

PubH 7407: Analysis of Categorical Data

- Syllabus: grading, TA's, topics covered.
- Lectures will be notes taken from text & (hopefully) posted the night before.
- We will follow the text pretty much in order covering non-starred sections, roughly one chapter per week.
- Sample SAS code is provided on Dr. Agresti's website. There is also a link to a large PDF file with R code. We will cover some aspects of fitting models in SAS.

1.1 Categorical response data

Response data considered in regression and ANOVA are continuous.

Examples:

- cholesterol level (milligrams per deciliter)
- lifetime of a lab rat (in weeks)
- money spent on breakfast cereal (U.S. \$)

A *categorical* variable takes on one of a (usually finite) number of categories, or levels. Examples:

- eye color (blue, brown, green, other)
- political affiliation (Democrat, Republican, other)
- cholesterol level (low, normal, high)

Note that a variable can be continuous or categorical depending on how it's defined.

1.1.1 Response or explanatory?

Variables can be classified as a *response* or *explanatory*.

In regression models we seek to model a response as a stochastic function of explanatory variables, or *predictors*.

In this course the response will be categorical and the predictors can be categorical, continuous, or discrete.

For example, if we wanted to model political affiliation as a function of gender and annual salary, the response would be (Republican, Democrat, other), and the two predictors would be annual salary (essentially continuous) and the categorical gender (male, female)

1.1.2 Nominal verses ordinal categorical variables

Nominal variables have no natural ordering to them. e.g. eye color (blue, brown, other), political affiliation (Democrat, Republican, other), favorite music type (jazz, folk, rock, rap, country, bluegrass, other), gender (male, female).

Ordinal variables have an obvious order to them. e.g. cancer stage (I, II, III, IV), a taste response to a new salsa (awful, below average, average, good, delicious).

Interval variables are ordinal variables that also have a natural scale attached to them. e.g. diastolic blood pressure, number of years of post high school education. Interval variables are typically discrete numbers that comprise an interval.

Read: Sections 1.1.3, 1.1.4, 1.1.5.

1.2 Distributions for categorical data

We review three distributions often used in the analysis of categorical data: binomial, multinomial, and Poisson distributions.

1.2.1 Binomial distribution

An observation is *binary* if it takes on one of two values (e.g. male/female, infected/non-infected, fail/pass). A random variable Y_i has a Bernoulli distribution with probability π if

$$P(Y_i = 1) = \pi \text{ and } P(Y_i = 0) = 1 - \pi.$$

Any binary variable can be written as Bernoulli by associating one category with 1 and the other category with 0.

Let Y_1, Y_2, \dots, Y_n be independent and identically distributed Bernoulli random variables with probability π . Then $Y = \sum_{i=1}^n Y_i$ has a binomial distribution with parameters n and π . We write this

$$Y \sim \text{bin}(n, \pi).$$

Y is the number of $\{Y_1, Y_2, \dots, Y_n\}$ that equal one. i.e. Y counts the number of 1's in the series Y_1, Y_2, \dots, Y_n .

Think of n identical experiments repeated independently of each other. e.g. $n = 5$ people are randomly chosen with replacement from all U of M students and classified as “undergraduate” ($Y_i = 0$) or “graduate” ($Y_i = 1$). Say we see $Y_1 = 0, Y_2 = 1, Y_3 = 1, Y_4 = 0, Y_5 = 0$. Then $Y = 0 + 1 + 1 + 0 + 0 = 2$.

The probability mass function of $Y \sim \text{bin}(n, \pi)$ is

$$P(Y = j) = \binom{n}{j} \pi^j (1 - \pi)^{n-j} \text{ for } j = 0, 1, \dots, n.$$

The mean and variance of Y is

$$E(Y) = n\pi, \quad \text{var}(Y) = n\pi(1 - \pi).$$

(Previous slide): If students are drawn *without* replacement, the number of undergraduates Y has a *hypergeometric* distribution.

1.2.2 Multinomial distribution

The binomial distribution deals with two categories. The multinomial distribution generalizes the binomial to C categories. Again, think of n independent trials, but now each trial results in an outcome $y_{ij} = 1$ if the outcome was the j^{th} category, and $y_{ij} = 0$ otherwise.

For example, let $C = 3$ outcomes, so the result of any trial is *one of* $\{1, 2, 3\}$. If the outcome is 1, then $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}) = (1, 0, 0)$. If the outcome is 2, then $\mathbf{y}_i = (0, 1, 0)$ and if the outcome is 3 then $\mathbf{y}_i = (0, 0, 1)$. Let $\mathbf{n} = \sum_{i=1}^n \mathbf{y}_i$. Then $\mathbf{n} = (n_1, \dots, n_C)$ gives the counts of the number out of n falling into each category.

Example: Let the outcome be political affiliation. Democrat corresponds to category 1, Republican is 2, and other is 3.

Let's say $n = 6$ people are sampled randomly and we see the categories 1,1,3,2,1,2. This corresponds to vectors $\mathbf{y}_1 = (1, 0, 0)$, $\mathbf{y}_2 = (1, 0, 0)$, $\mathbf{y}_3 = (0, 0, 1)$, $\mathbf{y}_4 = (0, 1, 0)$, $\mathbf{y}_5 = (1, 0, 0)$, and $\mathbf{y}_6 = (0, 1, 0)$. So $\mathbf{n} = (3, 2, 1)$, $n_1 = 3$ Democrats, $n_2 = 2$ Republicans, and $n_3 = 1$ other, out of $n = 6$.

Note that, always, $n_1 + n_2 + \cdots + n_C = n$.

Let $P(y_{ij} = 1) = \pi_j$. Note that $\pi_1 + \pi_2 + \cdots + \pi_C = 1$, so there are $C - 1$ free parameters in $\boldsymbol{\pi} = (\pi_1, \dots, \pi_C)$.

The pmf of \mathbf{n} is given by

$$p(n_1, \dots, n_C) = \frac{n!}{n_1! n_2! \cdots n_C!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_C^{n_C},$$

for $\sum_{j=1}^C n_j = n$ and each $0 \leq n_j \leq n$. We write

$$\mathbf{n} \sim \text{mult}(n, \boldsymbol{\pi}).$$

We have

$$E(n_j) = n\pi_j, \quad \text{var}(n_j) = n\pi_j(1 - \pi_j), \quad \text{cov}(n_j, n_k) = -n\pi_j\pi_k.$$

Marginally, $n_j \sim \text{bin}(n, \pi_j)$.

1.2.3 Poisson distribution

Some counts (essentially) have no fixed upper limit. The Poisson distribution counts events over time or space and is a limiting case of the binomial distribution when the number of trials n approaches infinity but events occur at a fixed rate μ (with units like surgeries/week, earthquakes/year, spelling errors per page, etc.)

We write $Y \sim \text{Pois}(\mu)$ with pmf

$$P(Y = j) = \frac{e^{-\mu} \mu^j}{j!} \text{ for } j = 0, 1, 2, \dots$$

For $Y \sim \text{Pois}(\mu)$ we have

$$E(Y) = \mu \text{ and } \text{var}(Y) = \mu.$$

Overdispersion: Often the variability associated with Poisson and binomial models is smaller than what is observed in real data. The increased variance can be attributed to unmeasured, or perhaps latent regressors in the model and thus the resulting count distribution is more correctly a *mixture* of binomial or Poisson distributions, with mixing weights being the proportion of outcomes resulting from specific (unaccounted for) covariate combinations.

We will discuss testing for overdispersion in specific models and remedies later on.

1.2.5 Connection between multinomial and Poisson distributions

Let $\mathbf{Y} = (Y_1, Y_2, Y_3)$ be independent Poisson with parameters (μ_1, μ_2, μ_3) .

e.g. (text): Y_1 is number of people that die in car accidents, Y_2 the number in airplane accidents, and Y_3 the number in railway accidents in Italy (over, say, a year). The total number of accidents $n = Y_1 + Y_2 + Y_3$ is $\text{Pois}(\mu_1 + \mu_2 + \mu_3)$.

Conditional on n , the distribution of (Y_1, Y_2, Y_3) is multinomial with parameters n and $\boldsymbol{\pi} = (\mu_1, \mu_2, \mu_3)/\mu_+$ where $\mu_+ = \mu_1 + \mu_2 + \mu_3$.

This is especially useful in log-linear models, covered in Chapters 8 and 9.

1.3 Inference for categorical data

1.3.1 Maximum likelihood estimation

Let the parameter vector for a model be $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ where p is the number of parameters in the model. Let the outcome variables be random variables denoted $\mathbf{y} = (y_1, \dots, y_n)$ and the probability model denoted

$$p(y_1, \dots, y_n | \boldsymbol{\beta}) = p(\mathbf{y} | \boldsymbol{\beta}).$$

The likelihood of $\boldsymbol{\beta}$, denoted $\mathcal{L}(\boldsymbol{\beta})$, is $\mathcal{L}(\boldsymbol{\beta}) = p(\mathbf{y} | \boldsymbol{\beta})$ thinking of data \mathbf{y} as fixed.

For example, if $\mathbf{n} = (n_1, \dots, n_c)$ is $\text{mult}(n, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_c)$, then $\boldsymbol{\beta} = (\pi_1, \pi_2, \dots, \pi_{c-1})$ because there are $c - 1$ free parameters in $\boldsymbol{\pi}$.

The likelihood of $\boldsymbol{\beta}$ is simply the probability of seeing the response data given $\boldsymbol{\beta}$:

$$\mathcal{L}(\boldsymbol{\beta}) = p(n_1, \dots, n_c | \boldsymbol{\beta}) = \binom{n}{n_1 \cdots n_c} \prod_{j=1}^c \pi_j^{n_j}.$$

The maximum likelihood estimator is that value of β that maximizes $\mathcal{L}(\beta)$ for given data:

$$\hat{\beta} = \operatorname{argmax}_{\beta \in \mathbf{B}} \mathcal{L}(\beta),$$

where \mathbf{B} is the set of values β can take on.

The MLE $\hat{\beta}$ makes the observed data as *likely as possible*. The estimator turns into an estimate when data are actually seen. For example, if $c = 3$ and $n_1 = 3$, $n_2 = 5$, $n_3 = 2$, then $\hat{\beta} = (\hat{\pi}_1, \hat{\pi}_2) = (0.3, 0.5)$ and of course $\hat{\pi}_3 = 1 - (\hat{\pi}_1 + \hat{\pi}_2) = 0.2$. Then $p(3, 5, 2 | \pi_1 = 0.2, \pi_2 = 0.5) \geq p(3, 5, 2 | \pi_1 = p_1, \pi_2 = p_2)$ for all values of p_1 and p_2 .

An estimator is random (i.e. before data are collected and seen they are random, and so then is any function of data) whereas an estimate is a fixed, known vector (like $(0.3, 0.5)$).

MLEs have nice properties for most (but not all) models (p. 9):

- They have large sample normal distributions:

$$\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_p(\boldsymbol{\beta}, \text{cov}(\hat{\boldsymbol{\beta}})) \text{ where } \text{cov}(\hat{\boldsymbol{\beta}}) = \left[-E \left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right) \right]_{p \times p}^{-1}.$$

- They are asymptotically consistent: $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$ (in probability) as the sample size $n \rightarrow \infty$.
- They are asymptotically efficient: $\text{var}(\hat{\beta}_j)$ is smaller than the corresponding variance of other (asymptotically) unbiased estimators.

Read: section 1.3.2.

Maximum likelihood for Poisson data

Let $Y_i \sim \text{Pois}(\lambda t_i)$ where λ is the unknown event rate and t_i are known exposure times. Assume the Y_1, \dots, Y_n are independent.

The likelihood of λ is

$$\begin{aligned}\mathcal{L}(\lambda) &= p(y_1, \dots, y_n | \lambda) = \prod_{i=1}^n p(y_i | \lambda) = \prod_{i=1}^n e^{-t_i \lambda} (t_i \lambda)^{y_i} / y_i! \\ &= \left[\prod_{i=1}^n \frac{t_i^{y_i}}{y_i!} \right] e^{-\lambda \sum_{i=1}^n t_i} \lambda^{\sum_{i=1}^n y_i} = g(\mathbf{t}, \mathbf{y}) e^{-\lambda \sum_{i=1}^n t_i} \lambda^{\sum_{i=1}^n y_i}.\end{aligned}$$

Then the log-likelihood is

$$L(\lambda) = \log g(\mathbf{t}, \mathbf{y}) - \lambda \sum_{i=1}^n t_i + \log(\lambda) \sum_{i=1}^n y_i.$$

Taking the derivative w.r.t. λ we get

$$L'(\lambda) = \frac{\partial L(\lambda)}{\partial \lambda} = - \sum_{i=1}^n t_i + \frac{1}{\lambda} \sum_{i=1}^n y_i.$$

Setting this equal to zero, plugging in \mathbf{Y} for \mathbf{y} , and solving for λ yields the MLE

$$\hat{\lambda} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n t_i}.$$

Now

$$\frac{\partial^2 L(\lambda)}{\partial \lambda^2} = - \frac{\sum_{i=1}^n y_i}{\lambda^2}.$$

Since $\sum_{i=1}^n Y_i \sim \text{Pois}(\lambda \sum_{i=1}^n t_i)$, we have

$$-E \left(\frac{\partial^2 L(\lambda)}{\partial \lambda^2} \right) = E \left(\frac{\sum_{i=1}^n Y_i}{\lambda^2} \right) = \frac{\lambda \sum_{i=1}^n t_i}{\lambda^2} = \frac{\sum_{i=1}^n t_i}{\lambda}.$$

The variance of $\hat{\lambda}$ is given by the “inverse” of this “matrix”

$$\text{var}(\hat{\lambda}) \doteq \frac{\lambda}{\sum_{i=1}^n t_i}.$$

The large sample normal result tells us

$$\hat{\lambda} \doteq N \left(\lambda, \frac{\lambda}{\sum_{i=1}^n t_i} \right).$$

The standard deviation of $\hat{\lambda}$ is estimated to be $\text{sd}(\hat{\lambda}) = \sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}}$.

Since we do not know λ , the standard deviation is estimated by the *standard error* obtained from estimating λ by its MLE:

$$\text{se}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{\sum_{i=1}^n t_i}} = \sqrt{\frac{\sum_{i=1}^n y_i}{[\sum_{i=1}^n t_i]^2}} = \frac{\sqrt{\sum_{i=1}^n y_i}}{\sum_{i=1}^n t_i}.$$

Example: Say that we record the number of adverse medical events (e.g. operating on the wrong leg) from a hospital over $n = 3$ different times: $t_1 = 1$ week in 2002, $t_2 = 4$ weeks in 2003, and $t_3 = 3$ weeks in 2005. We'll assume that the adverse surgical event rate λ (events/week) does not change over time and that event counts in different time periods are independent.

Then $Y_i \sim \text{Pois}(t_i\lambda)$ for $i = 1, 2, 3$. Say we observe $y_1 = 0$, $y_2 = 3$, and $y_3 = 1$. Then $\hat{\lambda} = (0 + 3 + 1)/(1 + 4 + 3) = 4/8 = 0.5$ event/week, or one event every other week. Also, $\text{se}(\hat{\lambda}) = \sqrt{4}/8 = 0.25$.

The large sample result tells us then (before data are collected and $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is random) that

$$\hat{\lambda} \overset{\bullet}{\sim} N(\lambda, 0.25^2),$$

useful for constructing hypothesis tests and confidence intervals.

1.3.3 Wald, likelihood ratio, and score tests

These are three ways to perform large sample hypothesis tests based on the model likelihood.

Wald test

Let \mathbf{M} be a $m \times p$ matrix. Many hypotheses can be written $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ where \mathbf{b} is a known $m \times 1$ vector.

For example, let $p = 3$ so $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. The test of $H_0 : \beta_2 = 0$ is written in matrix terms with $\mathbf{M} = (0, 1, 0)$ and $b = 0$. The hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 \text{ has } \mathbf{M} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The large sample result for MLEs is

$$\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_p(\boldsymbol{\beta}, \text{cov}(\hat{\boldsymbol{\beta}})).$$

So then

$$\mathbf{M}\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_m(\mathbf{M}\boldsymbol{\beta}, \mathbf{M}\text{cov}(\hat{\boldsymbol{\beta}})\mathbf{M}').$$

If $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ is true then

$$\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b} \overset{\bullet}{\sim} N_m(\mathbf{0}, \mathbf{M}\text{cov}(\hat{\boldsymbol{\beta}})\mathbf{M}').$$

So

$$W = (\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b})'[\widehat{\mathbf{M}\text{cov}(\hat{\boldsymbol{\beta}})\mathbf{M}'}]^{-1}(\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b}) \overset{\bullet}{\sim} \chi_m^2.$$

W is called the Wald statistic and large values of W indicate $\mathbf{M}\boldsymbol{\beta}$ is far away from \mathbf{b} , i.e. that H_0 is false. The p -value for $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ is given by $p - \text{value} = P(\chi_m^2 > W)$.

The simplest, most-used Wald test is the familiar test that a regression effect is equal to zero, common to multiple, logistic, Poisson, and ordinal regression models.

Score tests

In general, the $\text{cov}(\hat{\boldsymbol{\beta}})$ is a function of the unknown $\boldsymbol{\beta}$. The Wald test replaces $\boldsymbol{\beta}$ by its MLE $\hat{\boldsymbol{\beta}}$ yielding $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$. The score test replaces $\boldsymbol{\beta}$ by the the MLE $\hat{\boldsymbol{\beta}}_0$ obtained under the constraint imposed by H_0

$$\hat{\boldsymbol{\beta}}_0 = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbf{B}: \mathbf{M}\boldsymbol{\beta} = \mathbf{b}} \mathcal{L}(\boldsymbol{\beta}).$$

As with the Wald test, the resulting test statistic

$$S = (\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b})' [\mathbf{M}\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\mathbf{M}']^{-1} (\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b}) \overset{\bullet}{\sim} \chi_p^2.$$

Sometimes it is easier to fit the reduced model rather than the full model; the score test allows testing whether new parameters are necessary from a fit of a smaller model.

Likelihood ratio tests

The likelihood ratio test is easily constructed and carried out for nested models. The full model has parameter vector $\boldsymbol{\beta}$ and the reduced model obtains when $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ holds. A common example is when $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and we wish to test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ (e.g. a subset of regression effects are zero). Let $\hat{\boldsymbol{\beta}}$ be the MLE under the full model

$$\hat{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbf{B}} \mathcal{L}(\boldsymbol{\beta}),$$

and $\hat{\boldsymbol{\beta}}_0$ be the MLE under the constraint imposed by H_0

$$\hat{\boldsymbol{\beta}}_0 = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbf{B} : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}} \mathcal{L}(\boldsymbol{\beta}).$$

If $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ is true,

$$L = -2[\log \mathcal{L}(\hat{\boldsymbol{\beta}}_0) - \log \mathcal{L}(\hat{\boldsymbol{\beta}})] \overset{\bullet}{\sim} \chi_m^2.$$

The statistic L is the likelihood ratio test statistic for the hypothesis $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$. The smallest L can be is zero when $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. The more different $\hat{\boldsymbol{\beta}}$ is from $\hat{\boldsymbol{\beta}}_0$, the larger L is and the more evidence there is that H_0 is false. The p -value for testing H_0 is given by $p - \text{value} = P(\chi_m^2 > L)$.

To test whether additional parameters are necessary, LRT tests are carried out by fitting two models: a “full” model with all effects and a “reduced” model. In this case the dimension m of \mathbf{M} is the difference in the numbers of parameters in the two models.

For example, say we are fitting the standard regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

where $e_i \stackrel{iid}{\sim} N(0, \sigma^2)$. Then $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$ and we want to test $\boldsymbol{\beta}_1 = (\beta_2, \beta_3) = (0, 0)$, that the 2nd and 3rd predictors aren't needed. This test can be written using matrices as

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} .$$

The likelihood ratio test fits the full model above and computes $L_f = \log \mathcal{L}_f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma})$.

Then the reduced model $Y_i = \beta_0 + \beta_1 x_{i1} + e_i$ is fit and $L_r = \log \mathcal{L}_r(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ computed.

The test statistic is $L = -2(L_r - L_f)$; a p -value is computed as $P(\chi_2^2 > L)$. If the p -value is less than, say, $\alpha = 0.05$ we reject $H_0 : \beta_2 = \beta_3 = 0$.

Of course we wouldn't use this approximate LRT test here! We have outlined an approximate test, but there is well-developed theory that instead uses a different test statistic with an exact F -distribution.

Note that:

- The Wald test requires maximizing the unrestricted likelihood.
- The score test requires maximizing the restricted likelihood (under a nested submodel).
- The Likelihood ratio test requires both of these.

So the likelihood ratio test uses more information and both Wald and Score tests can be viewed as approximations to the LRT.

However, SAS can “automatically” perform Wald tests of the form $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ in a `contrast` statement and so I often use Wald tests because they’re easy to get. In large samples the tests are equivalent.

1.3.4 Confidence intervals

A plausible range of values for a parameter β_j (from $\boldsymbol{\beta}$) is given by a confidence interval (CI). Recall that a CI has a certain fixed probability of containing the unknown β_j before data are collected. After data are collected, nothing is random any more, and instead of “probability” we refer to “confidence.”

A common way of obtaining confidence intervals is by *inverting* hypothesis tests of $H_0 : \beta_k = b$. Without delving into why this works, a $(1 - \alpha)100\%$ CI is given by those b such that the p -value for testing $H_0 : \beta_k = b$ is larger than α .

For Wald tests of $H_0 : \beta_k = b$, the test statistic is $W = (\hat{\beta}_k - b)/\text{se}(\hat{\beta}_k)$. This statistic is approximately $N(0, 1)$ when $H_0 : \beta_k = b$ is true and the p -value is larger than $1 - \alpha$ only when $|W| < z_{\alpha/2}$ where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a $N(0, 1)$ random variable. This yields the well known CI

$$(\hat{\beta}_k - z_{\alpha/2}\text{se}(\hat{\beta}_k), \hat{\beta}_k + z_{\alpha/2}\text{se}(\hat{\beta}_k)).$$

The likelihood ratio CI operates in the same way, but the log-likelihood must be computed for all values of b . We'll explore the differences between inverting Wald, Score, and LRT for binomial data in the remainder of Chapter 1.