Discrete Response Model
Lecture 1

datascience@berkeley

Introduction to Categorical Data, and the Bernoulli and Binomial Probability Models

What Is a Categorical (Qualitative) Variable?

- What is a categorical (qualitative) variable?
  - Patient survival: yes or no
  - Customer retention: churn or not
  - Produce color choice: blue, green, yellow, …
  - Self-reported health condition rating: 1,2,3,4,5
  - Customer satisfaction: Satisfied, Neutral, Unsatisfied
  - Highest attained education level : HS, BS, MS, PhD (ordinal properties)
  - Annual income: <15,000, 15,000-<25,000, 25,000-<40,000, ≥40,000 (ordinal properties)
- The first three examples do not have a natural ordering, and the last four examples have a natural ordering. We call them ordinal variables.
- We will focus on binary response in Lecture 1 and 2.

Binary Response Variable Observed From a Homogeneous Population

Goal: Estimate the overall probability of observing one of two possible outcomes for this random variable.

- This is often equated with the "probability of success" for an individual item in the population.
- Equivalently, this is the overall prevalence of successes in the population because each item has the same probability of success.

Bernoulli and Binomial Probability Distributions

- Suppose $Y = 1$ is a success where the probability of a success is $P(Y = 1) = \pi$, $Y = 0$ is a failure.

- Goal: Estimate π.

Bernoulli probability mass function:

$$P(Y = y) = \pi^{y}(1-\pi)^{1-y} \quad \text{for } y = 0 \text{ or } 1$$

Notice that $P(Y = 1) = \pi$ and $P(Y = 0) = 1 - \pi$

Often, you observe multiple success/failure observations. Let Y_1, …, Y_n denote random variables for these observations. If the random variables are independent and have the same probability of success π, then we can use a binomial PMF for $W = \sum_{i=1}^{n} Y_i$.

Binomial Probability Distributions

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

for w = 0, 1, …, n

Notes:

- $\frac{n!}{w!(n-w)!} = \binom{n}{w}$ = n choose w
- W is a random variable denoting the number of "successes" out of n trials
- W has a fixed number of possibilities – 0, 1, …, n
- n is a fixed constant
- π is a parameter denoting the probability of a "success" with values between 0 and 1.

Required Conditions When Applying Binomial Probability Model

1. There are n identical trials.
2. Each trial has two possible outcomes, typically referred to as a success or failure.
3. The trials are independent of each other.
4. The probability of success, denoted by π, remains constant for each trial. The probability of a failure is $1-\pi$.
5. The random variable, W, represents the number of successes.

We will use two running (toy) examples to illustrate the concepts and techniques in this lecture.

As a reminder, please read the required reading before attending the live sessions.

An Example: Field Goal Kicking

Suppose a field goal kicker attempts five field goals during a game and each field goal has the same probability of being successful (the kick is made). Also, assume each field goal is attempted under similar conditions; i.e., distance, weather, surface,….

1. n identical trials: n = 5 field goals attempted under identical conditions.

2. Two possible outcomes of a trial: Each field goal can be made (success) or missed (failure).

3. The trials are independent of each other: The result of one field goal does not affect the result of another field goal.

An Example: Field Goal Kicking (cont.)

4. The probability of success, denoted by π, remains constant for each trial. The probability of a failure is 1-π: Suppose the probability a field goal is good is 0.6; i.e., P(success) = π = 0.6.

5. The random variable, W, represents the number of successes. Let W = number of field goals that are good. Thus, W can be 0, 1, 2, 3, 4, or 5. Because these five items are satisfied, the binomial probability mass function can be used, and W is called a binomial random variable.

Mean and Variance of Binomial Probability Distributions

$$E(W) = n\pi$$

$$Var(W) = n\pi(1-\pi)$$

Computing Probabilities of Binomial Probability Model

Field Goal Kicking (cont.)

Suppose $\pi = 0.6$, $n = 5$. What are the probabilities for each possible value of w?

Note: It is important that we write down the formula so that we know what we are computing before using a computer program or statistical package for the computation.

$$P(W=0) = \frac{n!}{w!(n-w)!}\pi^{w}(1-\pi)^{n-w}$$

$$= \frac{5!}{0!(5-0)!}0.6^{0}(1-0.6)^{5-0} = 0.4^{5} \approx 0.0102$$

Field Goal Kicking (cont.)

For W=0,…,5, we obtain

W	P(W = w)
0	0.0102
1	0.0768
2	0.2304
3	0.3456
4	0.2592
5	0.0778

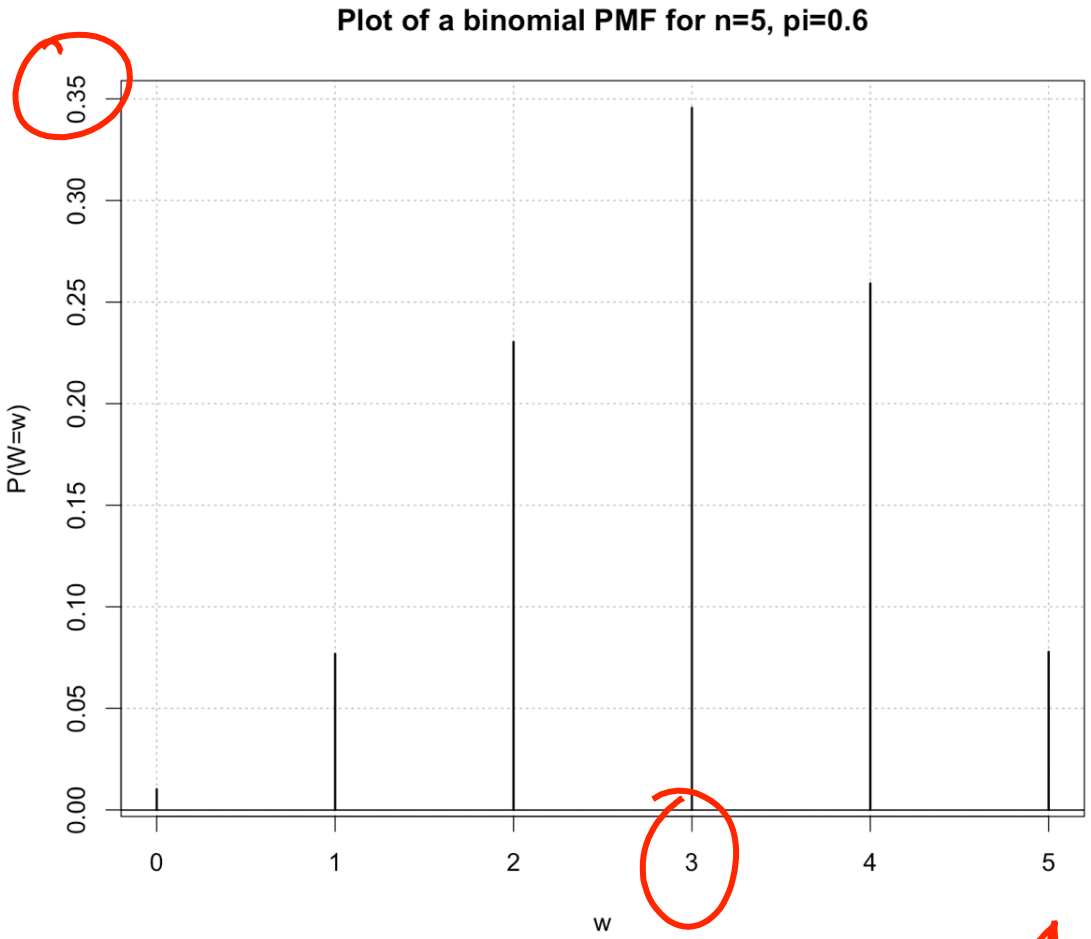
$E(W) = n\pi = 5*0.6 = 3$ and

$Var(W) = n\pi(1-\pi) = 5*0.6*(1 - 0.6) = 1.2$

Implementation in R:

```
> dbinom(x = 1, size = 5, prob = 0.6)
[1] 0.0768
> dbinom(x = 0:5, size = 5, prob = 0.6)
[1] 0.01024 0.07680 0.23040 0.34560 0.25920 0.07776
> pmf<-dbinom(x = 0:5, size = 5, prob = 0.6)
> pmf.df<-data.frame(w = 0:5, prob = round(x = pmf, digits = 4))
> pmf.df
  w   prob
1 0 0.0102
2 1 0.0768
3 2 0.2304
4 3 0.3456
5 4 0.2592
6 5 0.0778
```

Visualize the Results in R:




```
plot(x = pmf.df$w, y = pmf.df$prob, type = "h", xlab = "w",
   ylab = "P(W=w)", main = "Plot of a binomial PMF for n=5, pi=0.6",
   panel.first = grid(col="gray", lty="dotted"),
   lwd = 2)
abline(h = 0)
```

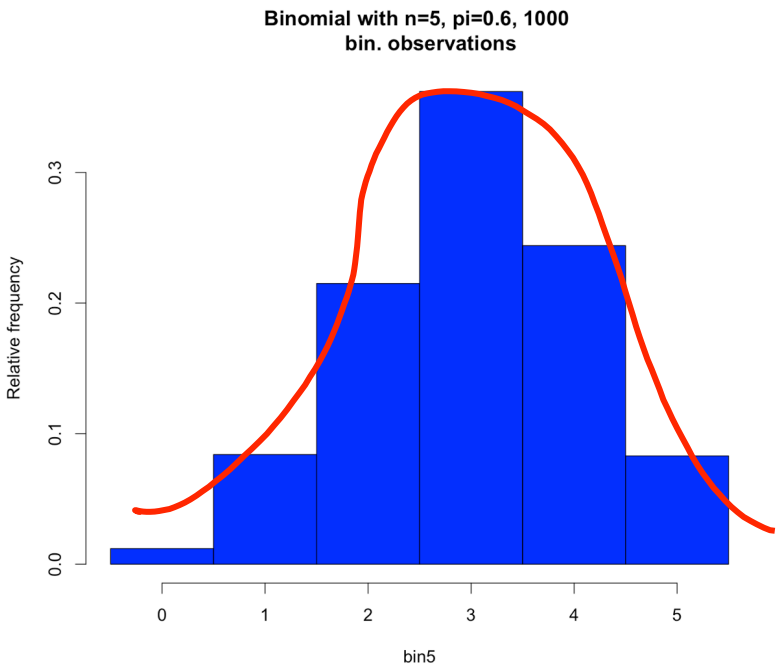
Simulating a Binomial Probability Model

Simulation: Binomial Probability Model:

- The purpose of this example is to show how one can "simulate" observing a random sample of observations from a population characterized by a binomial distribution.
- Why would someone want to do this?
  - It is easier to visualize an abstract, mathematical probability model.
  - It allows to examine the change of characteristics of the distribution as the defining parameters (i.e., pi and n, in the case of the binomial probability model) change.
- All the standard/common probability models are available in R.
- ➔ Talk about distribution, probability, and so on, in R here.

Simulation: Binomial Probability Model

```
> set.seed(4848)
> bin5<-rbinom(n = 1000, size = 5, prob = 0.6)
> bin5[1:20]
 [1] 3 2 4 1 3 1 3 3 3 4 3 3 3 2 3 1 2 2 5 2
> mean(bin5)
[1] 2.991
> var(bin5)
[1] 1.236155
> table(x = bin5)
x
  0   1   2   3   4   5
 12  84 215 362 244  83
```




```
hist(x = bin5, main = "Binomial with n=5, pi=0.6, 1000
    bin. observations", col="blue", probability = TRUE, breaks = -
    0.5:5.5, ylab = "Relative frequency")
```

- The shape of the histogram looks similar to the shape of the actual binomial distribution.
- The mean and variance are close to what we expect them to be!

Maximum Likelihood Estimation (1)

Maximum Likelihood Estimation

Suppose the success or failure of a field goal in football can be modeled with a Bernoulli(π) distribution. Let Y = 0 if the field goal is a failure and Y = 1 if the field goal is a success. Then the probability distribution for Y is:

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

where π denotes the probability of success.

Suppose we would like to estimate π for a 40-yard field goal. Let $y_1, \ldots, y_n$ denote a random sample of observed field goal results at 40 yards. Thus, these y_is are either 0s or 1s. Given the resulting data $(y_1, \ldots, y_n)$, the "likelihood function" measures the plausibility of different values of π:

Maximum Likelihood Estimation

Suppose we would like to estimate π for a 40-yard field goal. Let $y_1, \ldots, y_n$ denote a random sample of observed field goal results at 40 yards. Thus, these y_is are either 0s or 1s.

Given the resulting data $(y_1, \ldots, y_n)$, the "likelihood function" measures the plausibility of different values of π:

$$L(\pi \mid y_1, \ldots, y_n)$$

$$= P(Y_1 = y_1) * P(Y_2 = y_2) * \quad * P(Y_n = y_n)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i)$$

$$= \prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{1-y_i}$$

$$= \pi^{\sum_{i=1}^{n} y_i} (1-\pi)^{n-\sum_{i=1}^{n} y_i}$$

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

Maximum Likelihood Estimation (1)

Maximum Likelihood Estimation

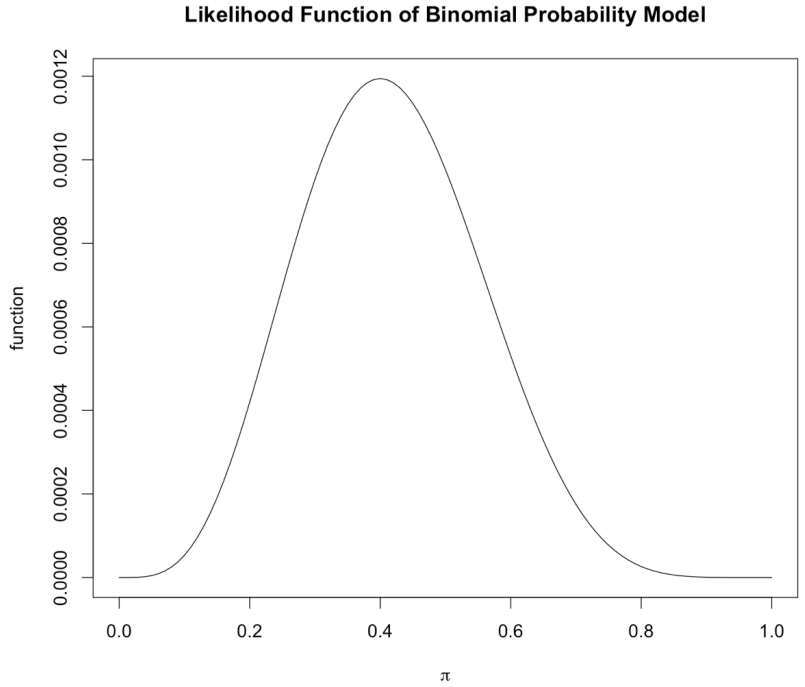
Suppose w = 4 and n = 10. Given this observed information, we would like to find the corresponding parameter value for π that produces the largest probability of obtaining this particular sample.

The following table can be formed to help find this parameter value:

π	$L(\pi \mid y_1,\ldots,y_n)$
0.2	0.000419
0.3	0.000953
0.35	0.001132
0.39	0.001192
0.4	0.001194
0.41	0.001192
0.5	0.000977

Maximum Likelihood Estimation (in R)

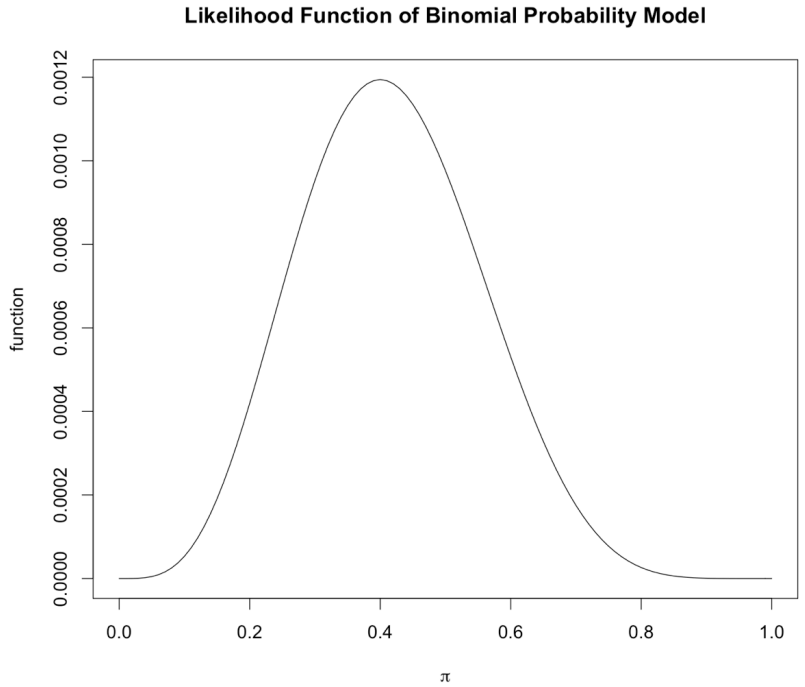
```
> sum.y<-4
> n<-10
> # Try different values of pi
> pi<-c(0.2, 0.3, 0.35, 0.39, 0.4, 0.41, 0.5)
> Lik<-pi^sum.y*(1-pi)^(n-sum.y)
> data.frame(pi, Lik)
    pi          Lik
1 0.20 0.0004194304
2 0.30 0.0009529569
3 0.35 0.0011317547
4 0.39 0.0011918935
5 0.40 0.0011943936
6 0.41 0.0011919211
7 0.50 0.0009765625
```




```
#Likelihood function plot
curve(expr = x^sum.y*(1-x)^(n-sum.y), xlim = c(0,1),
    xlab = expression(pi), ylab = "Likelihood
    function", main="Likelihood Function of Binomial Probability Model")
```

Maximum Likelihood Estimation (in R)

```
> sum.y<-4
> n<-10
> # Try different values of pi
> pi<-c(0.2, 0.3, 0.35, 0.39, 0.4, 0.41, 0.5)
> Lik<-pi^sum.y*(1-pi)^(n-sum.y)
> data.frame(pi, Lik)
    pi          Lik
1 0.20 0.0004194304
2 0.30 0.0009529569
3 0.35 0.0011317547
4 0.39 0.0011918935
5 0.40 0.0011943936
6 0.41 0.0011919211
7 0.50 0.0009765625
```




```
#Likelihood function plot
curve(expr = x^sum.y*(1-x)^(n-sum.y), xlim = c(0,1),
    xlab = expression(pi), ylab = "Likelihood
    function", main="Likelihood Function of Binomial Probability Model")
```

Note that $\pi = 0.4$ is the "most plausible" value of π for the observed data because this maximizes the likelihood function. Therefore, 0.4 is the maximum likelihood estimate (MLE).

Maximum Likelihood Estimation (2)

Finding MLE in General

In general, the MLE can be found as follows:

1. Find the natural log of the likelihood function, $\log[L(\pi \mid y_1,\ldots,y_n)]$
2. Take the derivative of $\log[L(\pi \mid y_1,\ldots,y_n)]$ with respect to π.
3. Set the derivative equal to 0 and solve for π to find the maximum likelihood estimate. Note that the solution is the maximum of $L(\pi \mid y_1,\ldots,y_n)$ provided certain "regularity" conditions hold (see Mood, Graybill, Boes, 1974).

For the field goal example:

$$\log[L(\pi \mid y_1,\ldots,y_n)] = \log\left[\pi^{\sum_{i=1}^n y_i}(1-\pi)^{n-\sum_{i=1}^n y_i}\right]$$

$$= \sum_{i=1}^n y_i \log(\pi) + (n - \sum_{i=1}^n y_i)\log(1-\pi)$$

where log means natural log.

$$\frac{\partial \log[L(\pi \mid y_1,\ldots,y_n)]}{\partial \pi} = \frac{\sum_{i=1}^n y_i}{\pi} - \frac{n - \sum_{i=1}^n y_i}{1-\pi} = 0$$

Finding MLE in General

$$\Rightarrow \frac{\sum_{i=1}^{n} y_i}{\pi} = \frac{n - \sum_{i=1}^{n} y_i}{1-\pi}$$

$$\Leftrightarrow \frac{1-\pi}{\pi} = \frac{n - \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i}$$

$$\Leftrightarrow \frac{1}{\pi} = \frac{n - \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i}$$

$$\Rightarrow \pi = \frac{\sum_{i=1}^{n} y_i}{n}$$

Therefore, the maximum likelihood estimator of π is the proportion of field goals made. To avoid confusion between a parameter and a statistic, we will denote the estimator as $\hat{\pi} = \sum_{i=1}^{n} y_i / n$.

Again, Maximum likelihood estimation will prove to be extremely important in this class, as it is used in pretty much all of the statistical models we will study.

Properties of MLE

Why are we interested in MLE?

It comes with many desirable statistical properties.

$\hat{\pi}$ will vary from sample to sample. We can mathematically quantify this variation for maximum likelihood estimators in general as follows:

- Asymptotic normality: For a large sample, maximum likelihood estimators can be treated as normal random variables.
- For a large sample, the variance of the maximum likelihood estimator can be computed from the second derivative of the log likelihood function.

Properties of MLE

Thus, in general for a maximum likelihood estimator $\hat{\theta}$ for θ, we can say that

$$\hat{\theta} \sim N\left(\theta, Var(\hat{\theta})\right)$$

for a large sample $Y_1, \ldots, Y_n$, where

$$Var(\hat{\theta}) = -\left[E\left(\frac{\partial^2 \log\left[L(\theta \mid Y_1, \ldots, Y_n)\right]}{\partial \theta^2}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}}$$

The use of "for a large sample" can also be replaced with the word "asymptotically." You will often hear these results talked about using the phrase "asymptotic normality of maximum likelihood estimators."

Properties of MLE

Thus, in general for a maximum likelihood estimator $\hat{\theta}$ for θ, we can say that

$$\hat{\theta} \sim N\left(\theta, Var(\hat{\theta})\right)$$

for a large sample $Y_1, \ldots, Y_n$, where

$$Var(\hat{\theta}) = -\left[E\left(\frac{\partial^2 \log\left[L(\theta \mid Y_1, \ldots, Y_n)\right]}{\partial \theta^2}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}}$$

The use of "for a large sample" can also be replaced with the word "asymptotically." You will often hear these results talked about using the phrase "asymptotic normality of maximum likelihood estimators."

Wald Confidence Interval

Wald Confidence Interval

Because $\hat{\theta} \sim N\left(\theta, Var(\hat{\theta})\right)$, we can rewrite this as a standardized statistic:

$$\frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \sim N(0,1)$$

This logic is covered many times at w203. Please make sure you understand this logic before proceeding to the next slide.

Concept Check (1 minute):

This logic is covered in w203. Make sure you understand the logic used here before proceeding to the next slide.

Because $\hat{\theta} \sim N\left(\theta, Var(\hat{\theta})\right)$, we can rewrite this as a standardized statistic:

$$\frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \sim N(0,1)$$

Wald Confidence Interval

Also, because we have a probability distribution here, we can quantify with a level of certainty that observed values of the statistic are within a particular range:

$$P\left(Z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} < Z_{1-\alpha/2} \right) \approx 1 - \alpha$$

where $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile from a standard normal. For example, if $\alpha = 0.05$, we have $Z_{0.975} = 1.96$.

Wald Confidence Interval

```
> qnorm(p = 1-0.05/2, mean = 0, sd = 1)
[1] 1.959964
```

Note that I specially chose $Z_{\alpha/2}$ and $Z_{1-\alpha/2}$ for symmetry. Of course, $Z_{\alpha/2} = -Z_{1-\alpha/2}$.

If we rearrange items within the P(·), we obtain

$$P\left(\hat{\theta} - Z_{1-\alpha/2}\sqrt{Var(\hat{\theta})} < \theta < \hat{\theta} + Z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}\right) \approx 1-\alpha$$

Thus, if α is chosen to be small, we are fairly certain the expression within P(·) will hold true. When we substitute the observed values of $\hat{\theta}$ and $Var(\hat{\theta})$ into the expression, we obtain the $(1-\alpha)100\%$ "Wald" confidence interval for θ as

$$\hat{\theta} - Z_{1-\alpha/2}\sqrt{Var(\hat{\theta})} < \theta < \hat{\theta} + Z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}$$

Notice this interval follows the typical form of a confidence interval for a parameter:

Estimator ± (distributional value)*(standard deviation of estimator)

Wald Confidence Interval

```
> qnorm(p = 1-0.05/2, mean = 0, sd = 1)
[1] 1.959964
```

Because $\hat{\pi}$ is a maximum likelihood estimator, we can use a Wald confidence interval for π:

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

Because $\hat{\pi}$ is close to 0 or 1, two problems may occur:

1) Calculated limits may be less than 0 or greater than 1, which is outside the boundaries for a probability.
2) When $\hat{\pi}$ = 0 or 1, $\hat{\pi}(1-\hat{\pi})/n$ = 0 for n > 0. This leads to the lower and upper limits to be exactly the same (0 for $\hat{\pi}$ = 0 or 1 for $\hat{\pi}$ = 1).

Example: Field Goal Kicking

Suppose $\sum_{i=1}^{n} y_i = w = 4$ and $n = 10$. The 95% confidence interval is

$$\hat{\pi} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.4 \pm 1.96\sqrt{\frac{0.4(1-0.4)}{10}}$$

$$0.0964 < \pi < 0.7036$$

Below is the implementation in R:

```
w<-4
n<-10
alpha<-0.05
pi.hat<-w/n

var.wald<-pi.hat*(1-pi.hat)/n
lower<-pi.hat - qnorm(p = 1-alpha/2) * sqrt(var.wald)
upper<-pi.hat + qnorm(p = 1-alpha/2) * sqrt(var.wald)
round(data.frame(lower, upper), 4)

#Quicker
round(pi.hat + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald), 4)
```

```
> round(pi.hat + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald), 4)
[1] 0.0964 0.7036
```

Interpretation of the CI

There are two problems:

1. The interval "works" if the sample size is large. The field goal kicking example has n = 10 only!
2. The discreteness of the binomial distribution often makes the normal approximation work poorly even with large samples.

- The result is a confidence interval that is often too "liberal." This means when 95% is stated as the confidence level, the true confidence level is often lower.
- There are "conservative" intervals. These intervals have a true confidence level larger than the stated level.
- The limitations of this particular confidence interval have been discussed for a long time in the statistical literature. There have been many alternative confidence intervals for π proposed.

Alternative Confidence Intervals and True Confidence Level

Alternatives (in Practice)

For n < 40, use the Wilson or Jeffrey's prior interval. Below is Wilson's interval. Jeffrey's prior interval is a Bayesian-based CI.

$$\pi \pm \frac{Z_{1-\alpha/2} n^{1/2}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + \frac{Z^2_{1-\alpha/2}}{4n}}$$

where

$$\pi = \frac{w + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}}$$

Alternatives (in Practice)

For n ≥ 40, use the Agresti-Coull (Agresti and Coull, 1998) interval

The (1-α)100% confidence interval is

$$\pi \pm Z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n+Z^2_{1-\alpha/2}}}$$

This is essentially a Wald interval where we add $Z^2_{1-\alpha/2}/2$ successes and $Z^2_{1-\alpha/2}/2$ failures to the observed data. In fact, when $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96 \approx 2$. Then

$$\pi = \frac{w+2^2/2}{n+2^2} = \frac{w+2}{n+4}$$

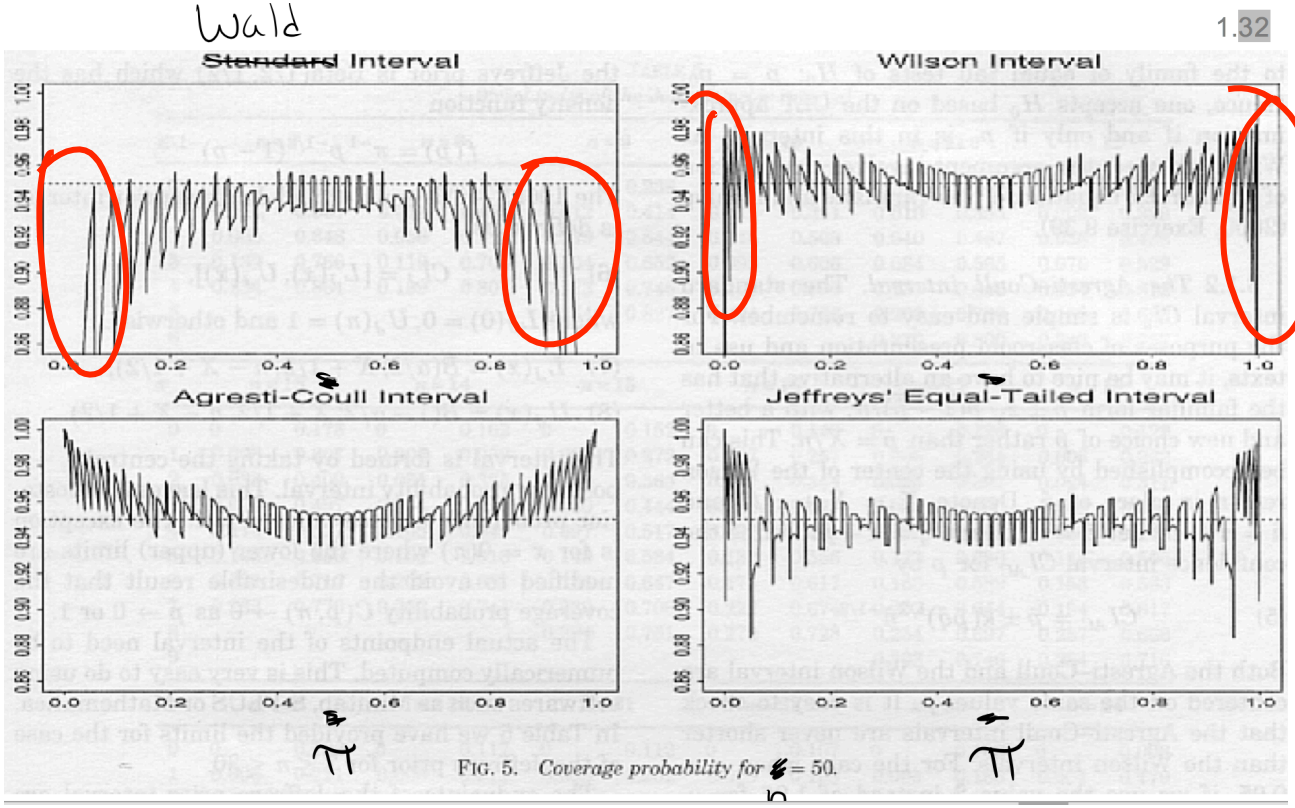
Thus, two successes and two failures are added. Also, notice how

$$\pi = \frac{w+Z^2_{1-\alpha/2}/2}{n+Z^2_{1-\alpha/2}}$$

can be thought of as an adjusted estimate of π. For values of w close to 0, $\pi > \hat{\pi}$ For values of w close to n, $\pi < \hat{\pi}$.

True Confidence Levels for Confidence Intervals

Below is a comparison of the performance of the four confidence intervals. The values on the y-axis represent the true confidence level (coverage) of the confidence intervals. Each of the confidence intervals are supposed to be 95%!



What Does True Confidence/Coverage Level Mean?

- Suppose a random sample of size n = 50 is taken from a population and a 95% Wald confidence interval is calculated.
- Suppose another random sample of size n = 50 is taken from the same population and a 95% Wald confidence interval is calculated.
- Repeat this process 10,000 times.
- We would expect 9,500 out of 10,000 (95%) confidence intervals to contain π.
- Unfortunately, this does not often happen. **It is guaranteed to happen only when n = ∞ for the Wald interval.**
- The true confidence or coverage level is the percent of times the confidence intervals contain or "cover" π.
- The plots show many possible values of π (0.0005 to 0.9995 by 0.0005). For example, the true confidence level using the Wald interval is approximately 0.90 for $\pi = 0.184$.

Calculate the True Confidence or Coverage Level in R

```
pi.hat<-w/n
pi.hat[1:10]
var.wald<-pi.hat*(1-pi.hat)/n
lower<-pi.hat - qnorm(p = 1-alpha/2) * sqrt(var.wald)
upper<-pi.hat + qnorm(p = 1-alpha/2) * sqrt(var.wald)
data.frame(w, pi.hat, lower, upper)[1:10,]
save<-ifelse(test = pi>lower, yes = ifelse(test =
    pi<upper, yes = 1, no = 0), no = 0)
save[1:10]
mean(save)
```

```
An estimate of the true confidence level is: 0.898
```

In this example, an estimate of the true confidence level is only 0.898 (and not 0.95)!

```
> data.frame(w, pi.hat, lower, upper)[1:20,]
    w pi.hat      lower     upper
1   9   0.18 0.07351063 0.2864894
2   9   0.18 0.07351063 0.2864894
3  10   0.20 0.08912769 0.3108723
4  11   0.22 0.10517889 0.3348211
5  12   0.24 0.12162077 0.3583792
6   9   0.18 0.07351063 0.2864894
7  11   0.22 0.10517889 0.3348211
8   5   0.10 0.01684577 0.1831542
9   8   0.16 0.05838385 0.2616161
10  6   0.12 0.02992691 0.2100731
11 16   0.32 0.19070178 0.4492982
12  8   0.16 0.05838385 0.2616161
13 11   0.22 0.10517889 0.3348211
14 10   0.20 0.08912769 0.3108723
15 15   0.30 0.17297982 0.4270202
16  5   0.10 0.01684577 0.1831542
17  7   0.14 0.04382187 0.2361781
18  8   0.16 0.05838385 0.2616161
19 11   0.22 0.10517889 0.3348211
20 10   0.20 0.08912769 0.3108723
```

Hypothesis Tests for π

Hypothesis Tests for π

Hypothesis:

$H_0: \pi = \pi_0$ vs. $H_a: \pi \neq \pi_0$

Test statistic: $Z_0 = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$

Reject H_0 if $|Z_0| > Z_{1-\alpha/2}$

This type of test is a type of "score test."

- Another way to perform a hypothesis test of $H_0: \pi = \pi_0$ vs. $H_a: \pi \neq \pi_0$ is a likelihood ratio test (LRT), which is very frequently used in the analysis of categorical data (beyond testing for π).

Hypothesis Tests for π (cont.)

The LRT statistic, Λ, is the ratio of two likelihood functions. The numerator is the likelihood function maximized over the parameter space restricted under the null hypothesis. The denominator is the likelihood function maximized over the unrestricted parameter space.

$$\Lambda = \frac{\text{Max. lik. when parameters satisfy } H_0}{\text{Max. lik. when parameters satisfy } H_0 \text{ or } H_a}$$

Wilks (1935, 1938) shows that $-2\log(\Lambda)$ can be approximated by a χ^2_u for a large sample and under H_o where u is the difference in dimension between the alternative and null hypothesis parameter spaces.

Example: Field Goal Kicking (cont.)

Suppose the hypothesis test $H_0: \pi = 0.5$ vs. $H_a: \pi \neq 0.5$ is of interest. Remember that $w = 4$ and $n = 10$.

The numerator of Λ is the maximum possible value of the likelihood function under the null hypothesis. Because $\pi = 0.5$ is the null hypothesis, the maximum can be found by just substituting $\pi = 0.5$ in the likelihood function.

$$L(\pi = 0.5 \mid y_1, \ldots, y_n) = 0.5^w (1 - 0.5)^{n-w}$$

Then

$$L(\pi = 0.5 \mid y_1, \ldots, y_n) = 0.5^4 (0.5)^{10-4} = 0.0009766$$

$$\Lambda = \frac{\text{Max. lik. when parameters satisfy } H_0}{\text{Max. lik. when parameters satisfy } H_0 \text{ or } H_a}$$

$$= \frac{0.0009766}{0.001194} = 0.8179$$

Then $-2\log(\Lambda) = -2\log(0.8179) = 0.4020$ is the test statistic value. The critical value is…

Example: Field Goal Kicking (cont.)

Then $-2\log(\Lambda) = -2\log(0.8179) = 0.4020$ is the test statistic value.

The critical value is $\chi^2_{1,0.95} = 3.84$ using $\alpha = 0.05$

```
> qchisq(p = 0.95, df = 1)
[1] 3.841459
```

Therefore, there is not sufficient evidence to reject the hypothesis that $\pi = 0.5$.

Introduction to Multiple Binary Variables

Introduction to Contingency Table

This section extends the methods from Section 1.1 to a heterogeneous setting where individual items come from one of two groups. Because there is still a binary response, we can summarize the sample in a 2×2 contingency table.

Free throws are typically shot in pairs. Below is a contingency table summarizing Larry Bird's first and second free throw attempts during the 1980–1981 and 1981–1982 NBA seasons.

		Second		
		Made	Missed	Total
First	Made	251	34	285
	Missed	48	5	53
	Total	299	39	338

Contingency Table

		Second		
		Made	Missed	Total
First	Made	251	34	285
	Missed	48	5	53
	Total	299	39	338

- 251 first and second free throw attempts were both made.
- 34 first free throw attempts were made, and the second were missed.
- 48 first throw attempts were missed, and the second free throw were made.
- 5 first and second free throw attempts were both missed,
- 285 first free throws were made regardless what happened on the second attempt,
- 299 second free throws were made regardless what happened on the first attempt,
- 338 free throw pairs were shot during these seasons,

Example 1

What types of questions would be of interest for this data?

In the clinical trials for the polio vaccine developed by Jonas Salk, two large groups were involved in the placebo-control phase of the study. The first group, which received the vaccination, consisted of 200,745 individuals. The second group, which received a placebo, consisted of 201,229 individuals. There were 57 cases of polio in the first group and 142 cases of polio in the second group.

	Polio	Polio free	Total
Vaccine	57	200,688	200,745
Placebo	142	201,087	201,229
Total	199	401,775	401,974

Polio vaccine clinical trials (data source: Francis et al., *American Journal of Public Health*, 1955)

Example 2

Clinical trials have been performed to evaluate the effectiveness of a number of HIV vaccines. The results of one trial in particular were discussed a lot in the media in 2009. Below is a contingency table summarizing the data used in the "modified intent-to-treat analysis" of the paper.

	HIV	HIV free	
Vaccine	51	8,146	8,197
Placebo	74	8,124	8,198
	125	16,270	16,395

(Vaccine, HIV)

Formulation of Contingency Table and Confidence Interval of Two Binary Variables

Notations and Model

- Let $Y_{11}, \ldots, Y_{n_1 1}$ be Bernoulli random variables for group 1 (row 1 of the contingency table).
- Let $Y_{12}, \ldots, Y_{n_1 2}$ be Bernoulli random variables for group 2 (row 2 of the contingency table).
- The number of "successes" for a group is represented by $W_j = \sum_{i=1}^{n_j} Y_{ij}$.
- W_j has a binomial distribution with success probability π_j and number of trials of n_j
- W_1 is independent of W_2; thus, we have an "independent binomial model". Some people refer to this as "independent binomial sampling" as a way to describe how the contingency table counts come about.
- The MLE of π_j is $\hat{\pi}_j = w_j / n_j$
- A "+" in a subscript is used to denote indices in a subscript that are being summed over. For example, $W_+ = W_1 + W_2$ is the total number of successes and $n_+ = n_1 + n_2$ is the total sample size. In fact, $W_j = Y_{+j}$.

		Response 1	Response 2	
Group	1	w_1	$n_1 - w_1$	n_1
	2	w_2	$n_2 - w_2$	n_2
		w_+	$n_+ - w_+$	n_+

		Response 1	Response 2	
Group	1	π_1	$1 - \pi_1$	1
	2	π_2	$1 - \pi_2$	1

Example: Larry Bird's Free Throws

```
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
```

```
rowSums(c.table) #n1 and n2

pi.hat.table<-c.table/rowSums(c.table)
```

```
> pi.hat.table
        Second
First         made     missed
  made   0.8807018 0.11929825
  missed 0.9056604 0.09433962
```

The estimated probability that Larry Bird makes his second free throw attempt is $\hat{\pi}_1 = 0.8807$, given that he makes the first, and $\hat{\pi}_2 = 0.9057$, given he misses the first.

Confidence Intervals for the Difference of Two Probabilities

Remember from Section 1.1 that the estimated probability of success $\hat{\pi}$ can be treated as an approximate normal random variable with mean π and variance $\pi(1-\pi)/n$ for a large sample. Using the notation in this week, this means that

$$\hat{\pi}_1 \sim N(\pi_1, \pi_1(1-\pi_1)/n_1) \text{ and}$$
$$\hat{\pi}_2 \sim N(\pi_2, \pi_2(1-\pi_2)/n_2)$$

for large n_1 and n_2.

Note: $\text{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \text{Var}(\hat{\pi}_1) + \text{Var}(\hat{\pi}_2)$ because $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent random variables. Some of you may have seen the following: Let X and Y be independent random variables and let a and b be constants. Then $\text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$.

The estimate of the variance is then

$$\text{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}$$

A $(1-\alpha)100\%$ Wald confidence interval for $\pi_1 - \pi_2$ is

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

Agresti and Caffo Adjustment to CI

Let $\pi_1 = \frac{w_1 + 1}{n_1 + 2}$ and $\pi_2 = \frac{w_2 + 1}{n_2 + 2}$

The Agresti-Caffo confidence interval is

$$\pi_1 - \pi_2 \pm Z_{1-\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_1 + 2} + \frac{\pi_2(1-\pi_2)}{n_2 + 2}}$$

Example: Larry Bird's Free Throws

```{r}
alpha<-0.05
pi.hat1<-pi.hat.table[1,1]
pi.hat2<-pi.hat.table[2,1]

#Wald
var.wald<-pi.hat1*(1-pi.hat1) / sum(c.table[1,]) + pi.hat2*(1-pi.hat2) /
sum(c.table[2,])

pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald)
```

```
-0.11218742  0.06227017
```

```
#Agresti-Caffo
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)
pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.AC)
```

```
-0.10353254  0.07781192
```

Therefore, the 95% Wald confidence interval is

$$-0.1122 < \pi_1 - \pi_2 < 0.0623$$

and the 95% Agresti-Caffo confidence interval is

$$-0.1035 < \pi_1 - \pi_2 < 0.0778$$

Testing the Difference of Two Probabilities

Hypothesis test of $H_0: \pi_1 - \pi_2 = 0$ vs. $H_a: : \pi_1 - \pi_2 \neq 0$

Test Statistic:

$$Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{\pi} = w_+ / n_+$

This test statistic has a standard normal distribution for a large sample. Therefore, you can reject H_0 if $|Z_0| > Z_{1-\alpha/2}$

Relative Risk

Why Is the Notion of Relative Risk Important?

A shortcoming of basing inference on $\pi_1 - \pi_2$ is that it measures a quantity whose meaning changes depending on the sizes of $\pi_1 - \pi_2$.

	Adverse reactions		
	Yes	No	Total
Drug	$\pi_1 = 0.510$	$1 - \pi_1 = 0.490$	1
Placebo	$\pi_2 = 0.501$	$1 - \pi_2 = 0.499$	1

$\pi_1 - \pi_2 = 0.510 - 0.501 = 0.009$

	Adverse reactions		
	Yes	No	Total
Drug	$\pi_1 = 0.010$	$1 - \pi_1 = 0.990$	1
Placebo	$\pi_2 = 0.001$	$1 - \pi_2 = 0.999$	1

$\pi_1 - \pi_2 = 0.010 - 0.001 = 0.009$

In the first scenario, an increase of 0.009 is rather small **relative** to the already sizable probabilities given for the two groups. On the other hand, the second scenario has a much larger adverse reaction probability for the drug group **relative** to the placebo group.

Defining Relative Risk

Convey the relative magnitudes of these changes better than differences allow.

$$RR = \frac{\pi_1}{\pi_2}$$

For scenario 2 above, the relative risk is RR = 0.010/0.001 = 10.

Interpretation:
An adverse reaction is **10** *times as likely* for those individuals taking the drug than those individuals taking the placebo, or the probability of an adverse reaction is 10 *times as large* for those individuals taking the drug than those individuals taking the placebo.

An adverse reaction is **nine** *times more likely* for individuals taking the drug than those individuals taking the placebo, or the probability of an adverse reaction is nine *times larger* for individuals taking the drug than those individuals taking the placebo.

Interpretation

- "One times as likely" is equivalent to $\pi_1/\pi_2 = 1$. In other words, they are equal. "Two times as likely" is equivalent to $\pi_1/\pi_2 = 2$. In other words, π_1 is twice the size of π_2; $\pi_1 = 2 \times \pi_2$.

- "Two times more likely" is equivalent to $\pi_1/\pi_2 = 3$. The "more" is what causes the difference from the previous interpretation.

- As another example, $\pi_1/\pi_2 = 1.5$ means that a success is 50% more likely for Group 1 than for Group 2. Alternatively, a success is 1.5 times as likely for Group 1 than for Group 2.

Relative Risk

MLE

The MLE of RR can be found by substituting MLEs of π_1 and π_2 into the equation for RR:

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{w_1 / n_1}{w_2 / n_2} = \frac{w_1 n_2}{w_2 n_1}$$

A $(1 - \alpha)100\%$ Wald confidence interval for log(RR) is

$$\log(\hat{\pi}_1 / \hat{\pi}_2) \pm Z_{1-\alpha/2} \sqrt{Var(\log(\hat{\pi}_1 / \hat{\pi}_2))}$$

where

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

$$= \frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}$$

MLE

$(1 - \alpha)100\%$ Wald confidence interval for RR is then

$$\exp\left[\log(\hat{\pi}_1 / \hat{\pi}_2) \pm Z_{1-\alpha/2}\sqrt{\text{Var}(\log(\hat{\pi}_1 / \hat{\pi}_2))}\right]$$

Odd Ratios

Notion of Odds

Odds are the probability of success divided by the probability of a failure.

		Response		
		1 = success	2 = failure	
Group	1	π_1	$1 - \pi_1$	1
	2	π_2	$1 - \pi_2$	1

- For Row 1, the "odds of a success" are $\text{odds}_1 = \pi_1/(1-\pi_1)$.
- For Row 2, the "odds of a success" are $\text{odds}_2 = \pi_2/(1-\pi_2)$.

Odds are a rescaling of the probability of success.

- If P(success) = 0.75, then the odds are 3 or "3 to 1 odds," that is, the probability of a success are three times as large as the probability of a failure.

Defining Odd-Ratio

$$\text{odds}_1 = \frac{\hat{\pi}_1}{1-\hat{\pi}_1} = \frac{w_1/n_1}{1-w_1/n_1} = \frac{w_1}{n_1-w_1}$$

$$\text{odds}_2 = \frac{\hat{\pi}_2}{1-\hat{\pi}_2} = \frac{w_2/n_2}{1-w_2/n_2} = \frac{w_2}{n_2-w_2}$$

		Response	
		1 = success	2 = failure
Group	1	w_1	$n_1 - w_1$
	2	w_2	$n_2 - w_2$

Odd Ratios

Defining Odd-Ratio

Odds ratio incorporate information from both Rows 1 and 2 into a single number:

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Odds ratios are VERY useful in categorical data analysis and will be used throughout this course.

Odd Ratios

MLE

$$OR = \frac{odds_1}{odds_2} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)}$$

$$= \frac{w_1/n_1(1-w_2/n_2)}{w_2/n_2(1-w_1/n_1)} = \frac{w_1(n_2-w_2)}{w_2(n_1-w_1)}$$

The estimate is a product of the counts on the "diagonal" (top left to bottom right) of the contingency table divided by a product of the counts on the off diagonal.

Interpretation

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

- Remember that Odd ratios is the ratio of two odds, comparing the odds of success relative to the odds of failure.
- The estimated odds of a success are OR times as large as in Group 1 than in Group 2.
- The estimated odds of a success are 1/OR times as large as in Group 2 than in Group 1.

Interpretation

$$\frac{(1-\hat{\pi}_1)/\hat{\pi}_1}{(1-\hat{\pi}_2)/\hat{\pi}_2} = \frac{\hat{\pi}_2(1-\hat{\pi}_1)}{\hat{\pi}_1(1-\hat{\pi}_2)}$$

- Consider the odds of a failure $(1-\pi_1)/\pi_1$, so the ratio of Group 1 to Group 2 becomes.

- The estimated odds of a failure are 1/OR times as large as in Group 1 than in Group 2 and OR times as large as in Group 2 than in Group 1.

MLE

- Because OR is a maximum likelihood estimate, we can use the "usual" properties of them to find the confidence interval.
- However, using the log(OR) often works better (i.e., its distribution is closer to being a normal distribution).

$$\text{Var}(\log(OR)) = \frac{1}{w_1} + \frac{1}{n_1 - w_1} + \frac{1}{w_2} + \frac{1}{n_2 - w_2}.$$

The $(1 - \alpha)100\%$ Wald confidence interval for log(OR) is

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

The $(1 - \alpha)100\%$ Wald confidence interval for OR is

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

Example: Larry Bird's Free Throw

		Second		
		Made	Missed	Total
First	Made	251	34	285
	Missed	48	5	53
	Total	299	39	338

$$OR = \frac{w_1(n_2 - w_2)}{w_2(n_1 - w_1)} = \frac{251 * 5}{48 * 34} = 0.7690.$$

Interpretation:

- The estimated odds of a made second free throw are **0.7690 times as large** when the first free throw is made than when the first free throw is missed.
- The estimated odds of a made second free throw are **1/0.7690 = 1.3 times as large** when the first free throw is missed than when the first free throw is made.

Example: Larry Bird's Free Throw

- In practice, present only one of these interpretations.
- Prefer the second interpretation and could rephrase it as "The estimated odds of a made second free throw are **30% larger** when the first free throw is missed than when the first free throw is made."

Incorrect Interpretation:

- "The estimated odds of a made second free throw are 1.3 times as **likely** … " is incorrect because "likely" means probabilities are being compared.
- Replacing "odds" with "probability" in any correct interpretation; remember, "odds" are not the same as probabilities.
- "The estimated odds are 1.3 times higher …" is incorrect because 1.3 means 30% times higher, not 130%. [Review the relative risk interpretation discussion if needed.]

Berkeley SCHOOL OF INFORMATION