Fisher's Exact Test

If the contingency table is 2 x 2, it is also possible to test independence (or equality of proportions) using Fisher's exact test.

(The listing is done exactly as for the chi-square test,) and the test statistic is the value in the (1,1) cell or some equivalent test statistic.

Here the test can be one-sided, unlike for the chi-square test.

Ex: In a game against South Florida in 2010, Corey Fisher made 1 three-point shot and missed 3, while Corey Stokes made 3 and missed 3. Find the p-values for (a) a two-sided test and (b) a one-sided test where the alternative is that Stokes succeeds more often.

Fisher's Exact Test Example (I)

Observed table:

$$\begin{array}{lcc|c} & \text{made} & \text{missed} & \\ \text{Fisher} & 1 & 3 & 4 \\ \text{Stokes} & 3 & 3 & 6 \\ \hline & 4 & 6 & \end{array}$$

Possibilities:

$$\begin{pmatrix} 0 & 4 \\ 4 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ 3 & 3 \end{pmatrix} \quad \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} \quad \begin{pmatrix} 3 & 1 \\ 1 & 5 \end{pmatrix} \quad \begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix}$$

$$\frac{\binom{4}{0}\binom{6}{4}}{\binom{10}{4}} \quad \frac{\binom{4}{1}\binom{6}{3}}{\binom{10}{4}} \quad \frac{\binom{4}{2}\binom{6}{2}}{\binom{10}{4}} \quad \frac{\binom{4}{3}\binom{6}{1}}{\binom{10}{4}} \quad \frac{\binom{4}{4}\binom{6}{0}}{\binom{10}{4}}$$

$$210 \rightarrow \binom{10}{4}$$

$$= \frac{15}{210} \quad = \frac{80}{210} \quad = \frac{90}{210} \quad = \frac{24}{210} \quad = \frac{1}{210}$$

probability of a table as likely or less likely than the observed table

(a) add all but $\frac{90}{210}$ to get $\frac{15+80+24+1}{210} \cong \boxed{0.57}$

(b) add the observed prob. and those to the left to get $\frac{80+15}{210} \cong \boxed{0.45}$

Fisher's Exact Test Example II

Verify our results for this example in R.

Note: The way we computed the two-tailed p-value is valid in general, but in symmetric situations, it reduces to doubling the shorter of the two one-tailed p-values.

More on Regression (I)

Recall that the standard multiple linear regression model is

predictor variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon, \quad \text{where}$$

response

$$\varepsilon \sim N(0, \sigma^2).$$

Assumptions: (1) Linearity, or correct form for the model

(2) Independent error terms ($\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$)

(3) Normally distributed errors

(4) Equal variances (homoscedasticity)

More on Regression II

Kernel regression (from last time) is a way to address violations of the linearity assumption.

Today we'll think about methods for addressing departures from the normal errors assumption.

Two possibilities: ① Theil's method for fitting a line in simple linear regression.

② Robust regression, where outliers (as can occur with non-normal data) are automatically downweighted.

Theil's Method

(I)

<u>Model</u>: $Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, where the x_i values are known and the ε_i values are iid w/ median 0. 不一定要 normal 的假设

<u>Q</u>: How does this differ from the usual model?

<u>A</u>: No assumption of normality!

We can find a distribution-free confidence interval for β (slope parameter) by building on Kendall's test of independence.

Theil's Method (II)

We get the CIs by inverting a test.

Hypotheses: $H_0: \beta = \beta_0$ vs $H_a: \beta \overset{<}{\underset{>}{\neq}} \beta_0$.

Procedure: Compute $D_i = Y_i - \beta_0 X_i$, $i = 1, 2, \dots, n$.

Under H_0, the D_i values are independent of the X_i values. This is since

$$D_i = Y_i - \beta_0 X_i = \underbrace{\alpha + \varepsilon_i}_{\text{iid}}.$$

Now test for evidence against H_0 by testing for an association between X_i and D_i.

Testing Example:

Ex: Test the theory that the slope is $\beta = 5$, using level $\alpha = .06$. Note that using the critical values -0.73 & 0.73 for $\hat{\tau}$ gives a level $-.056$ two-tailed test.

x	y
1	9
2	15
3	19
4	20
10	45
12	55

Soln:

x	y	D = y − 5x
1	9	4
2	15	5
3	19	4
4	20	0
10	45	−5
12	55	−5

concordant pairs is

$1.5 + 0 + 0 + 0 + 0.5$

$= 2$

$\Longrightarrow$

$$\hat{\tau} = 2\left(\frac{2}{\binom{6}{2}}\right) - 1$$

$$= \frac{4}{15} - 1 = -0.73$$

$\Longrightarrow$ Reject H_0!

Theil's Confidence Interval for β

The $100(1-\alpha)\%$ CI contains all values β_0 such that $H_0: \beta = \beta_0$ is not rejected when we do a two-sided level-α test.

Q: How can we compute this interval?

A: 1 Compute all pairwise slopes 2 & put them in order. It turns out that $\hat{\tau}$ is constant on the intervals between these pairwise slopes. Thus, the CI will go from the cth-smallest pairwise slope to the cth-largest pairwise slope for appropriate c.

Small Example

Ex: Verify that the claim about $\hat{\tau}$ is valid for these data:

X	Y
1	9
2	15
3	19

$9 \to 15$: 6; $15 \to 19$: 4; $\frac{10}{2} = 5$

Concordant unless $\beta_0 > 4$.

Soln:

$4 \quad 5 \quad 6 \quad \beta_0$

3 concordant pairs $\Rightarrow$ $\hat{\tau} = 1$

2 concordant pairs $\Rightarrow$ $\hat{\tau} = 1/3$

1 concordant pair $\hat{\tau} = -\frac{1}{3}$

0 concordant pairs $\Rightarrow$ $\hat{\tau} = -1$

assume indp. iid
但是没有个具体 dist family

Bigger Example

Ex: Find a 94% CI for β. Use the fact that Kendall's test rejects at level .056 (two tailed) if the # of concordant or discordant pairs is two or fewer.

X	Y
1	9
2	15
3	19
4	20
10	45
12	55

$\frac{6}{1} = 6$, $\frac{4}{1} = 4$, $\frac{1}{1} = 1$, $\frac{25}{6} \cong 4.17$, $\frac{10}{2} = 5$

$\frac{10}{2} = 5$, $\frac{5}{2} = 2.5$, $\frac{26}{7} \cong 3.71$, $\frac{35}{8} \cong 4.38$

$\frac{11}{3} \cong 3.67$, $\frac{30}{8} \cong 3.75$, $\frac{36}{9} = 4$

$\frac{36}{9} = 4$, $\frac{40}{10} = 4$

$\frac{46}{11} \cong 4.18$

Smallest: 1, 2.5, 3.67

Largest: 6, 5, 5

$\Rightarrow$ $\boxed{\text{The 94.4\% CI is } (3.67, 5.00).}$

Verify in R!

Robust Regression

(I)

First let's think about estimating the center μ of a distribution (univariate) given a random sample $X_1, X_2, \ldots, X_n$.

One estimator: Choose $\hat{\mu}_1$ to be the value that minimizes $L_1(c) = \sum (X_i - c)^2$.

Here $\hat{\mu}_1$ is the sample mean.

Another: Choose $\hat{\mu}_2$ to be the value that minimizes $L_2(c) = \sum |X_i - c|$.

Here $\hat{\mu}_2$ is the sample median (actually any sample median).

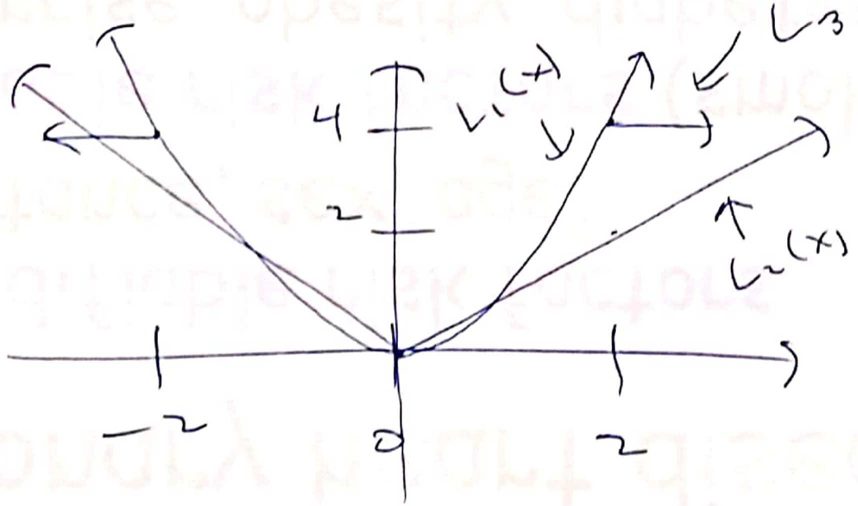
Robust Regression (II)

A third: Choose $\hat{\mu}_3$ to minimize

$$L_3(c) = \sum_i \rho(x_i - c),$$ where

$$\rho(x) = \begin{cases} x^2, & |x| \le d, \\ d^2, & |x| > d. \end{cases}$$

Pictures:



Intuition about L_3: Once $x_i - c$ exceeds 2, there's no further penalty, meaning that L_3 is outlier-resistant.

Applying This To Regression

We estimate the parameters in a regression model by minimizing the criterion

$$\sum_{i=1}^{n} \rho\left(e_i/s\right), \quad \text{where } e_i = \overset{\text{actual}}{Y_i} - \overset{\text{predicted}}{\hat{Y}_i} \text{ is the}$$

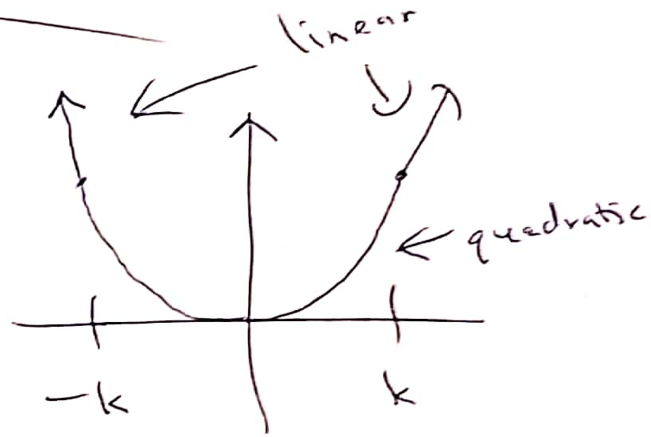
ith residual and s is a robust scale estimate like

$$s = \left.\frac{\text{median}\left|e_i - \text{median}(e_i)\right|}{0.6745}.\right\} \text{approximately unbiased for } \sigma \text{ in a normal model}$$

Examples: Using $\rho(x) = x^2$ gives regular least squares regression. Using $\rho(x) = |x|$ gives L^1-norm regression.

Some ρ Options in R (rlm)

(1) **Huber's**: (the default)



$$\rho_H(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le k, \\ k|e| - \frac{1}{2}k^2, & |e| > k. \end{cases}$$

(2) **Hampel's**: This has quadratic, linear, and <u>constant</u> pieces.

(3) **Bisquare**: Here $\rho_B(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left(1 - \left(\frac{e}{k}\right)^2\right)^3\right\}, & |e| \le k, \\ k^2/6, & |e| > k. \end{cases}$

Now try out some examples in R! Verify that robust regression handles outliers in a reasonable way.

Wald versus Agresti-Coull

Recall that the Wald CI for a proportion performs poorly, while the A-C CI performs well.

Verify this claim in R in two ways.

1. Monte Carlo simulation w/ p close to 0 or 1.

2. Actually computing the coverage probability of each interval for different values of n + p.

For simplicity, let's focus on 95% CIs.

Computing the Coverage Probability.

Suppose that when the # of successes (out of n) is i, the bounds are L_i and U_i.

$\uparrow$ for lower

$\leftarrow$ for upper

The coverage probability for a particular true proportion p is then

$$CP(p) = \sum_{i=0}^{n} \underbrace{\binom{n}{i} p^i (1-p)^{n-i}}_{\text{Prob. of } i \text{ successes}} \cdot \underbrace{I(L_i \le p \le U_i)}_{\text{Indicator for } p \text{ being in the interval for } i \text{ successes}} .$$