Chi-square Goodness of Fit

ST551 Lecture 17

Charlotte Wickham

2017-10-30

Finish last time's slides

What about discrete distributions?

The K-S test is only appropriate for continuous distirbutions (the hypothesized distribution is continuous).

But what about if our hypothesis if for a discrete distribution, e.g.:

- Discrete Uniform
- Bernoulli
- Poisson

Setting

Population: $Y \sim$ some **discrete** population distribution with p.m.f $p(y) = P(Y = y)$

Sample: n i.i.d from population, $Y_1, \ldots, Y_n$

Parameter: Whole p.m.f

(kind of like K-S but p.d.f. c.d.f)

Hypotheses: $H_0 : P(Y = y) = p_0(y)$, versus $H_A : P(Y = y) \neq p_0(y)$

defined by Discrete Uniform Bernoulli

Sample estimate of the p.m.f

The discrete sample based estimate of the probability mass function:

$$\hat{p}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i = y\}$$

$\widehat{P(Y=y)}$

The Chi-square Goodness of Fit test

Chi-square goodness of fit compares the estimated p.m.f to the hypothesized one.

Pearson's Chi-square statistic:

$$X(p_0) = \sum_{y} \frac{n\left(\hat{p}(y) - p_0(y)\right)^2}{p_0(y)}$$

all possible values of y

Under null hypothesis: $X(p_0)$ converges in distribution (as n goes to infinity) to χ^2 with $k-1$ degrees of freedom

k = number of possible values for Y.

An alternative presentation

- $j = 1, \ldots, k$ indexes possible values/categories for Y
- O_j be the observed number of values in category j
- $E_j = np_0(j)$ be the expected number of values in category j, based on the hypothesized distribution.

Pearson's Chi-square statistic:

$$X(p_0) = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Example: Dice rolling

I rolled a die 60 times and recorded how many times I got each side: 1, 2, 3, 4, 5, 6

Question: Is the die fair? That is, is $p_0(j) = 1/6$ for $j = 1, 2, 3, 4, 5, 6$?

$j = y$	1	2	3	4	5	6
O_j (Observed data)	20	11	6	7	6	10
E_j (Expected counts)	10	10	10	10	10	10

$$X(p_0) = \frac{(20-10)^2}{10} + \frac{(11-10)^2}{10} + \cdots + \frac{(10-10)^2}{10}$$

$$= 10 + \frac{1}{10} + \cdots + 0$$

$$= 14.2$$

Rejection region

for level α test:

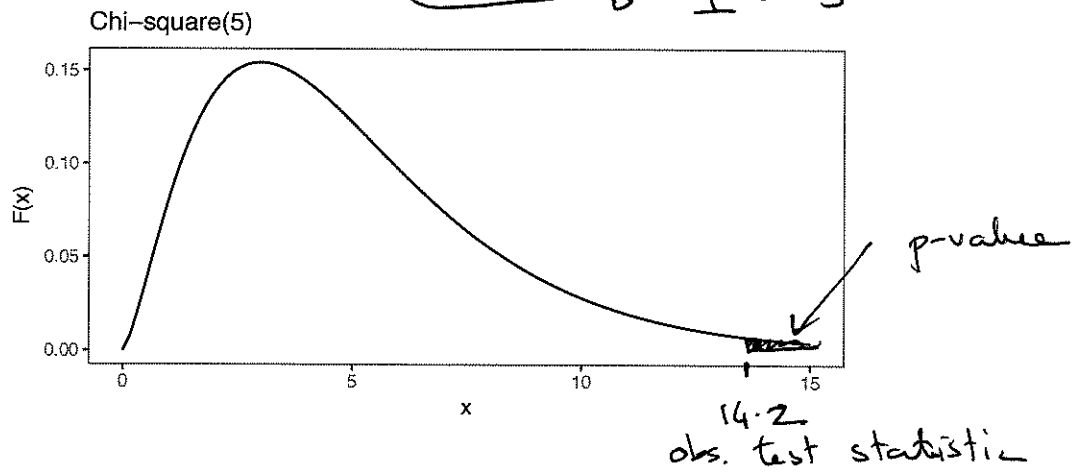
Reject H_0 for $X(p_0) > \chi^2_{(k-1)}(1-\alpha)$

Rejection region always to the right, p-values always area to right.
Why? We usualy are only interested in evidence of poor fit (not evidence of extra good fit)

Two-sided alternative $H_A: p(y) \neq p_0(y)$

Example: Dice rolling cont.

Compare to $\chi^2_{(5)}$ ← 6 − 1 = 5




```
qchisq(0.95, df = 5)
```

$\chi^2_{(5)}(1-\alpha)$ ≈

```
## [1] 11.0705
```

Reject H_0

```
1 - pchisq(chi_sq, df = 5)
```

14.2

```
## [1] 0.01438768
```

← p-value

In R:

```
rolls
```

```
##  [1] 6 5 2 2 6 2 1 1 5 1 2 2 1 1 3 4 5 1 2 4 1 6 6
## [24] 2 4 1 6 6 1 6 1 1 6 1 3 1 3 4 1 1 5 3 4 2 5 1
## [47] 6 3 1 4 3 2 5 1 6 2 1 4 1 2
```

```
chisq.test(table(rolls), p = rep(1/6, 6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(rolls)
## X-squared = 14.2, df = 5, p-value = 0.01439
```

Estimation of parameters

If the null hypothesis doesn't completely specify the distribution p_0, but specifes a family of distributions, $p_0(\theta_1, \theta_2, \ldots, \theta_d)$ where the θ are unknown parameters.

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

You can still use the Chi-square test with some modification

1. Estimate the parameters $\theta_1, \theta_2, \ldots, \theta_d$
2. Find E_j based on estimated parameters and p_0.
3. Compute Pearson's χ^2 statistic as usual
4. Compare statistic to a χ^2 with $k - d - 1$ degrees of freedom, where d is the number of parameters that were estimated.

Example: Poisson

I counted the number of passengers in $n = 40$ vehicles passing through an intersection.

Question: Is the number of passengers per vehicle distributed according to a Poisson distribution?

$k = 7$

$j =$	1	2	3	4	5	6	7
num of passengers Y	0	1	2	3	4	5	6
O_j	6	11	11	8	3	0	1
E_j							

(1) Estimate λ: $\hat{\lambda} = 1.875 = \overline{Y}$

```
mean(passengers)
```

```
## [1] 1.875
```

Example: Poisson

	0	1	2	3	4	5	6	k=8 7+
O_j	6	11	11	8	3	0	1	0
E_j	6.13	11.5	10.78	.	.	.	.	

```
dpois(0:6, lambda = mean(passengers)) * n
```

```
## [1]  6.1341987 11.5016225 10.7827711  6.7392319
## [5]  3.1590150  1.1846306  0.3701971
```

```
(1 - ppois(6, lambda = mean(passengers))) * n
```

```
## [1] 0.02917318
```

← E_8 = expected # cass with 7+ passengers

```
## [1] 25.22917
```

← $\chi^2(p_0)$

Example: Poisson

	0	1	2	3	4	5	6	7+
O_j	6	11	11	8	3	0	1	0
E_j	6.13	11.5	10.8					0.13

```
p_0 <- dpois(0:6, lambda = mean(passengers))
(E <- p_0 * n)
```

```
## [1]  6.1341987 11.5016225 10.7827711  6.7392319
## [5]  3.1590150  1.1846306  0.3701971
```

```
# For 7+ category
(E_7 <- (1 - sum(p_0)) * n)
```

$P(X \geq 7)$ if $X \sim \text{Poisson}(\lambda = \bar{Y})$

```
## [1] 0.1283331
```

Other points

- For binary data, the $X(p_0)$ statistic is equal to the square of the Z-statistic for testing a hypothesis regarding a binary proportion.
  (i.e. approximate Binomial test)

  Therefore, for two sided hypothesis testing $H_0 : p = p_0$ vs $H_A : p \neq p_0$, the χ^2 test and the z-test give the exact same result.

- The $X(p_0)$ statistic has an **asymptotic** (n large) χ^2 distribution.

  Therefore this test is approximate: the test is asymptotically exact.

  The approximation is generally considered appropriate when $E_j > 5$ for all j.

In our example some $E_j < 5$.
Consider collapsing categories

	0	1	2	3	4+

Next time

Starting two sample inference...