Introduction: Distributions and Inference for Categorical Data

Jinhwan Suk

Department of Mathematical Science, KAIST

May 27, 2020

Contents

- Bernoulli trials $\{Y_i\}$: independent and identical

$$P(Y_i = 1) = \pi \quad P(Y_i = 0) = 1 - \pi$$

- The total number of successes, $Y = \sum_{i=1}^{n} Y_i \sim bin(n, \pi)$

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

$$Y \approx \mathcal{N}(n\pi, n\pi(1 - \pi))$$

- Multinomial Distribution :

$$P(n_1, n_2, ..., n_{c-1}) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

Contents

It is used for counts of events that occur randomly over time or space.

- The Poisson pustulates :
  $g(x, w)$: probability of x changes in each interval of length w.
  1. $g(1, h) = \lambda h + o(h)$, λ : positive constant, $h > 0$
  2. $\sum_{x=2}^{\infty} g(x, h) = o(h)$
  3. The number of changes in nonoverlapping intervals are independent
- $g(x, w) = \frac{e^{-\lambda w}(\lambda w)^x}{x!}, \quad x = 0, 1, 2, \dots$
- $Y \approx \mathcal{N}(\mu, \sqrt{\mu}^2)$ as $\mu \to \infty$
- $Y \approx bin(n, \pi)$ if n is large and π is small.

Contents

Overdispersion is the presence of greater variability in a data set than would be expected based on a given statistical model.
($\because$ Poisson distribution is too simple.)

The probability of dying in fatal accident in the next week may vary due to many factors.

$$Y|\mu \sim Poisson(\mu), \quad \mathbb{E}\mu = \theta$$

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|\mu)], \quad Var(Y) = \mathbb{E}(Var(Y|\mu)) + Var(\mathbb{E}(Y|\mu))$$

$$\Rightarrow \mathbb{E}(Y) = \theta, \ Var(Y) = \theta + Var(\mu) > \theta$$

In chapter 4, 12, and 13, the book introduces methods for data that are dispersed.

Contents

Assume that $Y_1, Y_2, ..., Y_c$ are independent Poisson distributions with parameters $\mu_1, ..., \mu_c$. Then, the total $n = Y_1 + \cdots Y_c$ also has Poisson distribution.

If we assume a Poisson model but condition on n, $\{Y_i\}$ no longer have Poisson distribution.

$$
\begin{aligned}
P(Y_1 = n_1, ..., Y_c = n_c | \sum Y_j = n) &= \frac{P(Y_1 = n_1, ..., Y_c = n_c)}{P(\sum Y_j = n)} \\
&= \frac{\prod_i \exp(-\mu_i)\mu_i^{n_i} / n_i!}{\exp(-\sum \mu_j)(\sum \mu_j)^n / n!} \\
&= \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i}
\end{aligned}
$$

Contents

$\hat{\beta} = \arg\max_\beta l(\beta) = \arg\max_\beta L(\beta)$ is called MLE.

- Likelihood equation :

$$\frac{\partial L(\beta)}{\partial \beta} = 0$$

- $cov(\hat{\beta})$ denote the asymptotic covariance of $\hat{\beta}$.

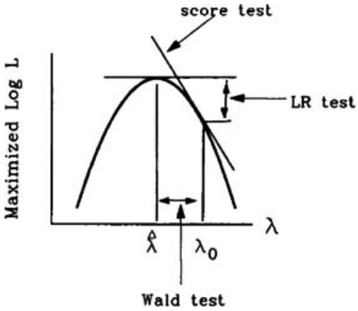
$$cov(\hat{\beta})^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, I_k) \ \ as \ n \to \infty$$

- SE denote the standard error of $\hat{\beta}$.
- Under regularity conditions, MLE is asymptotically efficient, i.e.,

$$\begin{aligned} cov(\hat{\beta}) = \mathcal{I}_n(\beta)^{-1} &= \mathbb{E}\left[\left(\frac{\partial L(\beta)}{\partial \beta}\right)^2\right]^{-1} \\ &= -\mathbb{E}\left(\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \end{aligned}$$

Contents

Situation : $H_0 : \beta = \beta_0$, realization of MLE is $\hat{\beta}$



Reject or Not??

Theorem (Likelihood Ratio Test)

For testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$, suppose $X_1, ..., X_n$ are iid $f(x|\beta)$, $\hat{\beta}$ is the MLE of β, and $f(x|\beta)$ satisfies the regularity conditions. Then under H_0, as $n \to \infty$,

$$-2\log\Lambda(X) \xrightarrow{d} \chi_1^2$$

where $\Lambda(X) = \frac{l(\beta_0)}{l(\hat{\beta})}$.
We would reject H_0 at level α if $-2\log\Lambda(x) > \chi^2_{1,\alpha}$

Lemma (Asymptotic efficiency)

Let $\hat{\beta}$ is the MLE of β, and $\tau(\beta)$ be a continuous function of β. Under the regularity conditions, $\tau(\hat{\beta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Theorem (Likelihood Ratio Test)

Under the same condition,

$$-2\log\Lambda(X) \xrightarrow{d} \chi^2_\nu$$

where $\Lambda(X) = \frac{\sup_{\Theta_0} l(\beta)}{\sup_{\Theta_0\cup\Theta_1} l(\beta)}$ *and* $\nu = dim(\Theta_0 \cup \Theta_1) - dim(\Theta_0)$.

proof on Shao's book Thm 6.5

Use the large-sample normality of MLE, i.e.,

$$z_n = \frac{\hat{\beta} - \beta_0}{SE} \xrightarrow{d} \mathcal{N}(0, I_k)$$

We would reject H_0 if and only if $z_n < -z_{\alpha/2}$ or $z_n > z_{\alpha/2}$

Example(Large sample binomial test)

Let $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(p)$. Consider testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$. The Wald test statistic $z_n =?$

The score function is defined to be

$$u(\beta) = \partial L(\beta)/\partial\beta, \quad \mathbb{E}u(\beta) = 0, \quad Var(u(\beta)) = \mathcal{I}_n(\beta)$$

By Central Limit Theorem,

$$u(\beta)/\sqrt{\mathcal{I}_n(\beta)} \xrightarrow{d} \mathcal{N}(0, I_k)$$

The score statistic $Z_S = u(\beta_0)/\sqrt{\mathcal{I}_n(\beta_0)}$

As $n \to \infty$, the Wald, likelihood-ratio, and score tests have certain asymptotic equivlences.

Contents

ML estimator of π is $\hat{\pi} = y/n$.

- Wald statistic :

$$z_W = \frac{\hat{\pi} - \pi_0}{SE} \approx \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$$

- Score statistic :

$$z_S = \frac{u(\hat{\pi})}{\sqrt{\mathcal{I}_n(\hat{\pi})}} \approx \frac{u(\pi_0)}{\sqrt{\mathcal{I}_n(\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

- Likelihood Ratio :

$$\begin{aligned} -2(L_0 - L_1) &= 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n-y)\log \frac{1-\hat{\pi}}{1-\pi_0}\right) \\ &= 2\left(y \log \frac{y}{n\pi_0} + (n-y)\log \frac{n-y}{n-n\pi_0}\right) \end{aligned}$$

Contents

Wald test statistic gives the interval of π_0 values for which

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

It performs poorly unless n is very large.

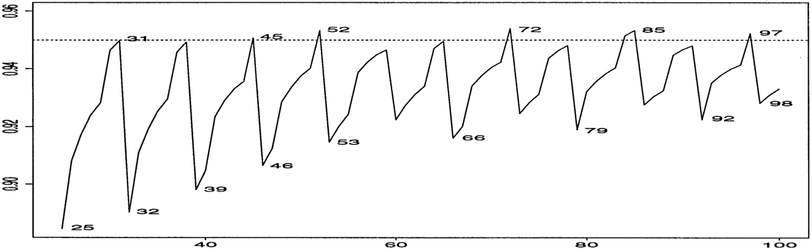


FIG. 1. *Standard interval; oscillation phenomenon for fixed* $p = 0.2$ *and variable* $n = 25$ *to* 100.

Statistical Inference for Binomial Parameters

Brown et al. 2001

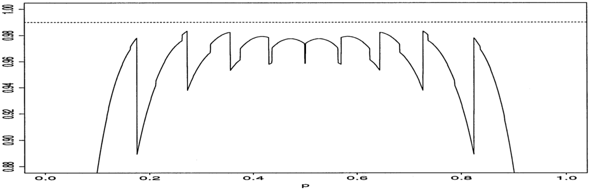


FIG. 4. *Coverage of the nominal* 99% *standard interval for fixed* $n = 20$ *and variable* p.

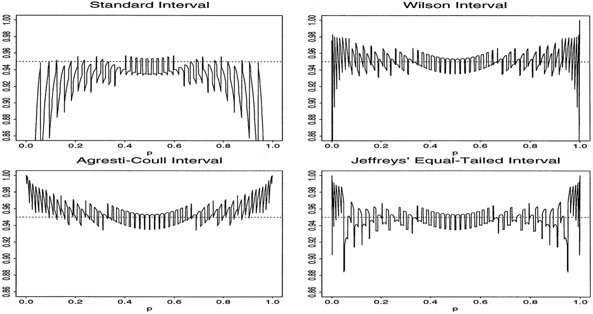


FIG. 5. *Coverage probability for* $n = 50$.

Score confidence interval : $|z_S| < z_{\alpha/2}$
Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}$$

$$\left(\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right)^2 = z^2_{\alpha/2}$$

$$\Leftrightarrow (n + z^2_{\alpha/2})\pi_0^2 - (2n\hat{\pi} + z^2_{\alpha/2})\pi_0 + n\hat{\pi}^2 = 0$$

$$\Leftrightarrow \pi_0 = \frac{(2n\hat{\pi} + z^2_{\alpha/2}) \pm \sqrt{(2n\hat{\pi} + z^2_{\alpha/2})^2 - 4(n + z^2_{\alpha/2})n\hat{\pi}^2}}{2(n + z^2_{\alpha/2})}$$

$$\Leftrightarrow \pi_0 = \hat{\pi}\left(\frac{n}{n+z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)$$
$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\frac{n}{n+z_{\alpha/2}^2} + \frac{1}{2}\frac{1}{2}\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right]}$$

The midpoint $\tilde{\pi}$ is a weighted average of $\hat{\pi}$ and $1/2$. And, the square of the coefficient of $z_{\alpha/2}$ is a weighted average of the variance.

Contents

One question asked each student whether he or she was a vegetarian. Of $n = 25$ students, $y = 0$ answered "yes"

The 95% interval for π is

- Wald : $0 \pm 1.96\sqrt{0.0 \times 1.0/25} = (0, 0)$
- LRT : $-2 \log \Lambda(x) = -50 \log(1 - \pi_0) \leq \chi^2_{1,0.05} = (0, 0.074)$
- Score : $(0.0, 0.133)$, more believable inference

For $H_0 : \pi = 0.5$, the score test statistic is

$$z_S = (0 - 0.5)/\sqrt{0.5 \times 0.5/25} = -5.0$$

For $H_0 : \pi = 0.1$, the score test statistic is

$$z_S = (0 - 0.1)/\sqrt{0.1 \times 0.9/25} = -1.67$$

If n is **small**, it would be better to use binomial distribution directly rather than its normal approximation.

Clopper-Pearson interval : Assume that θ is a parameter that we are interested in. From observed data, how can we set confidence interval for θ??

1. Set θ_0 as null hypothesis.
2. P-values exceed α in that null hypothesis?
3. If they do, θ_0 may be chosen properly.

The endpoints are the solutions in π_0 to the equations

$$\sum_{k=y}^{n} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \frac{\alpha}{2} \qquad \sum_{k=0}^{y} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \frac{\alpha}{2}$$

$$\Rightarrow \left(\left[1 + \frac{n-y+1}{yF_{2y,2(n-y+1)}(1-\alpha/2)} \right]^{-1}, \left[1 + \frac{n-y}{(y+1)F_{2(y+1),2(n-y)}(\alpha/2)} \right]^{-1} \right)$$

Contents

$H_0 : \pi_j = \pi_{j0},\ j = 1, ..., c$ where $\sum_j \pi_{j0} = 1$

- $\mu_j := \mathbb{E} n_j = n\pi_{j0}$, *expected frequencies*.
- Pearson statistic : $X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}$
- Observed value : $X_o^2 \Rightarrow$ P-value : $P_{H_0}(X^2 \geq X_o^2)$
- If n is large enough, $X^2 \approx \chi^2_{c-1}$

Example : Testing Mendel's Theory

- Mendel's Prediction : 75% yellow, 25% green
- Experiment result : $n_1 = 6022,\ n_2 = 2001,\ n = 8023$
- $X^2 = 0.015$, P-value : 0.90, not contradict null hypothesis.

- Mendel performed several experiments of this type. In 1936, R.A. Fisher summarized Mendel's results.
- $X_1^2, ..., X_k^2$: Results of each experiment performed independently, with degree of freedom $\nu_1, ..., \nu_k$
- $\sum_j X_j^2$ has a chi-squared distribution with df $= \sum_j \nu_j$.

Fisher obtained $P = 0.99996$, too good!

Goodness-of-fit tests can reveal not only when a fit is *inadequate*, but also when it is **better than random** fluctuations would have us expect.

- ML estimate for $\pi_0 = (\pi_{10}, \ldots, \pi_{c0})$ is $(n_1/n, \ldots, n_{c-1}/n)$
- Denote i-th observation by $Y_i = (Y_{i1}, \ldots, Y_{ic}), \; i = 1, \ldots, n.$

$$\sum_j Y_{ij} = 1 \quad Y_{ij} Y_{ik} = 0 \;\; \forall j \neq k$$

- $\mathbb{E} Y_{ij} = P(Y_{ij} = 1) = \pi_{j0} = \mathbb{E} Y_{ij}^2$
- $\mathbb{E} Y_i = \pi$ and $cov(Y_i) = \Sigma = (\sigma_{jk})$ where

$$\sigma_{jj} = Var(Y_{ij}) = \pi_{j0}(1 - \pi_{j0})$$

$$\sigma_{jk} = cov(Y_{ij}, Y_{ik}) = \mathbb{E} Y_{ij} Y_{ik} - \mathbb{E} Y_{ij} \mathbb{E} Y_{ik} = \pi_{j0}\pi_{k0}$$

$$\Sigma = diag(\pi_0) - \pi_0' \pi_0$$

- Define $\hat{\pi} := \bar{Y} = \frac{1}{n}\sum_i Y_i$, by CLT,

$$\sqrt{n}(\hat{\pi} - \pi_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

- Multiply $A \in \mathbb{R}^{c \times c-1}$ defined by $A_{ii} = 1, A_{ij} = 0$

$$\sqrt{n}(\hat{\pi} - \pi_0)A \xrightarrow{d} \mathcal{N}(0, A'\Sigma A)$$

- $\hat{\pi} = (n_1/n, \ldots, n_{c-1}/n)$, $\pi_0 = (\pi_{10}, \ldots, \pi_{c-1,0})$

$$n(\hat{\pi} - \pi_0)'\Sigma_0^{-1}(\hat{\pi} - \pi_0) \xrightarrow{d} \chi^2_{c-1}$$

- Σ_0^{-1} has (j,k)th element $\frac{1}{\pi_{c0}}$ when $j \neq k$ and $(\frac{1}{\pi_{j0}} + \frac{1}{\pi_{c0}})$ when $j = k$
- $n(\hat{\pi} - \pi_0)'\Sigma_0^{-1}(\hat{\pi} - \pi_0) = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j} =: X^2$

Contents

Recall that under dome regularity condition and under H_0, the ratio of the likelihood is

$$\Lambda(X) = \frac{\sup_{\Theta_0} l(\beta)}{\sup_{\Theta_0 \cup \Theta_1} l(\beta)} \leq 1 \text{ and } -2\log\Lambda(X) \xrightarrow{d} \chi^2_\nu$$

The likelihood ratio statistic G^2 is

$$G^2 = -2\log\frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}} = 2\sum_j \log(n_j/n\pi_{j0})$$

The larger the value of G^2, the greater the evidence against H_0.

When H_0 holds, the Pearson X^2 and the likelihood ratio G^2 both have χ^2_{c-1} distribution.

$$X^2 - G^2 \xrightarrow{P} 0$$

When H_0 is false, they need not take similar values.

Contents

Pearson's X^2 compares a sample distribution to a hypothetical one $\{\pi_{j0}\}$.

$$X^2 = \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j}$$

In some cases, $\pi_{j0} = \pi_{j0}(\theta),\ \theta \in \Theta \subset \mathbb{R}^p$

$\rightarrow$ cannot get $X^2 \Rightarrow$ replace μ_j by $\hat{\mu}_j$

The distribution of X^2 is a chi-squared distribution with $df = (c-1) - p$.

TABLE 1.1 Primary and Secondary Pneumonia Infections in Calves

	Secondary Infection[a]	
Primary Infection	Yes	No
Yes	30 (38.1)	63 (39.0)
No	0 (—)	63 (78.9)

$$H_0 : P(1st\ \text{infection}) = P(2nd\ \text{infection}|1st\ \text{infection})$$
$$\Leftrightarrow H_0 : \pi_{11} + \pi_{12} = \pi_{11}/(\pi_{11} + \pi_{12})$$

Under H_0, $\pi := \pi_{11} + \pi_{12}$, $\pi_{11} = \pi^2$, $\pi_{12} = \pi - \pi^2$, and $\pi_{22} = 1 - \pi$

$\hat{\pi} = \arg\max l(\pi) = (\pi^2)^{n_{11}}(\pi - \pi^2)^{n_{12}}(1 - \pi)^{n_{22}}$

$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22}) = 0.494 \Rightarrow X^2 = 19.7$ (df=1)

$P = 0.00001$: Strong evidence against H_0