Binary data – 2 group testing, power, and sample size

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)
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.
apgarlow	Categorizes the Apgar score into two categories: Normal (7-10) or Low (0-6).

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

Fundamental Rule of Data Analysis

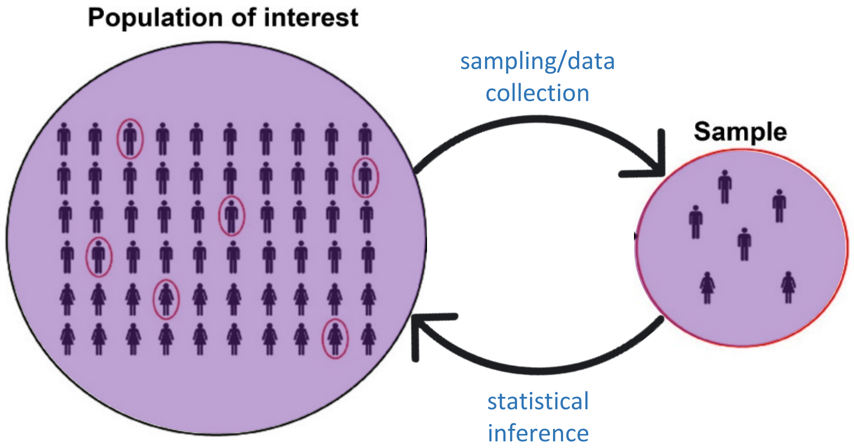
Different types of data require different statistical analyses.

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

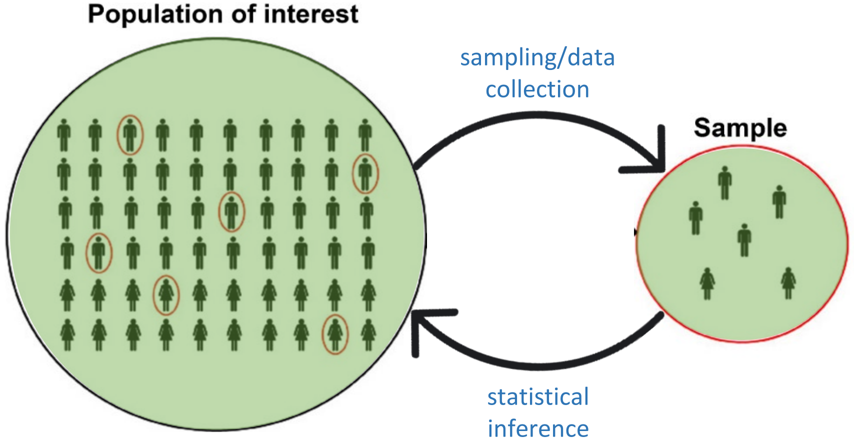
Two Groups: Notation

- 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

Two Groups: Testing

- We can test whether the two group proportions are equal with a hypothesis test

- There are two* different tests that can be used to relate the proportion in the category of interest in 2 groups:
  - Chi-squared (χ^2) test
  - Fisher's exact test

* Can also use confidence intervals for risk difference, risk ratio, or odds ratio.

Chi-squared (χ^2) Test of Independence

- Tests whether the outcome variable and grouping variable are independent

H_0: $p_1 = p_2$ $\quad\quad\quad$ H_A: $p_1 \neq p_2$

OR

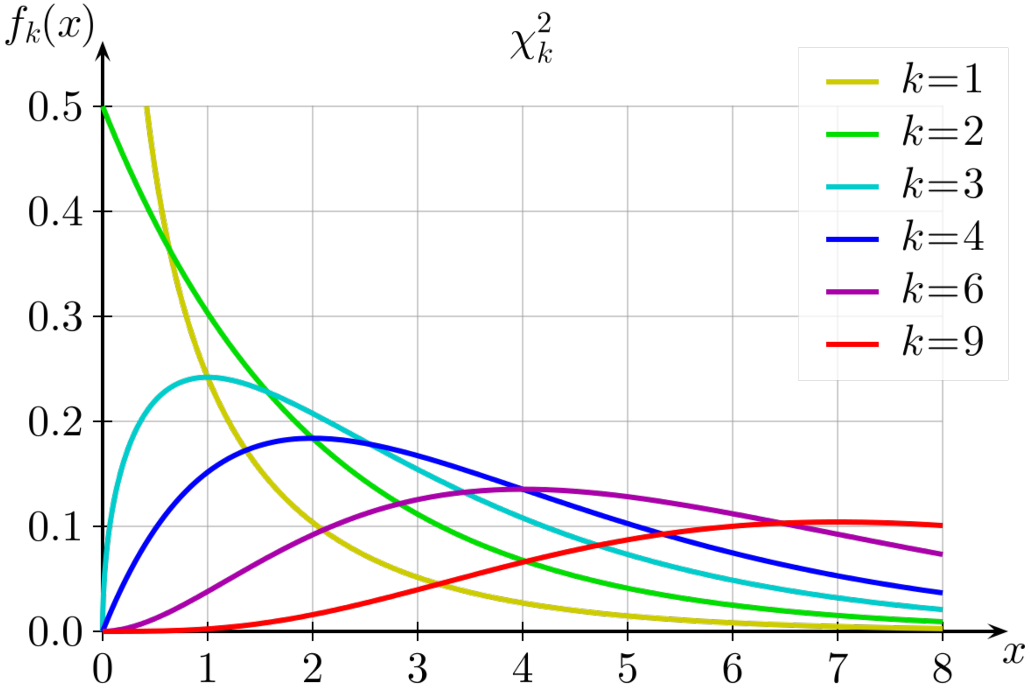
H_0: there is no association between [outcome variable] and [grouping variable]

H_A: there is an association between [outcome variable] and [grouping variable]

Chi-squared (χ^2) Test of Independence

- χ^2 distribution is used to calculate a p-value
  - If p-value ≤ α, reject H_0
  - If p-value > α, fail to reject H_0

χ^2 Distribution:



Chi-squared (χ^2) Test of Independence

- Only appropriate when no more than 20% of expected cell counts are <5 and no expected cell counts are <1

Expected Counts

- Given the row/column totals, we can calculate the cell counts we would expect if the proportion of "yes" outcomes is the same within each group (if the outcome variable and grouping variable are independent)

OBSERVED COUNTS:

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

EXPECTED COUNTS:

	Outcome		
Exposure	Yes	No	Total
Yes	$\frac{m_1 n_1}{N}$	$\frac{m_2 n_1}{N}$	n_1
No	$\frac{m_1 n_2}{N}$	$\frac{m_2 n_2}{N}$	n_2
Total	m_1	m_2	N

Expected Counts: Example

all calculated cell values >5; none less than 1

OBSERVED COUNTS:

Sex	Hemorrhage Yes	No	Total
Female	11	45	56
Male	4	40	44
Total	15	85	100

EXPECTED COUNTS:

Sex	Hemorrhage Yes	No	Total
Female	8.4	47.6	56
Male	6.6	37.4	44
Total	15	85	100

(56*15)/100=8.4 (56*85)/100=47.6

(44*15)/100=6.6 (44*85)/100=37.4

Example: Hemorrhage

- Use a χ^2 test to test whether the probability of germinal matrix hemorrhage is different in female and male infants.

let p1 be the probability of germinal matrix hemorrhage in female infants

let p2 be the prob. of gmh in male infants

H0: p1=p2

HA: p1!=p2

Example: Hemorrhage

χ^2 test for hemorrhage/sex:

```
Frequency table:
        hemorrhage
sex       Yes No
  Female   11 45
  Male      4 40

        Pearson's Chi-squared test

data:  .Table
X-squared = 2.1518, df = 1, p-value = 0.1424


Expected counts:
        hemorrhage
sex       Yes   No
  Female 8.4 47.6
  Male   6.6 37.4
```

p-value = 0.142

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that the probability of germinal matrix hemorrhage is different among male and female infants.

Example: Hemorrhage

χ^2 test for hemorrhage/sex:

```
Frequency table:
        hemorrhage
sex       Yes No
  Female   11 45
  Male      4 40

        Pearson's Chi-squared test

data:  .Table
X-squared = 2.1518, df = 1, p-value = 0.1424


Expected counts:
        hemorrhage
sex       Yes   No
  Female 8.4 47.6
  Male   6.6 37.4
```

p-value = 0.142

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that there is an association between sex and occurrence of germinal matrix hemorrhage.

Two Groups: Testing

- We can test whether the two group proportions are equal with a hypothesis test

- There are two* different tests that can be used to relate the proportion in the category of interest in 2 groups:
  - Chi-squared (χ^2) test
  - Fisher's exact test

* Can also use confidence intervals for risk difference, risk ratio, or odds ratio.

Fisher's Exact Test

- Tests whether the outcome variable and grouping variable are independent
- Must be used if expected counts are not large enough for χ^2 test

H_0: $p_1 = p_2$ H_A: $p_1 \neq p_2$

OR

H_0: there is no association between [outcome variable] and [grouping variable]

H_A: there is an association between [outcome variable] and [grouping variable]

Fisher’s Exact Test

Recall definition of a p-value:

Technical definition: Given the null hypothesis is true, the p-value is the probability of obtaining the results in your sample or more extreme

- Fisher’s exact test enumerates all possible ways you could arrange the inner cell counts to be “more extreme” than the observed table
  - “More extreme” = leads to $\hat{p}_1$ and $\hat{p}_2$ farther apart
- Fisher’s exact test p-value = Add up the probabilities of each of the “more extreme” tables
  - If p-value ≤ α, reject H_0
  - If p-value > α, fail to reject H_0

Example: Hemorrhage

- Use Fisher’s exact test to test whether the probability of germinal matrix hemorrhage is different in female and male infants.

Example: Hemorrhage

Fisher’s exact test for hemorrhage/sex:

```
Frequency table:
        hemorrhage
sex      Yes No
  Female  11 45
  Male     4 40

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.1683
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6515771 11.2791007
sample estimates:
odds ratio
  2.423891
```

p-value = 0.168

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that the probability of germinal matrix hemorrhage is different among male and female infants.

Example: Hemorrhage

Fisher's exact test for hemorrhage/sex:

```
Frequency table:
        hemorrhage
sex      Yes No
  Female  11 45
  Male     4 40

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.1683
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6515771 11.2791007
sample estimates:
odds ratio
  2.423891
```

p-value = 0.168

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that there is an association between sex and occurrence of germinal matrix hemorrhage.

Power and Sample Size

- Each type of hypothesis test has a formula for calculating the power of the test for a given sample size OR the necessary sample size to achieve a certain power
  - For two-group test of a proportion (binary data), the <u>formula</u> is different than when we were working with two-group t-tests of means (continuous data), but the main <u>underlying concepts</u> are the same

Factors Affecting Power

- Sample size
  - Larger sample size → have more information → higher power
- Effect size (difference between proportion in each group)
  - Bigger difference → easier to detect the difference → higher power
- Significance level
  - Larger α → threshold for rejection isn't as stringent → reject more often → higher power (and also more false positives)

Factors Affecting Necessary Sample Size

- Power
  - Want higher power → need more information → larger sample size necessary
- Effect size (difference between proportion in each group)
  - Smaller difference → harder to detect the difference → larger sample size necessary
- Significance level
  - Smaller α → threshold for rejection is more stringent → reject less often → need more subjects to maintain given power

Example: Apgar Score/Hemorrhage

- Researchers are using this study as pilot data to determine the power they will have in a new study. In the new study, they will recruit 70 infants with a low Apgar score and 70 infants with a normal Apgar score. What power will they have to detect a difference in the probability of germinal matrix hemorrhage between infants with a low Apgar score and infants with a normal score (using $\alpha = 0.05$)?

Example: Apgar Score/Hemorrhage

- Researchers are using this study as pilot data to determine the power they will have in a new study. In the new study, they will recruit 70 infants with a low Apgar score and 70 infants with a normal Apgar score. What power will they have to detect a difference in the probability of germinal matrix hemorrhage between infants with a low Apgar score and infants with a normal score (using $\alpha = 0.05$)?

Let p_1 be the probability of germinal matrix hemorrhage in all low birthweight infants with a low Apgar score.

Let p_2 be the probability of germinal matrix hemorrhage in all low birthweight infants with a normal Apgar score.

Example: Apgar Score/Hemorrhage

- Researchers are using this study as pilot data to determine the power they will have in a new study. In the new study, they will recruit 70 infants with a low Apgar score and 70 infants with a normal Apgar score. What power will they have to detect a difference in the probability of germinal matrix hemorrhage between infants with a low Apgar score and infants with a normal score (using $\alpha = 0.05$)?

Frequency table:

```
         hemorrhage
apgarlow Yes No
  Low     11 29
  Normal   4 56
```

Row percentages:

```
         hemorrhage
apgarlow        Yes         No
  Low    0.27500000 0.72500000
  Normal 0.06666667 0.93333333
```

Hypothesized p in Group 1 (p_1) = 0.275

Hypothesized p in Group 2 (p_2) = 0.0667

Sample size in each group = 70

Significance level = 0.05

Example: Apgar Score/Hemorrhage

Output from two-group proportion test power calculation:

```
              n = 70
             p1 = 0.275
             p2 = 0.067
      sig.level = 0.05
          power = 0.9132905
    alternative = two.sided

NOTE: n is number in *each* group
```

The new study will have 91.3% power to detect a difference in the probability of germinal matrix hemorrhage in infants with a low Apgar score and infants with a normal Apgar score.

Example: Toxemia Status/Apgar Score

- Researchers are using this study as pilot data to determine the number of subjects they need in a new study. In the new study, they want to have 80% power to show a difference in the probability of having an infant with a low Apgar score (as opposed to normal Apgar score) among mothers with and without a toxemia diagnosis (using $\alpha = 0.05$).

Example: Toxemia Status/Apgar Score

- Researchers are using this study as pilot data to determine the number of subjects they need in a new study. In the new study, they want to have 80% power to show a difference in the probability of having an infant with a low Apgar score (as opposed to normal Apgar score) among mothers with and without a toxemia diagnosis (using $\alpha = 0.05$).

Let p_1 be the probability of having an infant with a low Apgar score among all mothers with a toxemia diagnosis.

Let p_2 be the probability of having an infant with a low Apgar score among all mothers without a toxemia diagnosis.

Example: Toxemia Status/Apgar Score

- Researchers are using this study as pilot data to determine the number of subjects they need in a new study. In the new study, they want to have 80% power to show a difference in the probability of having an infant with a low Apgar score (as opposed to normal Apgar score) among mothers with and without a toxemia diagnosis (using $\alpha = 0.05$).

Frequency table:

```
        apgarlow
toxemia Low Normal
    Yes   7     14
    No   33     46
```

Row percentages:

```
        apgarlow
toxemia       Low    Normal
    Yes 0.3333333 0.6666667
    No  0.4177215 0.5822785
```

Hypothesized p in Group 1 (p_1) = 0.333

Hypothesized p in Group 2 (p_2) = 0.418

Power = 0.8

Significance level = 0.05

Example: Toxemia Status/Apgar Score

Output from two-group proportion test sample size calculation:

```
          n = 508.3171
         p1 = 0.333
         p2 = 0.418
  sig.level = 0.05
      power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

They need to recruit 1018 mothers (509 with toxemia and 509 without toxemia) to have 80% power to detect a difference in the probability of having a low Apgar score baby between mothers with and without toxemia.

More than Two Groups

- May be interested in comparing the probability of the outcome in more than two groups
- Can still use χ^2 test and Fisher's exact test

H_0: $p_1 = p_2 = p_3$ H_A: at least one proportion is different from the others

OR

H_0: there is no association between [outcome variable] and [grouping variable]

H_A: there is an association between [outcome variable] and [grouping variable]

Paired Samples

- Recall: Two groups of data are **paired** when each observation in the first group has a corresponding observation in the second group
- Oftentimes the two observations are within the same subject
  - Example: Recruit 100 subjects. In Year 1 of study, subjects do not have the flu shot. Record if each subject got the flu that year. In Year 2 of study, subjects are given the flu shot. Record if each subjects got the flu that year. Is getting the flu shot associated with getting the flu?
- In this scenario, we use **McNemar’s test**

Important Points

- Set up and interpretation of χ^2 test
    - Expected counts
- Set up and interpretation of Fisher's exact test
- When to use χ^2 vs. Fisher's exact test
- Performing power and sample size calculations for hypothesis test comparing proportion in two groups
- When there are >2 groups, χ^2 and Fisher's exact tests can still be performed (set up and interpretation is nearly identical)
- When data are paired, McNemar's test should be used