Asymptotic Inference for 2x2 Tables

Notation

When we are interested in the association between two binary variables, our contingency table reduces to a 2x2 table, and our notation simplifies:

		Y		
		1	2	Total
X	1	n_{11}	n_{12}	$\mathbf{n_1}$
	2	n_{21}	n_{22}	$\mathbf{n_2}$
	Total			$\mathbf{n}$

Typically, if we consider Y to be the response variable where $Y = 1$ denotes a "success", and X is the explanatory variable, we are interested in comparing the two success probabilities: $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$.

Example

Bozeman is known nationwide as one of the premier ski towns in the US. In fact, many students choose to attend MSU for that reason. With the hopes of increasing enrollment, MSU's advertising team created two brochures. One of the brochures has a skier on the front, and the other has a snowboarder on the front. One hundred and fifty California students were chosen at random, half were randomly assigned to receive the skier brochure, and half were randomly assigned to receive the snowboarder brochure. The advertising team wants to know if the probability a California student enrolls at MSU differs based on the type of brochure they receive. The data are summarized below.

```
dat <- matrix(c(17,14,58,61), nrow=2, ncol=2, byrow=TRUE,
              dimnames=list(c("Enrolled","Not Enrolled"),
                            c("Skier","Snowboarder")))
dat
```

```
##              Skier Snowboarder
## Enrolled        17          14
## Not Enrolled    58          61
```

Let π_1 be the probability a California student enrolls after receiving the skier brochure, and π_2 be the probability after receiving the snowboarder brochure. The maximum likelihood estimates of π_1 and π_2 are the sample proportions for the skier group and snowboarder group, respectively:

$\hat{\pi}_1 = \frac{17}{75} \approx 0.227$

$\hat{\pi}_2 = \frac{14}{75} \approx 0.187$

We will use this example to work through asymptotic statistical inference methods for three parameters of interest:

1. Difference in probabilities: $\pi_1 - \pi_2$
2. Relative risk (RR): π_1/π_2
3. Odds ratio (OR): $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

Difference in proportions

Parameter of interest: $\theta_D = \pi_1 - \pi_2$

Point estimate: $\hat{\theta}_D = \hat{\pi}_1 - \hat{\pi}_2$

Standard error:

$$SE(\hat{\theta}_D) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Under the assumption of $H_0 : \pi_1 = \pi_2$, we use the pooled sample proportion, $\hat{\pi}$, in place of $\hat{\pi}_1$ and $\hat{\pi}_2$ when calculating the standard error, resulting in the **null standard error**:

$$SE_0(\hat{\theta}_D) = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Asymptotic distribution of $\hat{\theta}_D$: For large samples,

$$\hat{\theta}_D \dot{\sim} N\left(\theta_D, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Example

Calculations:

```
pi_hat1 <- 17/75
pi_hat2 <- 14/75
pi_hat_pool <- (17+14)/(75+75)
se <- sqrt(pi_hat1*(1-pi_hat1)/75 + pi_hat2*(1-pi_hat2)/75)
null_se <- sqrt(pi_hat_pool*(1 - pi_hat_pool)*(1/75 + 1/75))
```

Point estimate:

```
pi_hat1 - pi_hat2
```

```
## [1] 0.04
```

Example Interpretation: The sample proportion of California students receiving the skier brochure that enrolled is 0.04 higher than the sample proportion of California students receiving the snowboarder brochure that enrolled.

Approximate 95% confidence interval:

```
pi_hat1 - pi_hat2 + c(-1,1) * qnorm(.975) * se
```

```
## [1] -0.08943899  0.16943899
```

Example Interpretation: We are 95% confident that the probability a California student would enroll at MSU after receiving the skier brochure is between 0.089 lower to 0.169 higher than the probability of enrolling after receiving the snowboarder brochure.

Standardized statistic under $H_0 : \pi_1 - \pi_2 = 0$:

```
z <- (pi_hat1 - pi_hat2)/null_se
z
```

```
## [1] 0.6049404
```

Example Interpretation: Our sample difference in proportions of 0.04 lies 0.60 standard errors above the null hypothesized value of zero.

Approximate p-value for $H_a : \pi_1 - \pi_2 \neq 0$:

```
2 * pnorm(-abs(z))
```

```
## [1] 0.5452186
```

Example Conclusion: We have no significant evidence that the probability of enrolling at MSU differs between the skier brochure and the snowboarder brochure among all California students.

Relative risk

Parameter of interest: $\theta_R = \dfrac{\pi_1}{\pi_2}$

Point estimate: $\hat{\theta}_R = \dfrac{\hat{\pi}_1}{\hat{\pi}_2}$

The sampling distribution of the sample relative risk, $\hat{\theta}_R$ is heavily right skewed. However, the sampling distribution of $\log \hat{\theta}_R$ is relatively symmetric and has an asymptotic normal distribution. Thus, inference for the relative risk is conducted on the log scale.

Standard error of $\log \hat{\theta}_R$:

$$SE(\log \hat{\theta}_R) = \sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}}$$

Example

Point estimate:

```
pi_hat1/pi_hat2
```

```
## [1] 1.214286
```

```
# Percent increase/decrease from denominator to numerator
(pi_hat1/pi_hat2 - 1)*100
```

```
## [1] 21.42857
```

Example Interpretation: The sample proportion of California students who enrolled at MSU after receiving the skier brochure is 21% higher than the sample proportion who enrolled after receiving the snowboarder brochure.

Approximate 95% confidence interval:

```
est_log <- log(pi_hat1/pi_hat2)
se_log <- sqrt((1-pi_hat1)/(75*pi_hat1) + (1-pi_hat2)/(75*pi_hat2))
CI_log <- est_log + c(-1,1) * qnorm(.975) * se_log
# Exponentiate to get back to original scale
exp(CI_log)
```

```
## [1] 0.646196 2.281799
```

Example Interpretation: We are 95% confident that the probability a California student enrolls at MSU after receiving a skier brochure is between 35.38% lower to 128.18% higher than the probability of enrolling after receiving the snowboarder brochure.

Odds ratio

Parameter of interest: $\theta_O = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

Point estimate: $\hat{\theta}_O = \dfrac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \dfrac{n_{11}n_{22}}{n_{12}n_{21}}$

The sampling distribution of the sample odds ratio, $\hat{\theta}_O$ is heavily right skewed. However, the sampling distribution of $\log \hat{\theta}_O$ is relatively symmetric and has an asymptotic normal distribution. Thus, as in the case of the relative risk, inference for the odds ratio is conducted on the log scale.

Standard error of $\log \hat{\theta}_O$:

$$SE(\log \hat{\theta}_O) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Example

Point estimate:

```
or <- (17*61)/(58*14)
or
```

```
## [1] 1.277094
```

```
# Percent increase/decrease from denominator to numerator
(or - 1)*100
```

```
## [1] 27.70936
```

Example Interpretation: The sample odds of enrolling at MSU for California students who received the skier brochure is 28% higher than the sample odds of enrolling after receiving the snowboarder brochure.

Approximate 95% confidence interval:

```
or_log <- log(or)
se_log <- sqrt(1/17 + 1/58 + 1/14 + 1/61)
CI_log <- or_log + c(-1,1) * qnorm(.975) * se_log
# Exponentiate to get back to original scale
exp(CI_log)
```

```
## [1] 0.5776054 2.8236718
```

Example Interpretation: We are 95% confident that the true odds that a California student enrolls at MSU after receiving a skier brochure is between 42.24% lower to 182.37% higher than the odds of enrolling after receiving the snowboarder brochure.