Binomial Data Summary Sheet

Reference Table

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	n_{00}	n_{01}	n_{0+}
	$X=1$	n_{10}	n_{11}	n_{1+}
	total	n_{+0}	n_{+1}	n

1 Frequency Measures

- Risk of disease

  $P(Y=1) = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{n_{+1}}{n}$

- Odds of disease

  $\text{Odds} = \frac{P(Y=1)}{P(Y=0)}$

  $\hat{\text{Odds}} = \frac{n_{+1}}{n_{+0}}$

- Conditional Risk of Disease:

  $\pi_0 = P(Y=1|X=0)$

  $\hat{\pi}_0 = \frac{n_{01}}{n_{0+}}$

  $\pi_1 = P(Y=1|X=1)$

  $\hat{\pi}_1 = \frac{n_{11}}{n_{1+}}$

2 Association Measures

- Relative Risk (RR)

  $RR = \frac{\pi_1}{\pi_0}$

  $\hat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_0}$

- Excess Relative Risk (ERR)

  $\text{ERR} = (\hat{RR} - 1)100$

- Risk Difference (RD)

  $RD = \pi_1 - \pi_0$

  $\hat{RD} = \hat{\pi}_1 - \hat{\pi}_0$

- Odds Ratio (OR)

  $OR = \frac{\pi_1/(1-\pi_1)}{\pi_0(1-\pi_0)}$

  $\hat{OR} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_0(1-\hat{\pi}_0)}$

Some comments...

- Note that RD and RR may yield *very* different results. Best to check both values and underlying probabilities
- The odds ratio is often used to approximate the RR when the RR cannot be computed directly from data (e.g. case control)

  - OR can be estimated consistently for biased samples; easily computed via logistic regression

  - accuracy depends on baseline risk; $\pi_0 \uparrow \Rightarrow OR$ accuracy $\downarrow$

3 Study Designs

1. Cohort Study (retro or prospective)

   RR, OR relevant

2. Case Control Study

   RR irrelevant, OR relevant

3. Cross Sectional Study

   RR, OR relevant

Set-up:
In order to understand how different study designs and analysis estimate the truth, assume the following to be true for the next four examples.
$RR = 2$, $P(X = 1) = 0.2$
$P(Y = 1|X = 0) = \pi_0 = 0.05$ $P(Y = 1|X = 1) = \pi_1 = 0.1$

3.1 Example 1: Prospective Cohort

Prospective Cohort Data

		Health Status		
		$Y = 0$	$Y = 1$	total
Treatment	$X = 0$	7613	385	7988
	$X = 1$	1808	194	579
	total	9421	579	10000

- $\hat{\pi}_0 = \frac{385}{7998} \approx 0.048$
- $\hat{\pi}_1 = \frac{194}{2002} \approx 0.096$
- $\hat{\pi}^* = P(Y = 1) = \frac{579}{10000} = 0.0579$
- $\hat{RD} = 0.096 - 0.048 = 0.048$
- $\hat{RR} = \frac{0.096}{0.048} = 2$
- $\hat{OR} = \frac{7613 \times 194}{385 \times 1808} \approx 2.12$

3.1.1 Comments:

- We observe that $\hat{RR}, \hat{OR}$, $\hat{\pi}_0, \hat{\pi}_1$ all accurately estimate the 'true' values

3.2 Example 2: Retrospective Cohort

Retrospective Cohort Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	4748	252	5000
	$X=1$	4465	535	5000
	total	9213	787	10000

- $\hat{\pi}_0 = \frac{252}{5000} \approx 0.0504$
- $\hat{\pi}_1 = \frac{535}{5000} \approx 0.107$
- $\hat{\pi}^* = P(Y=1) = \frac{787}{10000} = 0.0787$
- $\hat{RD} = 0.107 - 0.0504 = 0.0566$
- $\hat{RR} = \frac{0.107}{0.0504} = 2.12$
- $\hat{OR} = \frac{4748 \times 535}{252 \times 4465} \approx 2.26$

3.2.1 Comments:

- We observe that $\hat{RR}, \hat{OR}$, $\hat{\pi}_0, \hat{\pi}_1$ all accurately estimate the 'true' values

3.3 Example 3: Case-Control Study

Case-Control Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	4022	3358	7380
	$X=1$	978	1642	2620
	total	5000	5000	10000

- $\hat{\pi}_0 = \frac{3358}{7380} \approx 0.4555$ (biased)
- $\hat{\pi}_1 = \frac{1642}{2620} \approx 0.627$ (biased)
- $\hat{\pi}^* = P(Y=1) = \frac{5000}{10000} = 0.5$ (biased)
- $\hat{RD} = 0.627 - 0.455 = 0.172$ (biased)
- $\hat{RR} = \frac{0.627}{0.455} = 1.378$ (biased)
- $\hat{OR} = \frac{4022 \times 1642}{3358 \times 978} \approx 2.011$ (unbiased)

3.3.1 Comments:

- Only $\hat{OR}$ accurately estimates the true RR, OR

- All other estimates are biased because the selection method
- the $\hat{OR}$ is consistent because the exposure odds ratio is equivalent to the disease odds ratio i.e.

$$\begin{aligned} EOR &= \frac{ODDS(X=1|Y=1)}{ODDS(X=1|Y=0)} = \frac{P(X=1|Y=1)}{P(X=0|Y=1)}\frac{P(X=0|Y=0)}{P(X=1|Y=0)} \\ &= \frac{P(X=1,Y=1)}{P(Y=1)}\frac{P(Y=1)}{P(X=0,Y=1)}\frac{P(X=0,Y=0)}{P(Y=0)}\frac{P(Y=0)}{P(X=1,Y=0)} \\ &= \frac{P(X=1,Y=1)}{P(X=0,Y=1)}\frac{P(X=0,Y=0)}{P(X=1,Y=0)} = \frac{n_{11}/n \times n_{00}/n}{n_{01}/n \times n_{10}/n} \\ &= \frac{n_{11}n_{00}}{n_{01}n_{10}} = OR \end{aligned}$$

3.4 Example 4: Recall Bias

- If subjects are (uniformly) randomly misclassified, the odds ratio becomes biased towards the null (OR =1)
- Otherwise bias can be in either direction
- Recall Bias is one example of this

Suppose 20% of smokers without cancer misclassified as non-smokers

Recall-Bias Case Control Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	4231	3279	7312
	$X=1$	769	1729	2498
	total	5000	5000	10000

- $\hat{\pi}_0 = \frac{3279}{7312} \approx 0.448$ (biased)
- $\hat{\pi}_1 = \frac{1729}{2498} \approx 0.692$ (biased)
- $\hat{\pi}^* = P(Y=1) = \frac{5000}{10000} = 0.5$ (biased)
- $\hat{RD} = .692 - .448 = 0.244$ (biased)
- $\hat{RR} = \frac{.692}{0.448} = 1.544$ (biased)
- $\hat{OR} = \frac{4231 \times 1729}{3279 \times 769} \approx 2.91$ **(upward bias)**

3.4.1 Comments:

- Note that the $\hat{OR}$ is biased upward, since the misclassification accentuates results in a higher proportion of smokers getting cancer (or vice versus).

4 Other forms of Bias

- **Selection Bias**

  Sample is not representative of population of interest

- Confounding

  Unknown covariate C associated with X and Y

  C leads to bias

4.1 Bias Example: Alcohol confounded by Smoking

Lung Cancer- Alcohol Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	91	19	110
	$X=1$	19	91	110
	total	110	110	220

$$\hat{OR} = \frac{91 \times 91}{19 \times 19} = 22.94$$

There appears to be a huge effect. However, when stratified by smoking (see below)

$S = 0$: Alcohol/Cancer Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	90	9	99
	$X=1$	10	1	11
	total	100	10	110

$$\hat{OR} = \frac{90 \times 1}{9 \times 10} = 1$$

$S = 1$: Alcohol/Cacner Data

		Health Status		
		$Y = 0$	$Y = 1$	total
Treatment	$X = 0$	1	10	11
	$X = 1$	9	90	99
	total	10	100	110

$$\hat{OR} = \frac{90 \times 1}{9 \times 10} = 1$$

4.1.1 Comment

Notice That the two odds ratios are the same. This means that when adjusting for smoking, the odds of getting lung cancer are the same regardless if you drink or not. If you run the numbers looking at smoking's effect, it will be different for smokers vs. non smokers.