \cdots a_k^{X_k}}{\frac{n!}{X_1!X_2!\cdots X_k!} (\frac{X_1}{n})^{X_1} (\frac{X_2}{n})^{X_2} \ldots (\frac{X_k}{n})^{X_k}} \\
&= \left(\frac{na_1}{X_1}\right)^{X_1} \left(\frac{na_2}{X_2}\right)^{X_2} \cdots \left(\frac{na_k}{X_k}\right)^{X_k} \\
&= \left(\frac{m_1}{X_1}\right)^{X_1} \left(\frac{m_2}{X_2}\right)^{X_2} \cdots \left(\frac{m_k}{X_k}\right)^{X_k}
\end{aligned}$$

where $m_i = na_i = E(X_i|H_0)$ is the expected value of $X_i$ under the null hypothesis $H_0$.

- So the LR test rejects when

$$\Lambda = \left(\frac{m_1}{X_1}\right)^{X_1} \left(\frac{m_2}{X_2}\right)^{X_2} \cdots \left(\frac{m_k}{X_k}\right)^{X_k} < K.$$

- The test is often carried out in another form—Reject when

$$-2\log\Lambda = -2\sum_{i=1}^{k} X_i \log\left(\frac{m_i}{X_i}\right) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{m_i}\right) > C.$$

- We will cover how to choose the constant $C$ later. What do we need to know for choosing $C$?

# The $\chi^2$ (Chi-squared) test

- Instead of this LR test, Karl Pearson introduced another test for this problem, called the $\chi^2$ *test*.
- More specifically, the $\chi^2$ test rejects when

$$Q = \sum_{i=1}^{k} \frac{(X_i - m_i)^2}{m_i} > C.$$

- This *test statistic* is called the $\chi^2$ *statistic*.
- It is very intuitive. It measures how far the observed values are from the expected values, normalized by the size of the expected value.
- When the expected values under the null distribution are far away from the observed ones, then we reject the null.

# The relation between the $\chi^2$ test and the LR test.

- It turns out that *the $\chi^2$ test statistic* $Q$ is an approximation to the LR test statistic $-2\log\Lambda$.
- How?
- Recall Taylor expansion of a function $f(x)$ about a value $m$

$$f(x) = f(m) + f'(m)(x-m) + \frac{f''(m)}{2}(x-m)^2 + O\left((x-m)^3\right).$$

- Now consider the function $f(x) = x\log\left(\frac{x}{m}\right)$.

$$f(x) = (x-m) + \frac{(x-m)^2}{2m} + O\left((x-m)^3\right).$$

- Therefore, by applying Taylor's expansion aroung $m_i$ to each summand in the LR test statistic $-\log \Lambda$, we have

$$
\begin{aligned}
-2\log \Lambda &= \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{m_i}\right) \\
&= \sum_{i=1}^{k} 2\cdot\left[(X_i - m_i) + \frac{(X_i - m_i)^2}{2m_i} + O\left((X_i - m_i)^3\right)\right] \\
&= 2\cdot\left[\sum_{i=1}^{k} X_i - \sum_{i=1}^{k} m_i + \frac{1}{2}\sum_{i=1}^{k}\frac{(X_i - m_i)^2}{m_i}\right] + \text{Rem} \\
&= 2\left(n - n + \frac{Q}{2}\right) + \text{Rem} \\
&= Q + \text{Rem}.
\end{aligned}
$$

- Under the null hypothesis $H_0$, the remainder term is much smaller relative to $Q$ with high probability.
- Thus under $H_0$,

$$-2\log\Lambda \approx Q.$$

- Rejecting when

$$-2\log\Lambda > C$$

is approximately the same as rejecting when

$$Q > C.$$

# Choosing the constant $C$

- To fully specify the LR test or the $\chi^2$ test, we need to choose the constant $C$.
- Again, we do this by setting the Type I error rate to a given value $\alpha$.
- We need to find the constant $C$ such that (for the $\chi^2$ test),

$$P(Q > C|H_0) = \alpha.$$

- To this end, we need to know the sampling distribution of $Q$ under $H_0$.

# The sampling distribution of $Q$ under $H_0$

- We note that $Q = \sum_{i=1}^{k} \frac{(X_i - m_i)^2}{m_i}$ can only take a finite number of values.
- So one might consider enumerate all possible values of $(X_1, X_2, \ldots, X_k)$ to find the *exact* sampling distribution of $Q$ under $H_0$.
- But the computation will typically be tremendous, and needs to be done separately for each application, as the distribution depends on both $k$ and the actual null hypothesis $(a_1, a_2, \ldots, a_k)$.
- What can we do?

# Approximate sampling distribution of $Q$ (and $-2\log\Lambda$) under $H_0$

- We have learned earlier that the sampling distribution of $-2\log\Lambda$ under $H_0$ can be approximated by a $\chi^2$ distribution.
- The degree of freedom of the $\chi^2$ distribution is $p-q$. What are $p$ and $q$ here?
- It is the $\chi^2$ distribution with $k-1$ degrees of freedom. (Why?)
- Because $\chi^2$ is an approximation of $-2\log\Lambda$, the discrete sampling distribution of the $\chi^2$ statistic is also the same distribution.
- Next let us sketch a proof for the case when $k=2$.

- In this case the data are essentially $X_1 \sim \text{Binomial}(n, \theta)$. Note that $X_1 + X_2 = n$ trials. Alternatively we can write the data as $(X_1, n - X_1)$.
- Similarly, the null hypothesis can be written as $H_0 : \theta = a$.
- Therefore

$$
\begin{aligned}
Q &= \sum_{i=1}^{2} \frac{(X_i - m_i)^2}{m_i} \\
&= \frac{(X_1 - na)^2}{na} + \frac{((n - X_1) - n(1-a))^2}{n(1-a)} \\
&= \frac{(X_1 - na)^2}{na} + \frac{(X_1 - na)^2}{n(1-a)} \\
&= \left( \frac{X_1 - na}{\sqrt{na(1-a)}} \right)^2 .
\end{aligned}
$$

- Applying the Gaussian approximation (CLT) to the Binomial$(n, a)$, we see that $\frac{X_1 - na}{\sqrt{na(1-a)}} \sim_{approx} \text{N}(0, 1)$.

- Accordingly, for the $\chi^2$ test, we may choose a cutoff constant $C$ so that the Type I error rate

$$P(Q > C|H_0) \approx \alpha$$

which gives

$$C = F^{-1}_{\chi^2_{k-1}}(1-\alpha).$$

- Note that this cutoff constant can also be applied to the generalized LR test as

$$P(-2\log\Lambda > C|H_0) \approx P(Q > C|H_0) \approx \alpha.$$

- To sum up, the $\chi^2$ test with level $\alpha$ rejects when

$$Q = \sum_{i=1}^{n} \frac{(X_i - m_i)^2}{m_i} > C,$$

- while the LR test rejects when

$$-2\log\Lambda = 2\sum_{i=1}^{n} X_i \log(X_i/m_i) > C.$$

- What are the corresponding $p$-values?

# The $p$-values for the $\chi^2$ test and the LR test

- For the $\chi^2$ test,

$$\begin{aligned} p(\mathbf{X}) &= \inf\{\alpha : Q > F^{-1}_{\chi^2_{k-1}}(1-\alpha)\} \\ &= \inf\{\alpha : \alpha > 1 - F_{\chi^2_{k-1}}(Q)\} \\ &= 1 - F_{\chi^2_{k-1}}(Q). \end{aligned}$$

- For the LR test,

$$\begin{aligned} p(\mathbf{X}) &= \inf\{\alpha : -2\log\Lambda > F^{-1}_{\chi^2_{k-1}}(1-\alpha)\} \\ &= \inf\{\alpha : \alpha > 1 - F_{\chi^2_{k-1}}(-2\log\Lambda)\} \\ &= 1 - F_{\chi^2_{k-1}}(-2\log\Lambda). \end{aligned}$$

- The two $p$-values are often similar as $Q \approx -2\log\Lambda$.

# Example: Roulette wheel

| Roulette wheel outcome | Obs $(X_i)$ | Exp $(m_i)$ | Diff $(X_i - m_i)$ |
|---|---|---|---|
| 1 | 30 | 26.3 | 3.7 |
| 2 | 38 | 26.3 | 11.7 |
| 3 | 20 | 26.3 | -6.3 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 35 | 25 | 26.3 | -1.3 |
| 36 | 30 | 26.3 | 3.7 |
| 0 | 15 | 26.3 | -11.3 |
| 00 | 29 | 26.3 | 2.7 |
| Total | 1,000 | 1,000 | 0 |

- The LR test statistic

$$
\begin{aligned}
&-2\log\Lambda \\
=&\sum_{i=1}^{38} 2X_i \log(X_i/m_i) \\
=&2\cdot 30\log(30/26.3) + 2\cdot 38\log(38/26.3) + \cdots + 2\cdot 29\log(29/26.3) \\
\approx&51.2.
\end{aligned}
$$

- The $\chi^2$ test statistic

$$
\begin{aligned}
Q &= \sum_{i=1}^{38} (X_i - m_i)^2/m_i \\
&= (30-26.3)^2/26.3 + (38-26.3)^2/26.3 + \cdots + (29-26.3)^2/26.3 \\
&\approx 50.8
\end{aligned}
$$

- Is this significant evidence against the null?
- The p-values are both approximately 6.5%. (How to find the p-values?)

# Example: Weldon's dice data

| No. of Dice showing 5 or 6 | Observed ($X_i$) | Theory ($m_i$) | $X_i - m_i$ |
|---|---|---|---|
| 0 | 185 | 203 | -18 |
| 1 | 1149 | 1217 | -68 |
| 2 | 3265 | 3345 | -80 |
| 3 | 5475 | 5576 | -101 |
| 4 | 6114 | 6273 | -159 |
| 5 | 5194 | 5018 | 176 |
| 6 | 3067 | 2927 | 140 |
| 7 | 1331 | 1254 | 77 |
| 8 | 403 | 392 | 11 |
| 9 | 105 | 87 | 18 |
| 10 | 14 | 13 | 1 |
| 11 | 4 | *1* | 3 |
| 12 | 0 | *0* | 0 |
| Total | 26,306 | 26,306 | 0 |

There are a total of $k = 13$ categories.

# A rule of thumb when carrying out the two tests

- The $\chi^2$ approximation relies on the CLT.
- In this context, the CLT approximation is good when each of the $X_i$ is large enough.
- A rule of thumb is that $m_i > 5$ for all $i$.
- What if some of the $m_i$'s are smaller than 5?
- We *merge them together*, or with another category so that all categories have $m_i > 5$.

# Example: Weldon's dice data

| No. of Dice showing 5 or 6 | Observed ($X_i$) | Theory ($m_i$) | $X_i - m_i$ |
|---|---|---|---|
| 0 | 185 | 203 | -18 |
| 1 | 1149 | 1217 | -68 |
| 2 | 3265 | 3345 | -80 |
| 3 | 5475 | 5576 | -101 |
| 4 | 6114 | 6273 | -159 |
| 5 | 5194 | 5018 | 176 |
| 6 | 3067 | 2927 | 140 |
| 7 | 1331 | 1254 | 77 |
| 8 | 403 | 392 | 11 |
| 9 | 105 | 87 | 18 |
| *10-12* | *18* | *14* | *4* |
| Total | 26,306 | 26,306 | 0 |

After the grouping, we have $k = 11$ categories, and *we carry out the tests with this new $k$*.

- The $\chi^2$ statistic is

$$Q = \frac{(-18)^2}{203} + \frac{(-68)^2}{1217} + \frac{(-80)^2}{3345} + \cdots + \frac{18^2}{87} + \frac{4^2}{14} \approx 35.9.$$

- If we want to test the null at level $1\%$, with $k-1=10$ degrees of freedom, the cutoff constant

$$C = F^{-1}_{\chi^2_{10}}(1-0.01) = 23.2.$$

- So we will reject $H_0$ at the $1\%$ level.
- In fact, the chance that a $\chi^2_{10}$ random variable will exceed 35.9 is about .0000876. (So the $p$-value is .0000876 for the $\chi^2$ test.)
- Hence for any $\alpha$ larger than or equal to .0000876 we will reject $H_0$.
- Exercise: What is the LR test statistic $-2\log\Lambda$ and how does it compare to $Q$?