Analysis of Categorical Data
Extra Information

Christopher R. Bilder[1] and Thomas M. Loughin[2]

[1]University of Nebraska–Lincoln, Department of Statistics

[2]Simon Fraser University, Department of Statistics and Actuarial Science

www.chrisbilder.com/categorical

- Binary responses likely the most common type of categorical response
  - Define $Y = 1$ as a "success" with probability π
  - Define $Y = 0$ as a "failure" with probability $1 - \pi$
- Bernoulli distribution

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

  for $y = 0$ or 1
- $E(Y) = \pi$ and $Var(Y) = \pi(1 - \pi)$
- Binomial distribution
  - Observe multiple Bernoulli random variables, say $Y_1, \ldots, Y_n$, through repeated sampling or trials in identical settings
  - If all trials are identical and independent, $W = \sum_{i=1}^{n} Y_i$ has a binomial distribution:

$$P(W = w) = \binom{n}{w} \pi^w (1 - \pi)^{n-w}$$

    for $w = 0, \ldots, n$
  - $E(W) = n\pi$ and $Var(W) = n\pi(1 - \pi)$
- Goal: Estimate π

- Given observed data, what is the most plausible value of π?
- Maximum likelihood estimation
  - Likelihood function measures the plausibility of different values of π
  - Bernoulli setting

$$\begin{aligned} L(\pi|y_1,\ldots,y_n) &= P(Y_1 = y_1) \times \cdots \times P(Y_n = y_n) \\ &= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\ &= \pi^{w}(1-\pi)^{n-w} \end{aligned}$$

  - Binomial setting: $L(\pi|w) = P(W = w) = \binom{n}{w}\pi^{w}(1-\pi)^{n-w}$
- The value of π which maximizes the likelihood function is considered to be the most plausible
  - Maximum likelihood estimate (MLE)
  - Derive MLE to be $\hat{\pi} = w/n$
  - For more complicated likelihood functions, will need to use numerical iterative methods

- Maximum likelihood estimators have a normal distribution for a large sample
  - Suppose $\hat{\theta}$ is MLE of θ
  - Mean is θ
  - $Var(\hat{\theta})$ is estimated by

$$-E\left(\frac{\partial^2}{\partial\theta^2}\log[L(\theta|W)]\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

    where $\log(\cdot)$ is the natural log function
- Bernoulli/binomial:
  - $\hat{\pi} = w/n$ is MLE
  - Mean is π
  - Estimated variance is

$$\begin{aligned}\widehat{Var}(\hat{\pi}) &= -E\left\{\frac{\partial^2\log\left[L(\pi|W)\right]}{\partial\pi^2}\right\}^{-1}\Bigg|_{\pi=\hat{\pi}} = -E\left\{-\frac{W}{\pi^2}+\frac{n-W}{(1-\pi)^2}\right\}^{-1}\Bigg|_{\pi=\hat{\pi}} \\ &= \left[\frac{n}{\pi}-\frac{n}{1-\pi}\right]^{-1}\Bigg|_{\pi=\hat{\pi}} = \frac{\hat{\pi}(1-\hat{\pi})}{n}\end{aligned}$$

- See Casella and Berger (2002) for more details about maximum likelihood estimation

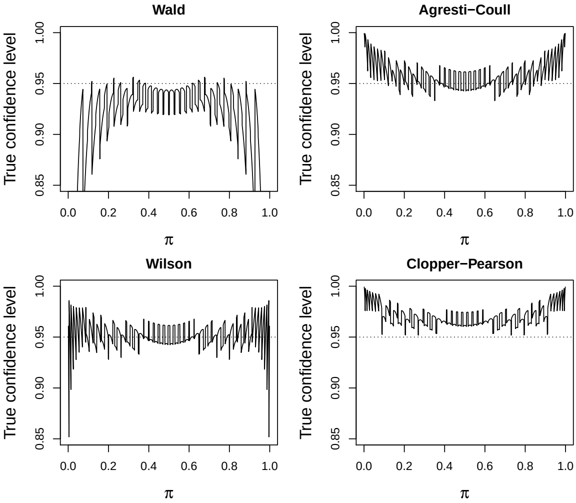
- Wald interval
  - Use large-sample normality of maximum likelihood estimator
  - $(1-\alpha)100\%$ confidence interval for π

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

    where Z_a is the a^{th} quantile from a standard normal distribution (e.g., $Z_{0.975} = 1.96$)
  - Problems:
    - Limits may be less than 0 or greater than 1
    - When $w = 0$ or n, $\sqrt{\hat{\pi}(1-\hat{\pi})/n} = 0$, leading to an interval of (0,0) or (1,1)
    - True confidence level (coverage) is very often less than $(1-\alpha)100\%$

Example: True confidence levels, interval for π (ConfLevel4Intervals.R)

- $n = 40$ and $\alpha = 0.05$
- When $\pi = 0.157$, true confidence level is 0.8759 for Wald interval
- Plots for $0 < \pi < 1$:



- Wilson (score) interval
  - $H_0 : \pi = \pi_0$ vs. $H_a : \pi \neq \pi_0$
  - Score statistic
$$Z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$
  - Approximate with a standard normal distribution and use $\pm Z_{1-\alpha/2}$ as critical values
  - Invert the test to find interval
    - Find all possible values for π_0 that lead to a "do not reject" of H_0
    - Results in
$$\tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + \frac{Z^2_{1-\alpha/2}}{4n}}$$
    where
$$\tilde{\pi} = \frac{w + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}}$$
  - Benefits:
    - Limits always between 0 and 1
    - Decent true confidence level properties

Example: Corn seed germination (Corn.R)

- My garden

- Planted 64 corn seeds of a particular variety in one $4' \times 4'$ raised bed
- Followed seed packet directions
- After 21 days, 48 seeds had sprouted (7-14 days was period given on seed packet)

Example: Corn seed germination (Corn.R)

```
> w <- 48
> n <- 64
> alpha <- 0.05
> pi.hat <- w/n
> pi.hat
[1] 0.75
> pi.tilde <- (w + qnorm(p = 1 - alpha/2)^2/2)/(n + qnorm(p = 1 -
    alpha/2)^2)
> pi.tilde
[1] 0.7358
> wilson <- pi.tilde + qnorm(p = c(alpha/2, 1 - alpha/2)) * sqrt(n)/(n +
    qnorm(p = 1 - alpha/2)^2) * sqrt(pi.hat * (1 - pi.hat) +
    qnorm(p = 1 - alpha/2)^2/(4 * n))
> round(wilson, digits = 4)
[1] 0.6318 0.8399
> library(package = binom)
> binom.confint(x = w, n = n, conf.level = 1 - alpha, methods = "wilson")
  method  x  n mean  lower  upper
1 wilson 48 64 0.75 0.6318 0.8399
```

- Compare to 95% Wald interval: $0.6439 < \pi < 0.8561$

- Denote π_1 and π_2 as the probabilities of a success for the two groups
- 2×2 contingency tables

		Response		
		Success	Failure	Total
Group	1	π_1	$1 - \pi_1$	1
	2	π_2	$1 - \pi_2$	1

		Response		
		Success	Failure	Total
Group	1	w_1	$n_1 - w_1$	n_1
	2	w_2	$n_2 - w_2$	n_2

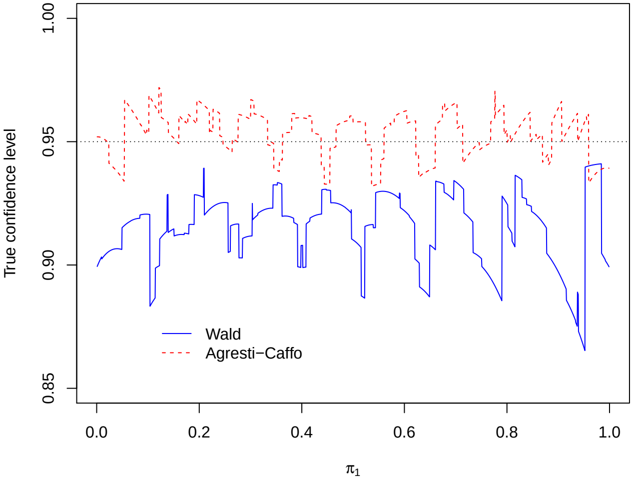
- $W_j \sim \text{Binomial}(n_j, \pi_j)$ for $j = 1, 2$
  - MLE for π_j: $\hat{\pi}_j = w_j / n_j$
  - $\hat{\pi}_j \dot{\sim} N(\pi_j, \widehat{Var}(\hat{\pi}_j))$ for large n_j, where $\widehat{Var}(\hat{\pi}_j) = \hat{\pi}_j(1 - \hat{\pi}_j)/n_j$
- $(1 - \alpha)100\%$ Wald interval

$$\hat{\pi}_1 - \hat{\pi}_2 \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

- Problems with Wald interval:
  - Limits may be less than -1 or greater than 1
  - When $w_j = 0$ or n_j, the $\hat{\pi}_j(1 - \hat{\pi}_j)/n_j$ part of the variance becomes 0
  - True confidence level (coverage) is very often less than $(1 - \alpha)100\%$

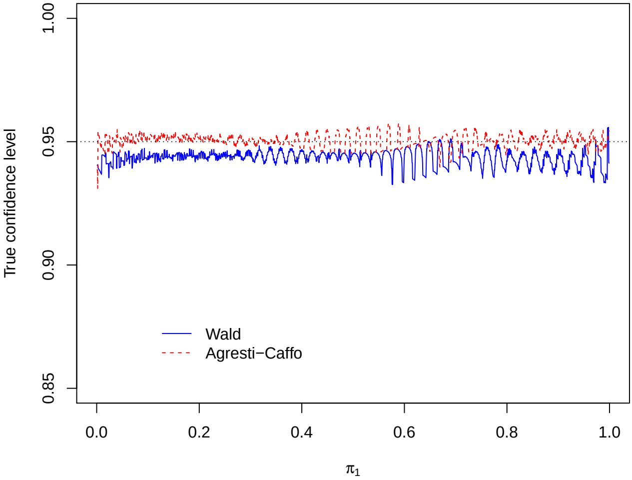
Example: True confidence levels, interval for $\pi_1 - \pi_2$ (ConfLevelTwoProb.R)

- $n_1 = n_2 = 10$, $\pi_2 = 0.4$, and $\alpha = 0.05$



Example: True confidence levels, interval for $\pi_1 - \pi_2$ (ConfLevelTwoProb.R)

- $n_1 = n_2 = 50$, $\pi_2 = 0.4$, and $\alpha = 0.05$



- $(1-\alpha)100\%$ Agresti-Caffo interval

$$\tilde{\pi}_1 - \tilde{\pi}_2 \pm Z_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1+2} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2+2}}$$

  where

$$\tilde{\pi}_1 = \frac{w_1+1}{n_1+2} \text{ and } \tilde{\pi}_2 = \frac{w_2+1}{n_2+2}$$

  - Benefit: True confidence level is much closer to $(1-\alpha)100\%$ than Wald
- Score interval
  - $H_0 : \pi_1 - \pi_2 = d$ vs. $H_a : \pi_1 - \pi_2 \neq d$
  - Invert test
  - Performs similarly to Agresti-Caffo interval
  - No closed form expression
  - See p. 57 of Bilder and Loughin (2014)

Example: Larry Bird free throws (Bird.R)

```
> c.table <- array(data = c(251, 48, 34, 5), dim = c(2, 2),
     dimnames = list(First = c("made", "missed"), Second = c("made",
         "missed")))
> c.table
        Second
First    made missed
  made    251     34
  missed   48      5
> c.table[1, 2]  #Row 1, column 2 count
[1] 34
> pi.tilde1 <- (c.table[1, 1] + 1)/(sum(c.table[1, ]) + 2)
> pi.tilde2 <- (c.table[2, 1] + 1)/(sum(c.table[2, ]) + 2)
> var.AC <- pi.tilde1 * (1 - pi.tilde1)/(sum(c.table[1, ]) +
     2) + pi.tilde2 * (1 - pi.tilde2)/(sum(c.table[2, ]) +
     2)
> alpha <- 0.05
> pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1 - alpha/2)) *
     sqrt(var.AC)
[1] -0.10353  0.07781
```

Example: Larry Bird free throws (Bird.R)

```
> library(PropCIs)
> wald2ci(x1 = c.table[1, 1], n1 = sum(c.table[1, ]), x2 = c.table[2,
    1], n2 = sum(c.table[2, ]), conf.level = 0.95, adjust = "AC")



data:

95 percent confidence interval:
 -0.10353  0.07781
sample estimates:
[1] -0.01286
```

- With 95% confidence, the difference in the probability of success on the second attempt is between -0.1035 and 0.07781 when the first free throw is made vs. when the first free throw is missed
- Wald: $-0.1122 < \pi_1 - \pi_2 < 0.0623$; use `adjust = "Wald"` with `wald2ci()`
- Could enter values of w_1, n_1, w_2, n_2 directly into R rather than use contingency table structure

Example: Larry Bird free throws (Bird.R)

- What if the data was not already summarized in a contingency table format?

Observation	First	Second
1	Made	Made
2	Missed	Made
3	Made	Made
⋮	⋮	⋮
338	Made	Missed

- Suppose `all.data2` contains this form of the data

```
> bird.table2 <- xtabs(formula = ~first + second, data = all.data2)
> bird.table2

        second
first    made missed
  made    251     34
  missed   48      5

> # table(all.data2$first, all.data2$second) #This also works
```

- Proceed with using `bird.table2` object in place of `c.table`

- Meaning of $\pi_1 - \pi_2$ changes depending on the sizes of these probabilities
  - Two examples:
    1. $\pi_1 = 0.51$ and $\pi_2 = 0.50$
    2. $\pi_1 = 0.011$ and $\pi_2 = 0.001$
  - Both have $\pi_1 - \pi_2 = 0.01$, but
    1. Difference is small relative to size of probabilities
    2. Difference is large relative to size of probabilities
- Relative risk
  - $RR = \pi_1/\pi_2$
    1. $RR = 0.51/0.50 = 1.02$
    2. $RR = 0.011/0.001 = 11.0$
  - Interpretation for 2.:
    - A success is 11 times as likely for group 1 than for group 2
    - A success is 10 times more likely for group 1 than for group 2
- What if $RR = 1$?

- MLE: $\widehat{RR} = \hat{\pi}_1/\hat{\pi}_2$
- Wald confidence interval
  - Normal approximation is better for $\log(\hat{\pi}_1/\hat{\pi}_2)$ than for $\hat{\pi}_1/\hat{\pi}_2$
  - Estimated variance

$$\widehat{Var}(\log(\hat{\pi}_1/\hat{\pi}_2)) = \frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}$$

  - Interval for $\log(RR)$

$$\log(\hat{\pi}_1/\hat{\pi}_2) \pm Z_{1-\alpha/2}\sqrt{\frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}}$$

  - Interval for RR

$$\exp\left[\log(\hat{\pi}_1/\hat{\pi}_2) \pm Z_{1-\alpha/2}\sqrt{\frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}}\right]$$

  - What if w_1 or $w_2 = 0$? Possible ad-hoc solutions:
    - Add 0.5 to the count
    - Add 0.5 to all counts

Example: HIV vaccine (HIVvaccine.R)

```
> c.table <- array(data = c(51, 74, 8146, 8124), dim = c(2, 2),
     dimnames = list(Trt = c("vaccine", "placebo"), Response = c("HIV",
         "No HIV")))
> c.table
         Response
Trt        HIV No HIV
  vaccine  51   8146
  placebo  74   8124
> n1 <- sum(c.table[1, ])
> n2 <- sum(c.table[2, ])
> pi.hat1 <- c.table[1, 1]/n1
> pi.hat2 <- c.table[2, 1]/n2
> pi.hat1/pi.hat2
[1] 0.6893
```

- Article said "cut the risk of becoming infected with HIV by more than 31 percent"

Example: HIV vaccine (HIVvaccine.R)

```
> alpha <- 0.05
> var.log.RR <- 1/c.table[1, 1] - 1/n1 + 1/c.table[2, 1] - 1/n2
> RR.ci <- exp(log(pi.hat1/pi.hat2) + qnorm(p = c(alpha/2, 1 -
    alpha/2)) * sqrt(var.log.RR))
> round(RR.ci, 2)

[1] 0.48 0.98

> rev(round(1/RR.ci, 2))

[1] 1.02 2.07
```

- With 95% confidence,
  - HIV infection is between 0.48 and 0.98 times as likely for the vaccine group than for the placebo group
  - the probability of HIV infection is between 0.48 and 0.98 times as large for the vaccine group than for the placebo group
  - the vaccine reduces the probability of HIV infection by 2% to 52%
  - HIV infection is between 1.02 to 2.07 times as likely for the placebo group than for the vaccine group
  - HIV infection is between 0.02 to 1.07 times more likely for the placebo group than for the vaccine group
  - the probability of HIV infection is between 0.02 to 1.07 times larger for the placebo group than for the vaccine group

Example: HIV vaccine (HIVvaccine.R)

- The `twoby2()` function from the `Epi` package produces the same calculations

```
> library(package = Epi)
> twoby2(c.table, alpha = 0.05)
2 by 2 table analysis:
------------------------------------------------------
Outcome   : HIV
Comparing : vaccine vs. placebo

        HIV No HIV    P(HIV) 95% conf. interval
vaccine  51   8146    0.0062    0.0047   0.0082
placebo  74   8124    0.0090    0.0072   0.0113

                                   95% conf. interval
             Relative Risk:  0.6893    0.4831   0.9834
         Sample Odds Ratio:  0.6873    0.4805   0.9832
    Probability difference: -0.0028   -0.0055  -0.0001

        Asymptotic P-value: 0.0401
------------------------------------------------------
```

- Numerical iterative methods are used to determine regression parameter estimates
- Convergence decided by looking at ratio of successive residual deviances
  - Define $D^{(k)}$ as the residual deviance at iteration k
  - Convergence occurs when

$$\frac{\left|D^{(k)} - D^{(k-1)}\right|}{0.1 + \left|D^{(k)}\right|} < \epsilon$$

  where ϵ is small (`glm()` uses $\epsilon = 10^{-8}$)
- What if convergence does not occur?
  - Try a larger number of iterations (`glm()` uses `maxit = 25`)
  - Convergence may not be possible due to problems with the data

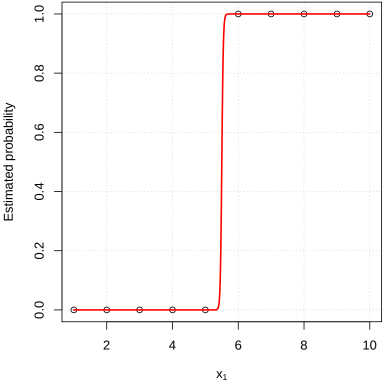
Example: Complete separation (Non-convergence.R)

- An explanatory variable(s) perfectly separates the data between y = 0 and 1 values
- MLE(s) is infinite

```
> set1 <- data.frame(x1 = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10), y = c(0,
    0, 0, 0, 0, 1, 1, 1, 1, 1))
> set1
   x1 y
1   1 0
2   2 0
3   3 0
4   4 0
5   5 0
6   6 1
7   7 1
8   8 1
9   9 1
10 10 1
```

Example: Complete separation (Non-convergence.R)

```
> mod.fit1 <- glm(formula = y ~ x1, data = set1,
     family = binomial(link = logit))
Warning:  glm.fit:  algorithm did not converge
Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
> mod.fit1$coefficients
(Intercept)          x1
     -245.8        44.7
```



- Use `trace = TRUE` in `glm()` to see iteration history

- R may indicate convergence occurs even with complete separation!
  - In previous example with a larger number of iterations, R will indicate convergence occurs
    - Reason: Because $\hat{\pi}$ values are so close to 0 or 1, there will be little change to $D^{(k)}$ for successive iterations despite $\hat{\beta}_1$ continuing to change
    - Still will print:
      `glm.fit:  fitted probabilities numerically 0 or 1 occurred`
  - What can you do?
    - Construct a plot like on previous slide
    - Use a stricter convergence criteria (smaller ϵ – change `epsilon` argument value) to determine if regression parameter estimates change for a larger number of iterations
    - Check if $\hat{\pi}$ values are very close to 0 or 1
- Alternative approaches if convergence does not occur
  - Exact logistic regression – See Section 6.2.3 of Bilder and Loughin (2014)
  - Include a "penalty" in the likelihood function – See Section 2.2.7 of Bilder and Loughin (2014)

Analysis of Categorical Data
Extra Information

Christopher R. Bilder[1] and Thomas M. Loughin[2]

[1]University of Nebraska–Lincoln, Department of Statistics

[2]Simon Fraser University, Department of Statistics and Actuarial Science

www.chrisbilder.com/categorical

R Index