Binary data – 2 group estimation

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

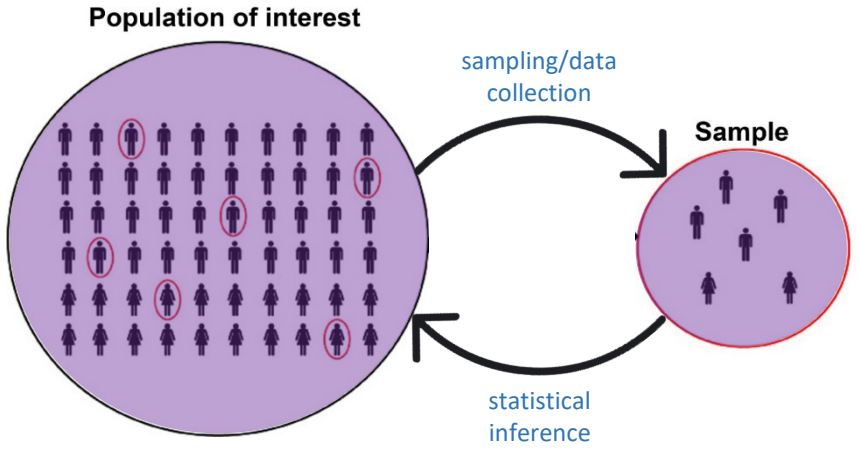
One-group vs. Two-group

- We've discussed estimation and testing for a binary variable in one group
  - Compare the proportion in the category of interest to a known value

- Often we're interested in comparing a binary variable in two groups
  - Compare the proportion in the category of interest in one population to the proportion in the category of interest in another population

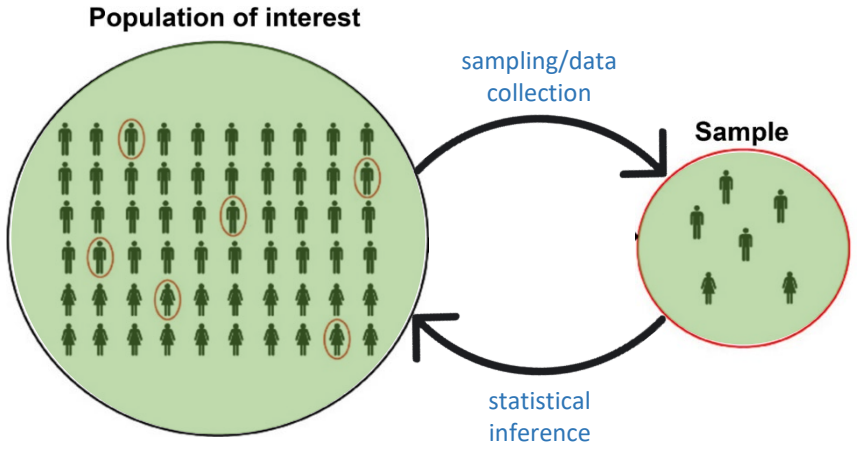
Two Groups: Notation

- p_1 = population proportion in category of interest in group 1



- $\hat{p}_1$ = sample proportion in category of interest in group 1

- p_2 = population proportion in category of interest in group 2



- $\hat{p}_2$ = sample proportion in category of interest in group 2

Calculating Sample Proportions

- Typically we think of one binary variable as the "exposure" and the other binary variable as the "outcome"
- The two groups of interest are the exposed and unexposed
- Proportions are the percentage of "Yes" outcomes

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	n_1
No	c	d	n_2
Total	m_1	m_2	N

$$\hat{p}_1 = \frac{a}{n_1} \qquad \hat{p}_2 = \frac{c}{n_2}$$

Two Groups: Estimation

- We want a single number representing the relationship between the two proportions
- When working with continuous data (mean in each group), we used the difference in means ($\mu_1 - \mu_2$)
- There are three common statistics used to relate the proportion in the category of interest in 2 groups:
  - Risk difference
  - Risk ratio
  - Odds ratio

Risk Difference (RD)

- Risk difference = difference between Group 1 proportion and Group 2 proportion

RD in population:

$$p_1 - p_2$$

RD in sample:

$$\hat{p}_1 - \hat{p}_2$$

- Interpretation is in terms of the difference in percentage of “Yes” outcomes between the two groups
- What value of RD indicates no difference in the percentage of “Yes” outcomes in each group?

Risk Ratio (RR)

- Risk ratio (also called relative risk) = ratio of Group 1 proportion to Group 2 proportion

RR in population:

$$\frac{p_1}{p_2}$$

RR in sample:

$$\frac{\hat{p}_1}{\hat{p}_2}$$

- Interpretation is in terms of the how many times larger the percentage of “Yes” outcomes is in group 1 compared to the percentage of “Yes” outcomes in group 2
- What value or RR indicates no difference in the percentage of “Yes” outcomes in each group?

Odds Ratio (OR)

- Odds ratio = ratio of odds in Group 1 to odds in Group 2

Aside: Odds

- Odds refers to how many times more likely the "Yes" outcome is compared to the "No" outcome
- Odds ≠ probability
- If p = probability then odds = $p/(1-p)$
- Probabilities range between 0 and 1; odds range between 0 and ∞

Odds Ratio (OR)

- Odds ratio = ratio of odds in group 1 to odds in group 2

Aside: Odds

- Example: If the odds of winning is 3 then the probability of winning is 3 times more likely than the probability of losing
- Example: If the odds of winning is 0.33 then the probability of winning is 33% as large as the probability of losing.

Odds Ratio (OR)

- Odds ratio = ratio of odds in group 1 to odds in group 2

Aside: Odds

- Odds of "Yes" outcome in group 1: $p_1/(1-p_1)$
- Odds of "Yes" outcome in group 2: $p_2/(1-p_2)$

Odds Ratio (OR)

- Odds ratio = ratio of odds in group 1 to odds in group 2

OR in population:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

OR in sample:

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

- Interpretation is in terms of the how many times larger the odds of the "Yes" outcome is in group 1 compared to the odds of "Yes" outcome in group 2
- What value of OR indicates no difference in the percentage of "Yes" outcomes in each group?

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
sex	Sex of the baby (Male, Female)
birthwt	Birth weight of the baby (g)
gestage	Gestational age (weeks)
hemorrhage	Germinal matrix hemorrhage in the baby (Yes, No)
toxemia	Toxemia diagnosis for the mother (Yes, No)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Hemorrhage

- Use the risk difference, risk ratio, and odds ratio to compare the proportion of female infants with germinal matrix hemorrhage to the proportion of male infants with germinal matrix hemorrhage.

Example: Hemorrhage, RD

2x2 table:

sex	hemorrhage Yes	No
Female	11	45
Male	4	40

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Female	11	45	56
	Male	4	40	44
	Total	15	85	100

$\hat{p}_1 =$ 11/56=0.1964 $\hat{p}_2 =$ 4/44=0.0909

Risk difference: 0.1964-0.0909=0.1055

The difference between the percentage of female infants with germinal matrix hemorrhage and the percentage of male infants with germinal matrix hemorrhage is 10.55%

Example: Hemorrhage, RR

2x2 table:

sex	hemorrhage Yes	No
Female	11	45
Male	4	40

$$\hat{p}_1 = \frac{11}{56} = 0.1964 \qquad \hat{p}_2 = \frac{4}{44} = 0.0909$$

Risk ratio: 0.1964/0.0909=2.16

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Female	11	45	56
	Male	4	40	44
	Total	15	85	100

The percentage of female infants with germinal matrix hemorrhage is 2.16 times larger than the percentage of male infants with germinal matrix hemorrhage.

Example: Hemorrhage, OR

2x2 table:

sex	hemorrhage Yes	No
Female	11	45
Male	4	40

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Female	11	45	56
	Male	4	40	44
	Total	15	85	100

$$\hat{p}_1 = \frac{11}{56} = 0.1964 \qquad \hat{p}_2 = \frac{4}{44} = 0.0909$$

Odds ratio: 0.1964/(1-0.1964) / 0.0909/(1-0.0909) = 2.44

The odds of germinal matrix hemorrhage among female infants is 2.44 times larger than the odds of germinal matrix hemorrhage among male infants.

Example: Hemorrhage

- Use the risk difference, risk ratio, and odds ratio to compare the proportion of male infants with germinal matrix hemorrhage to the proportion of female infants with germinal matrix hemorrhage.

p1 is probability of germinal matrix hemorhage among male infants

p2 is prob. of gmh among female infants

(same as before, but switched groups 1 and 2)

Example: Hemorrhage, RD

2x2 table:

sex	hemorrhage Yes	No
Male	4	40
Female	11	45

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Male	4	40	44
	Female	11	45	56
	Total	15	85	100

$\hat{p}_1 =$ 0.0909 $\hat{p}_2 =$ 0.1964

Risk difference: 0.0909-0.1964=-0.1055

The difference in probability of germinal matrix hemorrhage between male and female infants (male-female) is -10.55%

Example: Hemorrhage, RR

2x2 table:

sex	hemorrhage Yes	No
Male	4	40
Female	11	45

$$\hat{p}_1 = \frac{4}{44} = 0.0909 \qquad \hat{p}_2 = \frac{11}{56} = 0.1964$$

Risk ratio: 0.0909/0.1964=0.46

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Male	4	40	44
	Female	11	45	56
	Total	15	85	100

The percentage of male infants with germinal matrix hemorrhage is 0.46 times the percentage of female infants with germinal matrix hemorrhage.

Example: Hemorrhage, OR

2x2 table:

sex	hemorrhage Yes	No
Male	4	40
Female	11	45

$$\hat{p}_1 = \frac{4}{44} = 0.0909 \qquad \hat{p}_2 = \frac{11}{56} = 0.1964$$

Odds ratio: 0.41

		OUTCOME (hemorrhage)		
		Yes	No	Total
EXPOSURE (gender)	Male	4	40	44
	Female	11	45	56
	Total	15	85	100

The odds of germinal matrix hemorrhage among male infants is 0.41 times the odds of germinal matrix hemorrhage among female infants.

Confidence Intervals for RD, RR, & OR

- We can calculate confidence intervals for the risk difference (RD), risk ratio (RR), and odds ratio (OR)
- Focus on CI concepts and interpretation; don't worry about the underlying formulas

Interpretation:
"We are ___% confident that the (RD/RR/OR) comparing _____ to _____ in the population is between _____ and _____."

Plausible Values

- The CI contains what we consider to be plausible values for the measure of interest (RD, RR, or OR)
- Gives us a standardized way to answer the question, “Is the probability of the outcome of interest the same in the two groups?”
- For RD: If 0 is not in the CI, we’re confident that the answer is “No”.
  For RR: If 1 is not in the CI, we’re confident that the answer is “No”.
  For OR: If 1 is not in the CI, we’re confident that the answer is “No”.

Example: Hemorrhage

- Calculate 95% confidence intervals for the risk difference, risk ratio, and odds ratio to compare the proportion of female infants with germinal matrix hemorrhage to the proportion of male infants with germinal matrix hemorrhage.

Example: Hemorrhage

- Calculate 95% confidence intervals for the risk difference, risk ratio, and odds ratio to compare the proportion of female infants with germinal matrix hemorrhage to the proportion of male infants with germinal matrix hemorrhage.

2x2 table:

	hemorrhage	
sex	Yes	No
Female	11	45
Male	4	40

Example: Hemorrhage, RD

risk difference	RD CI lower	RD CI upper
0.10551948	-0.02880509	0.23984405

95% CI for RD: (-0.029, 0.240)

We are 95% confident that the risk difference comparing the probability of germinal matrix hemorrhage in females to the probability of germinal matrix hemorrhage in males is between -0.029 and 0.240.

Example: Hemorrhage, RR

risk ratio RR	CI lower RR	CI upper
2.160714	0.738113	6.325165

95% CI for RR: (0.738, 6.325)

We are 95% confident that the risk ratio comparing the probability of germinal matrix hemorrhage in females to the probability of germinal matrix hemorrhage in males is between 0.738 and 6.325.

Example: Hemorrhage, OR

odds ratio OR	CI lower OR	CI upper
2.4444444	0.7209083	8.2885834

95% CI for OR: (0.721, 8.289)

We are 95% confident that the odds ratio comparing the odds of germinal matrix hemorrhage in females to the odds of germinal matrix hemorrhage in males is between 0.721 and 8.289.

Example: Hemorrhage

95% CI for RD: (-0.029, 0.240)
95% CI for RR: (0.738, 6.325)
95% CI for OR: (0.721, 8.289)

Since the 95% CI for the risk difference includes 0 and the 95% CI for the risk ratio and odds ratio include 1, there is not sufficient evidence to say that the probability of germinal matrix hemorrhage is different in male and female infants.

Study Design Considerations

- So far, everything we've done in this course has assumed that our sample is a random selection of subjects from the population
- This isn't always the reality
  - We need to observe enough "Yes" and "No" outcomes and enough exposed and unexposed subjects in order to make valid statistical inferences
  - Financial constraints, time constraints, and the prevalence of the outcome/exposure can all necessitate non-random sampling
- There are two main ways to do non-random sampling from the population. We refer to these as cohort and case-control studies.

Types of Studies

COHORT STUDIES

- Select subjects based on exposure status
- Randomly select of a certain number of exposed subjects; randomly select a certain number of unexposed subjects
- Observe outcomes on these subjects
- **RD, RR, and OR are all appropriate**

CASE-CONTROL STUDIES

- Select subjects based on outcome
- Randomly select of a certain number of subjects who have the outcome of interest; randomly select a certain number of subjects who don't have the outcome of interest
- Retrospectively determine exposure status of these subjects
- **RD and RR are NOT appropriate; only OR is appropriate**

Types of Studies

COHORT STUDY EXAMPLE

- Researchers are interested in comparing the probability of developing lung cancer among smokers and non-smokers. They recruit 100 smokers and 100 non-smokers, and they follow them to determine whether they develop lung cancer or not.

CASE-CONTROL STUDY EXAMPLE

- Researchers are interested in comparing the probability of developing lung cancer among smokers and non-smokers. They recruit 100 people with lung cancer and 100 people without lung cancer, and they ask them whether they have smoked in their life or not.

Study Design Considerations

- Recap:
  - When the sample is randomly selected from the population, RD, RR, and OR can be used
  - If the sample is randomly selected *within levels of the exposure* (cohort study), RD, RR, and OR can be used
  - If the sample is randomly selected *within levels of the outcome* (case-control study), only the OR can be used

Important Points

- Set up of 2x2 table and calculation of sample proportions
- Calculation and interpretation of point estimates and confidence intervals for…
  - Risk difference
  - Risk ratio
  - Odds ratio
- Random vs. non-random sampling
  - Non-random sampling strategies: cohort and case-control studies
  - When each measure (RD, RR, OR) is appropriate to use