Basic Discrete Distribution and Inference for Categorical Data

CF Jeff Lin, MD., PhD.

February 17, 2006

Common Discrete Distributions

1. Binomial
2. Poisson
3. Multinomial

Bernoulli Distribution

1. A random variable Y has a Bernoulli distribution if

$$X = \begin{cases} 1, & \text{with probability } \pi; \\ 0, & \text{with probability } 1-\pi. \end{cases} \tag{1}$$

2. That is

$$f(y;\pi) = \begin{cases} \pi^y (1-\pi)^{1-y}, & \text{for } x = 0, 1; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

$$\mathcal{E}(Y) = \pi \tag{3}$$

$$\mathbf{Var}(Y) = \pi(1-\pi) \tag{4}$$

Binomial Distribution

1. The binomial distribution, one of the more useful distributions, is based on the idea of a Bernoulli.

2. Many application refer to a fixed number n observations.

3. Let $y_1, y_2, \ldots, y_n$ denote responses has the for n **repeated independent** and **identical trials** such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$.

4. We use the generic labels "success" and "failure" for outcomes 1 and 0.

Binomial Distribution

5. **Identical trials** means that the probability of success π is the same for each trial. **independent trials** means that Y_is are independent random variables.

6. These are often called **Bernoulli trials**.

7. A Bernoulli trial is an experiment with two and only two, possible outcomes.

Binomial Distribution

8. The total number successes, $Y = \sum_i^n Y_i$ has the **binomial distribution** as

$$\begin{aligned} P[Y = y] &= P[\text{exactly } y \text{ successes in the } n \text{ trials}] \\ &= \binom{n}{y} \pi^y (1-\pi)^{n-y}. \end{aligned} \tag{5}$$

which index n and parameter π, denote by $Bin(n, \pi)$.

Binomial Distribution

9. Formally, a random variable Y is defined to have binomial distribution if the discrete density function of Y is given by

$$f(y;n,\pi) = \begin{cases} \binom{n}{y}\pi^y(1-\pi)^{n-y}, & \text{for } y = 0,1,2,\cdots,n; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathcal{E}(Y) = n\pi \tag{7}$$

$$\mathbf{Var}(Y) = n\pi(1-\pi) \tag{8}$$

where the **parameter** n, π satisfies $0 \leq \pi \leq 1$, π is often denoted by p and $1-\pi$ is often denoted by q.

Example: Dice probabilities

1. Suppose we are interested in finding the probability of obtaining at least one 6 in four rolls of a fair dice.

2. This experiment can be modeled as a sequence of four Bernoulli trials with success probability $p = \frac{1}{6} = P(\text{die shows 6})$.

3. Define the random variable X by

   $X =$ total number of6s in four rolls.

4. Then $X \sim Bin(n = 4, p = \frac{1}{6})$ and

$$P(\text{at least one 6}) \;=\; P(X > 0) = 1 - P(X = 0) = 1 - \left(\frac{5}{6}\right)^4 = 0.518$$

Binomial Distribution

5. The skewness is described by
   $\mathcal{E}[(Y-\mu)^3/\sigma^3] = (1-2\pi)/\sqrt{n\pi(1-\pi)}$.

6. The distribution converges to normality as n increases, for fixed π.

Binomial Distribution

7. There is no guarantee that successive binary observations are independent or identical.

8. Thus, occasionally, we will utilize other distributions.

9. For example, we sample binary outcomes without replacement from a finite population, we use **hypergeometric distribution**.

Poisson Distribution

1. A random variable Y is defined to have **Poisson distribution** if the discrete density function of Y is given by

$$f(y;\mu) = \begin{cases} \frac{e^{-\mu}\mu^y}{y!} & \text{for } y = 0, 1, 2, \cdots, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

$$\mathcal{E}(Y) = \mu \tag{10}$$

$$\mathbf{Var}(Y) = \mu \tag{11}$$

It satisfies $\mathcal{E}(Y) = \mathbf{Var}(Y) = \mu$.

Poisson Distribution

2. It is unimodal with mode equal to the integer part of μ.

3. Its skewness is described by $\mathcal{E}(Y-\mu)/\sigma^3 = 1/\sqrt{\mu}$.

4. The distribution approaches normality as μ increases.

5. The Poisson distribution has a single parameter μ, sometimes called the **intensity parameter**.

Poisson Distribution

6. The Poisson distribution provides a realistic model for many random phenomena.

7. Since the values of a Poisson random variable are the nonnegative integers, any random phenomenon for which a count of some sort is of interest is a candidate for modeling by assuming a Poisson distribution.

8. This distribution is usually associated with rare events, counts of events that occur randomly over time or space, which outcomes in disjoint periods are independent, such as a count might be the number of a fatal traffic accidents per week.

Poisson Distribution

9. For example, if we are modeling a phenomenon in which we are waiting for an occurrence, the number of occurrences in a given time intervals can sometimes be modeled by the Poisson distribution.

10. On the basic assumptions on which Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time.

11. It makes sense to assume that the longer we wait, the more likely occurrence will enter.

Example: Mercy Hospital

1. Patients arrive at the emergency room of Mercy Hospital at the average rate of 6 per hour on weekend evenings.

2. What is the probability of 4 arrivals in 30 minutes on a weekend evening?

$$\begin{aligned} \mu &= \frac{6}{\text{hour}} = \frac{3}{\text{half-hour}} \\ y &= 4 \\ f(y=4) &= \frac{3^4 e^{-3}}{4!} = 0.1680 \end{aligned}$$

Poisson Sampling

Poisson Sampling Example: Sprots Injury

1. An investigator plan to study the sports injury rate of college students in a department of physical education in one academic year.

2. The data will consist of monthly counts of the number of sports injuries.

3. The result is shown in Table 1.

4. **Poisson distribution** is a potential probability model for the number of sports injuries in any given month.

Poisson Sampling Example: Sprots Injury

Table 1: Counts of sports injury in one year

Month	1	2	3	4	5	6	7	8	9	10	11	12
Count	28	56	51	28	6	5	19	13	13	10	7	11

Poisson Sampling

Poission Variables for Categorical Data

Table 2: Poission Variables for Categorical Data

Level of a Discrete Variable	1	2	$\cdots$	i	$\cdots$	C
Observed Count	y_1	y_2	$\cdots$	y_i	$\cdots$	y_C

Poisson Sampling

1. Observed counts $y_i, i = 1, \ldots, C$, in the C cells of a contingency table.

2. For instance, these might be observations for the C levels of a single categorical variable, or for $C = IJ$ cells of a two-way table.

Poisson Sampling

1. The Poisson sampling model for count, y_i, assumes that they are independent Poisson random variables.

2. The joint probability function for y_i is then the product of the probabilities for the C cells.

$$f(y_i; \mu_i, i = 1, \ldots, C) = \prod_1^C \frac{e^{-\mu_i} \mu^{y_i}}{y_i!} \tag{12}$$

3. The total sample size $n = \sum y_i$ also has a Poisson distribution, with parameter $\sum \mu_i$, where μ_i are called **expected frequencies**.

Poisson Sampling Distrobution

1. An important **sampling distribution** for categorical data treats count of each level as an as an independent Poisson observation.

2. The sampling scheme is called **Poisson sampling**.

3. A key feature of the Poisson distribution is that its variance increases as the mean does.

4. Sample count tend to vary more when their average level is higher.

Overdispersion of Poisson Sampling Distrobution

1. In practice, count observations often have variance exceeding the mean, rather than equaling he mean as the Poisson requires.

2. This phenomenon is called **overdispersion**.

3. The assumption of Poisson sampling is often too simplistic, because of factors such as overdispersion.

4. Nevertheless, Poisson sampling assumption produce useful results, albeit in a approximate manner, in a wide variety of categorical data analysis.

Multinomial Distribution

1. A random variable $\underline{\boldsymbol{Y}} = (Y_1, Y_2, \ldots, Y_C)^T$ has multinomial distribution, for some integers $C \geq 2, n \geq 1$ and some $0 \leq \pi \leq 1$ such that $\pi_1 + \pi_2 + \cdots + \pi_C = 1$ and $1 \leq i \leq C; y_1 + y_2 + \cdots + y_C = n$, as

$$P(Y_1, Y_2, \ldots, Y_C) = \begin{cases} \frac{n!}{y_1! y_2! \cdots y_C!} \pi^{y_1} \pi^{y_2} \cdots \pi_c^{y_C}, & \text{for } y_i = 0, 1, 2, \cdots, n, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

$$\mathcal{E}(Y_j) = n\pi_j, \tag{14}$$

$$\mathbf{Var}(Y_j) = n\pi_j(1 - \pi_j), \tag{15}$$

$$\mathbf{Cov}(Y_j, Y_k) = -n\pi_j\pi_k. \tag{16}$$

Multinomial Distribution

2. Some trials have more than two possible outcomes.

3. Suppose that each of n independent, identical trials can have outcome in any of C categories (The discrete response variable has distinct C levels).

4. Let $Y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise.

5. Then $\underline{\boldsymbol{Y}}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iC})^T$ represents a multinomial trial, with $\sum_j Y_{ij} = 1$; for instance, $(0, 0, 1, 0)$ denotes outcomes in category 3 of four possible categories.

6. Note that Y_{iC} is redundant, being linearly dependent on the others.

Multinomial Distribution

7. Let $n_j = \sum_i Y_{ij}$ denote the umber of trials having outcome in category j.

8. The counts $(n_1, n_2, \ldots, n_C)^T$ have **multinomial distribution**.

9. Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in categories j for each trial.

Multinomial Distribution

10. The multinomial probability mass function is

$$P(n_1, n_2, \ldots, n_C) = \frac{n!}{n_1! n_2! \cdots n_C!} \pi^{n_1} \pi^{n_2} \cdots \pi_C^{n_C}. \tag{17}$$

11. Since $\sum_j n_j = n$, this is $(C-1)$-dimensional, with $n_C = n - (n_1 + n_2 + \cdots + n_{C-1})$.

12. The binomial distribution is the special case with $C = 2$ and the **trinomial distribution** is another special case with $C = 3$.

Overdispersion

1. In practice, count observations often exhibit variability exceeding the predicted by the binomial or Poisson.

2. This phenomenon is called **overdispersion**.

3. We assumed above that each person has the same probability of dying in a fatal accident in the next week.

4. More realistically, these probabilities vary, due to factors such as amount of time spent driving, whether the person wears a seat belt, and geographical location.

5. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Overdispersion

6. Suppose that Y is random variable with variance $\mathbf{Var}(Y \mid \mu)$ for given μ but μ itself varies because of unmeasured factors such as those just described.

7. Let $\theta = \mathcal{E}(\mu)$.

Overdispersion

8. Then unconditionally,

$$\mathcal{E}(Y) = \mathcal{E}[\mathcal{E}(Y \mid \mu)], \tag{18}$$

$$\mathbf{Var}(Y) = \mathcal{E}[\mathbf{Var}(Y \mid \mu)] + \mathbf{Var}[\mathcal{E}(Y \mid \mu)]. \tag{19}$$

9. When Y is conditional Poisson (give μ), for instance, then

$$\mathcal{E}(Y) = \mathcal{E}(\mu) + \theta, \tag{20}$$

$$\text{and} \quad \mathbf{Var}(Y) = \mathcal{E}(\mu) + \mathbf{Var}(\mu) = \theta + \mathbf{Var}(\mu) > \theta. \tag{21}$$

Overdispersion

10. Assuming a Poisson distribution for a count variable is often too simplistic, because of factors has the cause overdispersion.

11. The **negative binomial** is a related distribution for count data that permit the variance to exceed the mean.

Overdispersion

12. Analyzes assuming binomial (or multinomial) distributions are also sometimes invalid because of overdispersion.

13. This might happen because the true distribution is a mixture of different binomial distributions, with the parameter varying because of unmeasured variables.

Overdispersion

14. To illustrate, suppose that an experiment exposes pregnant mice to a toxin and then after a week observes the number of fetuses in each mouse's litter that show signs of malformation.

15. Let n_i denote the number of fetuses in the litter for mouse i.

16. The mice also vary according to other factors that may not be measured, such as their weight, overall health, and genetic makeup.

17. Extra variation then occurs because of the variability from liter to liter in the probability π of malformations might cluster near 0 and near n_i, showing more dispersion than expected for binomial sampling with a single value of π.

Overdispersion

18. Overdispersion could also occur when π varies among featuses in a litter according to some distribution.

19. Use **beta-binomial distribution** to adjust overdispersion for binomial distribution.

20. Use **negative binomial distribution** to adjust overdispersion for Poisson distribution.

Connection Between Poisson and Multinomial Distributions

1. Suppose $Y_1, Y_2, \ldots, Y_C$ are independent Poisson distribution with mean μ_i.

2. The joint probability mass function for $\underline{\boldsymbol{Y}} = (Y_1, Y_2, \ldots, Y_C)^T$ is the product of the n mass functions of Poisson distribution.

3. The total $n = X = \sum_i Y_i$ also has a Poisson distribution (9), with parameter $\sum \mu_i$. Such as,

$$f(X = \sum_i y_i = n; \mu_i) = \begin{cases} \frac{e^{-\sum_i \mu_i} (\sum_i \mu_i)^n}{n!}, & \text{for } X = \sum_i Y_i = n, \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

Connection Between Poisson and Multinomial Distributions

4. An unusual feature of Poisson sampling is that the total sample size $n = \sum Y_i$ is random, rather than fixed.

5. If we start with the Poisson model but condition on the total sample size n, n no longer have Poisson distribution, since each Y_i cannot exceed n, condition on n, Y_i are no longer independent, since the value of one affects the possible range for the others.

Connection Between Poisson and Multinomial Distributions

6. Given that $n = \sum Y_i$, the conditional probability of a set Y_i satisfying this condition is **multinomial distribution** as

Connection Between Poisson and Multinomial Distributions

$$P\left[Y_i = y_i, i = 1, \ldots, c \mid n = \sum y_i \right]$$

$$= \frac{P[\, y_i, i = 1, \ldots, c \,]}{P[\, n = \sum y_i \,]}$$

$$= \frac{\prod_i \left[\exp(-\mu_i)\ \mu_i^{y_i} / y_i! \right]}{\exp\left(\sum_i -\mu_i\right) \left(\sum_i \mu_i\right)^n / n!} \tag{23}$$

$$= \left(\frac{n!}{\prod_i y_i!} \right) \prod_i \pi_i^{y_i}, \quad \text{where } \pi_i = \frac{\mu_i}{\sum_i \mu_i}. \tag{24}$$

Connection Between Poisson and Multinomial Distributions

7. We denote the vector, $\underline{\mathbf{y}} = (y_1, y_2, \ldots, y_c)^T$, and mean vector, $\underline{\boldsymbol{\pi}} = (\pi_i, \ldots, \pi_c)^T$ with the multinomial distribution as

$$\underline{\mathbf{y}} \sim Multin(n, \underline{\boldsymbol{\pi}}) \tag{25}$$

characterized by the sample size n and the cell probability π_i.

Connection Between Poisson and Multinomial Distributions

8. Many categorical data analyses assume a multinomial distribution.

9. Such analyses usually have the same parameter estimates as those of analyses assumption Poisson distribution, because in the likelihood functions.

10. Note: this statement will be more clear when we discuss the sampling distributions.

Basic Statistical Inference for Categorical Data

Basic Statistical Inference
Likelihood Functions and
Maximum Likelihood Estimation

Likelihood Functions and Maximum Likelihood Estimation

1. We often use **maximum likelihood** for parameter estimation in categorical data analysis.

Likelihood Functions and Maximum Likelihood Estimation

2. Under weak regularity conditions, such as the parameter space having fixed dimension with true value falling in its interior, maximum likelihood estimators have desirable properties:

   (a) They have large-sample normal distributions.
   (b) They are asymptotically consistent.
   (c) They Converge to the parameter as n increases.
   (d) They are asymptotically efficient, producing large-sample standard errors no greater than those form other estimation methods.

Likelihood Functions and Maximum Likelihood Estimation

3. The method of maximum likelihood (ML) was developed by R. A. Fisher (1922, 1925) and largely replaced more ad hoc method (such as least squares and method of moments) as the standard estimation method.

Likelihood Functions and Maximum Likelihood Estimation

4. The strength of ML is its inherent logic, its extremely widely scope, and its high efficiency under wide conditions.

5. The main weakness of ML is that the entire distribution of the data must be modeled.

6. This involves extra assumptions whose failure might have adverse effects on estimation precision.

Likelihood Functions and Maximum Likelihood Estimation

7. Given the data, for a chosen probability distribution the **likelihood function** is the probability of those data, treated as a function of the unknown parameter.

8. The maximum likelihood (ML) estimate is the parameter value that maximizes this function.

9. This is the parameter value under which the data observed have the highest probability of occurrence.

10. The parameter value that maximizes the likelihood function also maximizes the log of that function.

Likelihood Functions and Maximum Likelihood Estimation

11. It is simpler to maximize the log likelihood since it is a sum rather than a product of terms.

12. we denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$.

Likelihood Functions and Maximum Likelihood Estimation

13. The likelihood function is $L(\beta)$ and the log-likelihood function is $\ell(\beta) = \log[\mathbf{L}(\beta)]$.

14. For many models, $\ell(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0.

15. The ML estimate is then the solution of the likelihood equation, $\partial\ell(\beta)/\partial\beta$.

16. Often, β is multidimensional, denoted by $\underline{\boldsymbol{\beta}}$, and $\underline{\hat{\boldsymbol{\beta}}}$ is the solution of a set of likelihood equations.

Likelihood Functions and Maximum Likelihood Estimation

17. Let $SE(\hat{\beta})$ denote the standard error of $\hat{\beta}$, and let $\mathbf{Cov}(\underline{\hat{\boldsymbol{\beta}}})$ denote the asymptotic covariance matrix of $\underline{\hat{\boldsymbol{\beta}}}$. Under regularity conditions (Rao, 1973, p.364) is the inverse of the **Fisher's information matrix**.

18. The (j,k) element of the Fisher's information matrix is

$$-\mathcal{E}\left(\frac{\partial^2 \ell(\underline{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_k}\right). \tag{26}$$

19. The standard errors (SE) are the square roots of the diagonal elements for the inverse information matrix.

Likelihood Functions and Maximum Likelihood Estimation

20. The greater the curvature of the log likelihood, the smaller the standard errors.

21. This is the reasonable, since large curvature implies that the log likelihood drops quickly as $\underline{\beta}$ moves away from $\underline{\hat{\beta}}$; hence, the data would have been much more likely to occur if $\underline{\beta}$ took a value near $\underline{\hat{\beta}}$ rather than a value far from $\underline{\hat{\beta}}$.

Likelihood Function and ML Estimate for Binomial Parameter

Likelihood Function and ML Estimate for Binomial Parameter

1. The part of a likelihood function involving the parameter is called he **kernel**.

2. Since he maximization of the likelihood is with respect to the parameters, the rest is irrelevant. Consider the binomial distribution.

3. The binomial coefficient $\binom{n}{y}$ has no influence on where the maximum occurs with respect to π.

4. Thus, we ignore it and treat the kernel as the likelihood function.

Likelihood Function and ML Estimate for Binomial Parameter

5. The binomial log likelihood is then

$$\ell(\pi) = \log[\pi^y(1-\pi)^{n-y}] = y\log(\pi) + (n-y)\log(1-\pi). \quad (27)$$

6. Differentiating with respect to π yields

$$\frac{\partial \ell(\pi)}{\partial \pi} = \frac{y}{\pi} - \frac{(n-y)}{(1-\pi)} = \frac{y - n\pi}{\pi(1-\pi)} \quad (28)$$

7. Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Likelihood Function and ML Estimate for Binomial Parameter

8. Calculating $\partial^2 \ell(\pi)/\partial \pi^2$, taking the expectation, and combining terms, we get

$$-\mathcal{E}\left[\frac{\partial^2 \ell(\pi)}{\partial \pi^2}\right] = \mathcal{E}\left[\frac{y}{\pi^2} + \frac{(n-y)}{(1-\pi)^2}\right] = \frac{n}{\pi(1-\pi)}. \tag{29}$$

9. Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is no surprise.

Likelihood Function and ML Estimate for Binomial Parameter

10. Since $\mathcal{E}(Y) = n\pi$ and $\mathbf{Var}(Y) = n\pi(1-\pi)$, the distribution of $\hat{\pi} = y/n$ has mean and standard error as

$$\mathcal{E}(\hat{\pi}) = \pi, \quad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}. \tag{30}$$

Wald, Likelihood Ratio, and Score Tests

Wald, Likelihood Ratio, and Score Tests

1. Three standard ways exist to use the likelihood function to perform large-sample inference.

2. We introduce these for a significance test of a null hypothesis $H_0 : \beta = \beta_0$ and then discuss their relation to interval estimation.

3. They all exploit the large-sample normality of ML estimators.

Wald Test

1. The first method is **Wald statistic (Wald test)**.

2. With non-null standard error SE of $\hat{\beta}$, the test statistic

$$Z_W = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \tag{31}$$

has an approximate standard normal distribution when $\beta = \beta_0$. one refers Z_W to the standard normal table to obtain one- or two-sided p-value.

Wald Test

3. Equivalently, for the two-sided alternative, Z_W^2 has a chi-squared null distribution with 1 degree of freedom (df).

4. The p-value is then the right-tailed chi-squared probability above the observed value.

5. This type statistic, using the non-null standard error, is called a **Wald statistic** (Wald, 1943).

Wald Test

6. The multivariate extension for the Wald test of $H_0 : \underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\beta}}_0$ has test statistic

$$X^2_W = (\hat{\underline{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}_0)^T [\mathbf{Cov}(\hat{\underline{\boldsymbol{\beta}}})]^{-1} (\hat{\underline{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}_0). \tag{32}$$

   (The prime (T) on a vector or matrix denotes the transpose.)

7. The non-null covariance is based on the curvature (26) of the log likelihood at $\hat{\underline{\boldsymbol{\beta}}}$.

Wald Test

8. The asymptotic multivariate normal distribution for $\underline{\hat{\boldsymbol{\beta}}}$ implies an asymptotic chi-squared distribution for X^2_W.

9. The df equal the rank of $\mathbf{Cov}(\underline{\hat{\boldsymbol{\beta}}})$, which is the number of non-redundant parameters in $\underline{\boldsymbol{\beta}}$.

Likelihood Ratio Test (LRT)

1. A second general-purpose method uses the likelihood function through the ratio of two maximizations as **likelihood ratio test (likelihood ratio statistic)**:

   (a) The maximum over the possible parameter values under H_0.

   (b) The maximum over the possible parameter value permitting H_0 or an alternative H_A to be true.

Likelihood Ratio Test (LRT)

2. Let L_0 denote the maximized value of the likelihood function under H_0

3. Let L_1 denote the maximized value generally (i.e. under $H_0 \cup H_1$).

Likelihood Ratio Test (LRT)

4. For instance, for parameter $\underline{\boldsymbol{\beta}} = (\underline{\boldsymbol{\beta}}_1, \underline{\boldsymbol{\beta}}_0)$, and $H_0 : \underline{\boldsymbol{\beta}}_0 = \underline{\mathbf{0}}$,
   - L_1 is the likelihood function calculated at the $\underline{\boldsymbol{\beta}}$ for which the data would have been most likely.
   - L_0 is the likelihood function calculated at the $\underline{\boldsymbol{\beta}}_1$ value for which the data would be most likely, when $\underline{\boldsymbol{\beta}}_0 = \underline{\mathbf{0}}$.

Likelihood Ratio Test (LRT)

5. Then L_1 is always at least as large as L_0, since L_0 results from maximizing over a restricted set of parameter values.

6. The ratio $\Lambda = L_0/L_1$ of the maximized likelihood cannot exceed 1.

7. Wilks (1935, 1938) showed that $-2\log\Lambda$ has a limiting null chi-squared distribution, as $n \to \infty$.

8. The degrees of freedom (**df**) equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_1$ and under H_0.

Likelihood Ratio Test (LRT)

9. The likelihood-ratio test statistic equals

$$X^2_{LR} = -2\log\Lambda = -2log(L_0/L_1) = -2(\ell_0 - \ell_1) \tag{33}$$

where ℓ_0 and ℓ_1 denote the maximized log-likelihood functions.

Score Test

1. The third method is the **score statistic**, due to R.A. Fisher and C.R. Rao.

2. The score test is based on the slope and expected curvature of the log-likelihood function $\ell(\beta)$ at the null value β_0.

3. It uterizes the size of the score function

$$\mathcal{U}(\beta) = \partial\ell(\beta)/\partial(\beta), \tag{34}$$

   evaluated at β_0.

4. The (β) tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0.

Score Test

5. Denote $-\mathcal{E}[\partial^2 \ell(\beta)/\partial \beta^2]$ (i.e., the Fisher's information) evaluated at β_0 by $\boldsymbol{I}(\beta_0)$.

6. The score statistic is the ratio of $\mathcal{U}(\beta_0)$ to its null SE, which is $[\boldsymbol{I}(\beta_0)]^{1/2}$.

7. This has an approximate standard normal null distribution.

Score Test

8. The chi-squared form of the score statistic is

$$X^2_{SC} = \frac{[\mathcal{U}(\beta_0]^2}{\boldsymbol{I}(\beta_0)} = \frac{[\partial \ell(\beta)/\partial(\beta)]^2}{-\mathcal{E}[\partial^2 \ell(\beta)/\partial \beta^2]}, \tag{35}$$

where the partial derivative notation reflects derivatives with respect to β that are evaluated at β_0.

Score Test

9. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to $\underline{\boldsymbol{\beta}}$ and the inverse information matrix, both evaluated at the H_0 estimates (i.e., assuming that $\underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\beta}}_0$).

Wald, Likelihood Ratio, and Score Tests

1. We consider the three tests $H_0 : \beta = \beta_0$ for the univariate case for $\beta_0 = 0$.

2. The Wald test uses the behavior of $\ell(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/SE)^2$.

3. The SE of $\hat{\beta}$ depends on the curvature of $\ell(\beta)$ at $\hat{\beta}$.

Wald, Likelihood Ratio, and Score Tests

4. The score test is based on the slope and curvature of $\ell(\beta)$ at β_0.

5. The likelihood-ratio test combines information about $\ell(\beta)$ at both β and $\beta_0 = 0$.

6. In a sense, this statistic uses the most information of the three types of test statistic and is the most versatile.

Wald, Likelihood Ratio, and Score Tests

7. As $n \to \infty$, the Wald test, likelihood ratio test, and score test have certain asymptotic equivalences (Cox and Hinkley, 1974, sec, 9.3).

8. For small to moderate sample sizes, the likelihood ratio test is usually more reliable than the Wald test.

Constructing Confidence Intervals

Constructing Confidence Intervals

1. In practice, it is more informative to construct confidence intervals for parameters than to test hypothesis about their values.

2. For any of the three test methods, a confidence interval results form inverting the test.

3. For instance, a 95% confidence interval for β is the set of β_0 for which the test of $H_0 : \beta = \beta_0$ has a p-value exceeding 0.05.

Constructing Confidence Intervals

4. Let $Z_{1-\alpha}$ denote the Z-score form the standard normal distribution having right-tailed probability α; this is the $100(1-\alpha)$ percentile of that distribution.

5. Let $\chi^2_{\text{df},(1-\alpha)}$ denote the $100(1-\alpha)$ percentile of the chi-squared distribution with degrees of freedom df. $100(1-\alpha)\%$ confidence intervals based on asymptotic normality use $Z_{1-\alpha/2}$, for instance $Z_{1-0.05/2} = 1.96$ for 95% confidence.

Wald Confidence Intervals

6. The Wald confidence interval is the set of β_0 for which

$$| \hat{\beta} - \beta_0 | / SE < Z_{1-\alpha/2}. \tag{36}$$

7. This gives the interval $\hat{\beta} \pm Z_{1-\alpha/2}(SE(\hat{\beta}))$.

Likelihood Ratio Confidence Intervals

8. The likelihood ratio based confidence interval is the set of β_0 for which

$$X^2_{LR} = -2[\ell(\beta_0) - \ell(\beta)] < \boldsymbol{\chi}^2_{1,(1-\alpha)}. \tag{37}$$

[Recall that $\boldsymbol{\chi}^2_{1,(1-\alpha)} = Z^2_{1-\alpha/2}$.]

Likelihood Ratio Confidence Intervals

9. When $\hat{\beta}$ has a normal distribution, the log-likelihood function has a parabolic shape (i.e., a second-degree polynomial).

10. For small samples with categorical data, $\hat{\beta}$ may be far from normality and the log-likelihood function can be far from as symmetric, parabolic shape curve.

11. This can also happen with moderate to large samples

12. when a model contains many parameters.

13. In such case, inference based on asymptotic normality of $\hat{\beta}$ may have inadequate performance.

Constructing Confidence Intervals

14. A marked divergence in results of Wald and likelihood ratio inference indicates that the distribution of $\hat{\beta}$ may not be close to normality.

15. In may such cases, inference can instead utilize an exact small-sample distribution or "higher-order" asymptotic methods that improve on simple normality. (e.g., Pierce and Peters 1992).

Constructing Confidence Intervals

16. The Wald confidence interval is most common in practice because it is simple to construct using ML estimates and standard errors reported by statistical software.

17. The likelihood ratio based intervals becoming more widely available in software and is preferable for categorical data with small to moderate n.

18. For the best known statistical model, regression for a normal response, the three types of inference necessarily provide identical results.

Statistical Inference for Binomial Parameters

Statistical Inference for Binomial Parameters

1. For binomial parameter π, we obtain the likelihood function and ML estimator $\hat{\pi} = y/n$. Consider $H_0 : \pi = \pi_0$.

2. Since H_0 has a single parameter, we use the normal rather than chi-squared forms of Wald and score test statistics.

3. They permit tests against one-side as well as two-sided alternatives.

4. The Wald statistic is

$$Z_W = \frac{\hat{\pi} - \pi_0}{SE(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \tag{38}$$

Statistical Inference for Binomial Parameters

5. Evaluation the binomial score (28) and information (29) at π_0 yields

$$\mathcal{U}(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \quad \boldsymbol{I} = \frac{n}{\pi_0(1-\pi_0)}. \tag{39}$$

6. The normal form of the score statistic simplifies to

$$Z_{\mathsf{SC}} = \frac{\mathcal{U}(\pi_0)}{[\boldsymbol{I}(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}. \tag{40}$$

Statistical Inference for Binomial Parameters

7. Whereas the Wald statistic Z_W uses the standard error evaluate at $\hat{\pi}$, the score statistic Z_{SC} uses it evaluated at π_0.

8. The score statistic is preferable, as it uses the actual SE rather than an estimate.

9. Its null sampling distribution is closer to standard normal than that of the Wald statistic.

Statistical Inference for Binomial Parameters

10. The binomial log-likelihood function (27) equals
    $\ell_0 = y \log \pi_0 + (n-y) \log(1-\pi_0)$ under H_0 and
    $\ell_1 = y \log \hat{\pi} + (n-y) \log(1-\hat{\pi})$ more generally.

11. The likelihood ratio test statistic simplifies to

$$X^2_{LR} = -2(\ell_0 - \ell_1) = 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n-y) \log \frac{1-\hat{\pi}}{1-\pi_0} \right). \tag{41}$$

Statistical Inference for Binomial Parameters

12. Expressed as

$$X^2_{LR} = -2(\ell_0 - \ell_1) = 2\left(y \log \frac{y}{n\pi_0} + (n-y) \log \frac{n-y}{n - n\pi_0}\right). \quad (42)$$

13. It compares observed success and failure counts to fitted (i.e. null) counts by

$$2 \sum \text{observed } \log \frac{\text{observed}}{\text{fitted}}. \quad (43)$$

Statistical Inference for Binomial Parameters

14. We'll see that this formula also holds for tests about Poisson and multinomial parameters.

15. Since no unknown parameters occur under H_0 and one occurs under H_A, (43) has an asymptotic chi-squared distribution with df $= 1$.

Intervals for a Binomial Parameter

1. A significance test merely indicates whether a particular π value (such as $\pi = 0.5$) is plausible.

2. We learn more by using a confidence interval to determine the range of plausible values.

Intervals for a Binomial Parameter

3. Inverting the Wald test statistic gives the interval of π_0 values for which $|Z_W| < Z_{1-\alpha/2}$, or

$$\hat{\pi} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{44}$$

4. Historically, this was one of the first confidence interval used for any parameter (Laplace 1812, p.283).

Intervals for a Binomial Parameter

5. Unfortunately, it performs poorly unless n is very large (e.g., Brown et al., 2001).

6. The actual coverage probability usually falls below the nominal confidence coefficient, much below when π is near 0 or 1.

7. A simple adjustment that adds $\frac{1}{2}(Z_{1-\alpha/2})^2$ observations of each type to the sample before using this formula contains π_0 values for which $\mid z_s \mid < Z_{1-\alpha/2}$.

Intervals for a Binomial Parameter

8. The score confidence interval contains π_0 values for which $|z_{SC}| < Z_{1-\alpha/2}$.

9. Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0) / \sqrt{\pi_0(1-\pi_0)/n} = \pm Z_{1-\alpha/2}. \tag{45}$$

10. These are quadratic in π_0.

Intervals for a Binomial Parameter

11. First discussed by E. B. Wilson (1927), this score interval

$$\hat{\pi}\left(\frac{n}{n+Z^2_{1-\alpha/2}}\right)+\frac{1}{2}\left(\frac{Z^2_{1-\alpha/2}}{n+Z^2_{1-\alpha/2}}\right)$$
$$\pm Z_{1-\alpha/2}\sqrt{\frac{1}{n+Z^2_{1-\alpha/2}}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+Z^2_{1-\alpha/2}}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{Z^2_{1-\alpha/2}}{n+Z^2_{1-\alpha/2}}\right)\right]}. \quad (46)$$

Intervals for a Binomial Parameter

12. The midpoint $\tilde{\pi}$ of the interval is a weighted average of $\hat{\pi}$ and $\frac{1}{2}$, where the weight $n/(n+Z^2_{1-\alpha/2})$

13. given $\hat{\pi}$ increases as n increases. Combining terms, this midpoint equals $\tilde{\pi} = (y+Z^2_{1-\alpha/2})/(n+Z^2_{1-\alpha/2})$.

14. This is the sample proportion for an adjusted sample that add $Z^2_{1-\alpha/2}$ observations, half of each type.

Intervals for a Binomial Parameter

15. The square of the coefficient of $Z^2_{1-\alpha/2}$ in this formula is a weighted average of the variance of a ample proportion when $\pi = \hat{\pi}$ and the variance of a sample proportion when $\pi = 1/2$, using the adjusted sample size $n = Z^2_{1-\alpha/2}$in place of n.

16. This interval has much better performance than the Wald interval.

Intervals for a Binomial Parameter

17. The likelihood ratio based confidence interval is more complex computationally, but simple in principle.

18. It is the set of π_0 for which the likelihood ratio test has a p-value exceeding α.

19. Equivalent, it is the set of π_0 for which double the log likelihood drops by less than $\boldsymbol{\chi}^2_{1,(1-\alpha)}$ from its value at the ML estimate $\hat{\pi} = y/n$.

20. That is

$$X^2_{LR} = -2(\ell_0 - \ell_1) - 2[\ell(\pi_0) - \ell(\hat{\pi})] \leq \boldsymbol{\chi}^2_{1,1-\alpha} = 3.94. \tag{47}$$

Exact Small-Sample Inference

Exact Small-Sample Inference

1. With modern computational power, it is not necessary to rely on large-sample approximation for the distribution of statistics such as $\hat{\pi}$.

2. Tests and confidence intervals can use the binomial distribution directly rather than its normal approximation.

3. Such inferences occur naturally for small samples, but apply for any n.

Exact Small-Sample Inference

4. We illustrate by testing $H_0 : \pi = 0.5$ against $H_A : \pi \neq 0.5$ for the survey results, $y = 0$, with $n = 25$.

5. We noted that the score statistic equals $z = -5.0$.

6. The exact p-value for this statistic, based on the null $Bin(25, 0.5)$ distribution, is

   $$P(\,|z|\, \geq 5.0) = P(Y = 0 \text{ or } Y = 25) = 0.5^{25} + 0.5^{25} = 0.00000006.$$

   $100(1-\alpha)\%$ confidence intervals consists of all π_0 for which p-values exceed α in exact binomial tests.

Exact Small-Sample Inference

7. The best known interval (**Clopper and Person**, 1934) uses the tail method for forming confidence intervals.

8. It requires each one-sided p-value to exceed $\alpha/2$.

Exact Small-Sample Inference

9. The lower and upper endpoints are the solutions in π_0 to the equations

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

and

$$\sum_{k=0}^{y} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \alpha/2, \tag{48}$$

except that the lower bound is 0 when $y = 0$ and he upper bound is 1 when $y = n$.

Exact Small-Sample Inference

10. When $y = 1, 2, \ldots, n-1$ for connections between binomial sums and the incomplete beta function and related cumulative distribution functions (cdf's) of beta and F distribution, the confidence interval equals

$$\left[1 + \frac{n-y+1}{y\ F_{2y,2(n-y+1),\alpha/2}}\right]^{-1} < \pi < \left[1 + \frac{n-y+1}{(y+1)\ F_{2(y+1),2(n-y),(1-\alpha/2)}}\right]^{-1}, \tag{49}$$

where $F_{a,b,c}$ denotes the c quantile form the F distribution with degrees of freedom a and b.

11. When $y = 0$ with $n = 25$, the **Clopper-Pearson** 95% **confidence interval** for π is $(0, 0.137)$.

Exact Small-Sample Inference

12. In principle this approach seems ideal.

13. However, there is a serious complication.

14. Because of discreteness, the actual coverage probability for any π is at least as large as the nominal confidence level (Casella and Berger, 2001, p.434; Neyman, 1935) and it can be much greater.

15. Similarly, for a test of $H_0 : \pi = \pi_0$ at a fixed desired size α such as 0.05, it is not usually possible to achieve that size.

16. There is a finite number of possible samples, and hence a finite number of possible p-values, of which 0.05 may not be one.

Exact Small-Sample Inference

17. In testing H_0 with fixed π_0, one can pick a particular α that can occur as a p-value.

18. For interval estimation, however, this is not a option.

19. This is because constructing the interval corresponds to inverting an entire range of π_0 values in $H_0 : \pi = \pi_0$ and each distinct π_0 value can have its own set of possible p-value; that is, there is not a single null parameter value π_0 as in one test.

Exact Small-Sample Inference

20. For any fixed parameter value, the actual coverage probability coverage probability can be much more larger than the nominal confidence level.

21. The coverage probabilities are too low for the Wald method, whereas he Clopper-Pearson method errs in the opposite direction.

22. The score method behaves well, except for some π values close to 0 and 1.

23. Its coverage probabilities tend to be near the nominal level, not being consistently conservative or liberal. this a good method unless is very close o0 or 1.

Exact Small-Sample Inference

24. In discrete problems using small-sample distributions, shorter confidence intervals usually result from inverting a single two-sided test rather than two one-sided tests.

25. The interval is then the set of parameter values for which the p-value of a two-sided test exceeds α.

Exact Small-Sample Inference

26. For the binomial parameter, see Blaker (2000).

27. For observed outcome y_0, with Blaker's approach the p-value is the minimum of the two one-sided binomial probabilities $P(Y \geq y_0)$ and $P(Y \leq y_0)$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability.

28. The interval is computationally more complex, although available in software (Blaker gave S-Plus functions).

29. The result is still conservative, but less than the Clopper-Pearson interval.