Binary data – numerical and graphical summaries, binomial distribution, 1 group estimation

Low Birth Weight Data

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

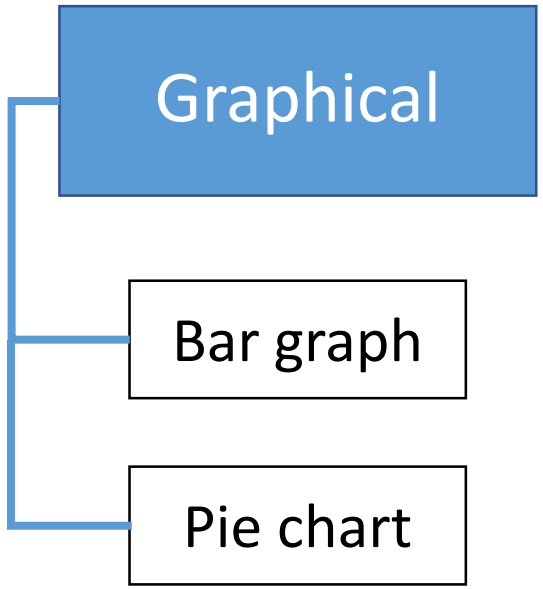
Variable	Description
sex	Sex of the baby (Male, Female)
birthwt	Birth weight of the baby (g)
gestage	Gestational age (weeks)
hemorrhage	Germinal matrix hemorrhage in the baby (Yes, No)
toxemia	Toxemia diagnosis for the mother (Yes, No)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system 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.

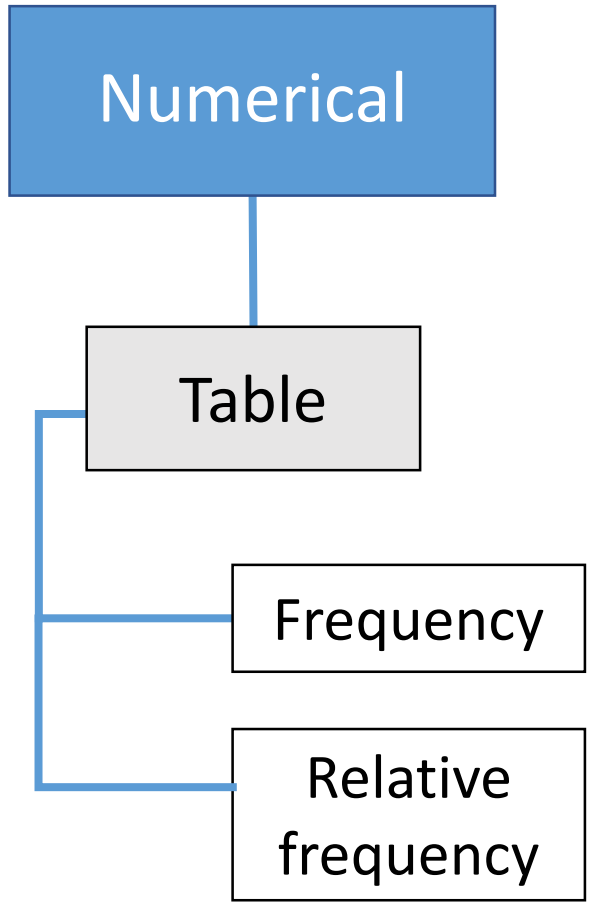
Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

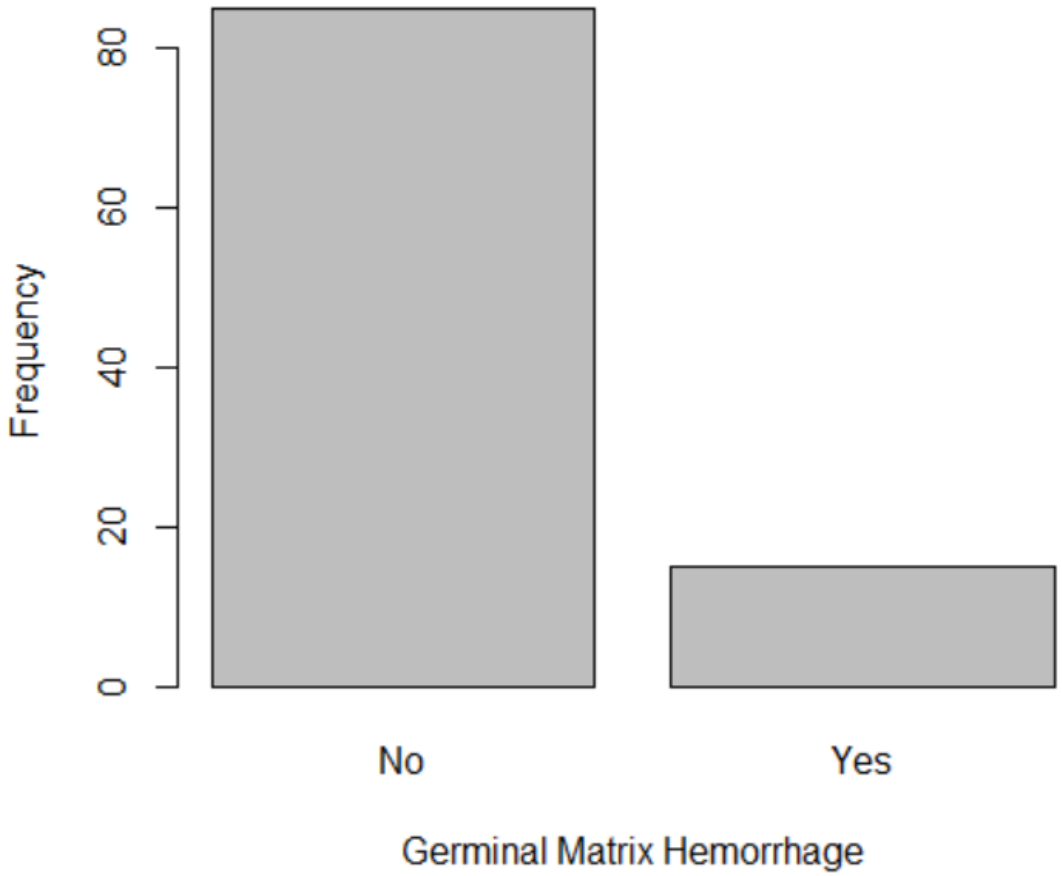
Summaries of Binary Data





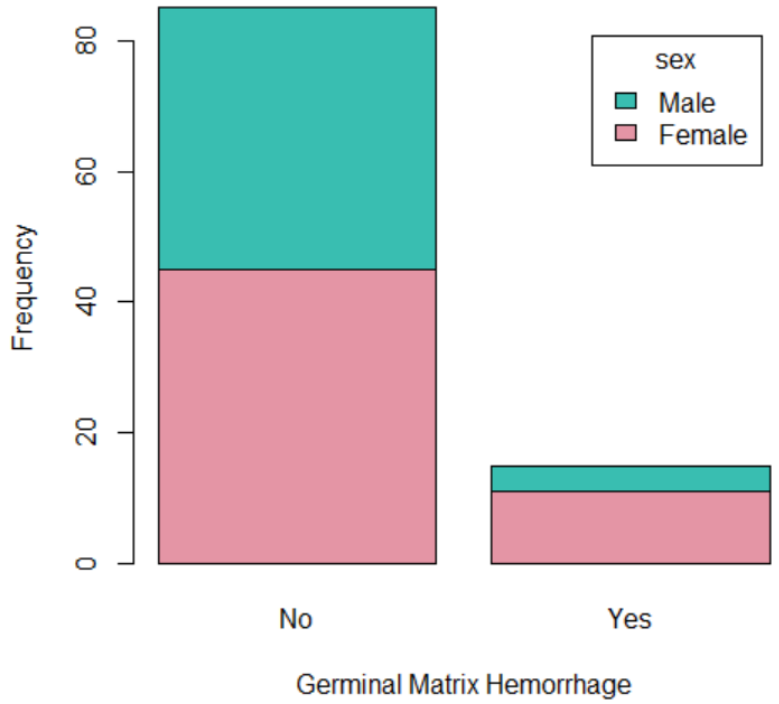
Bar Graph

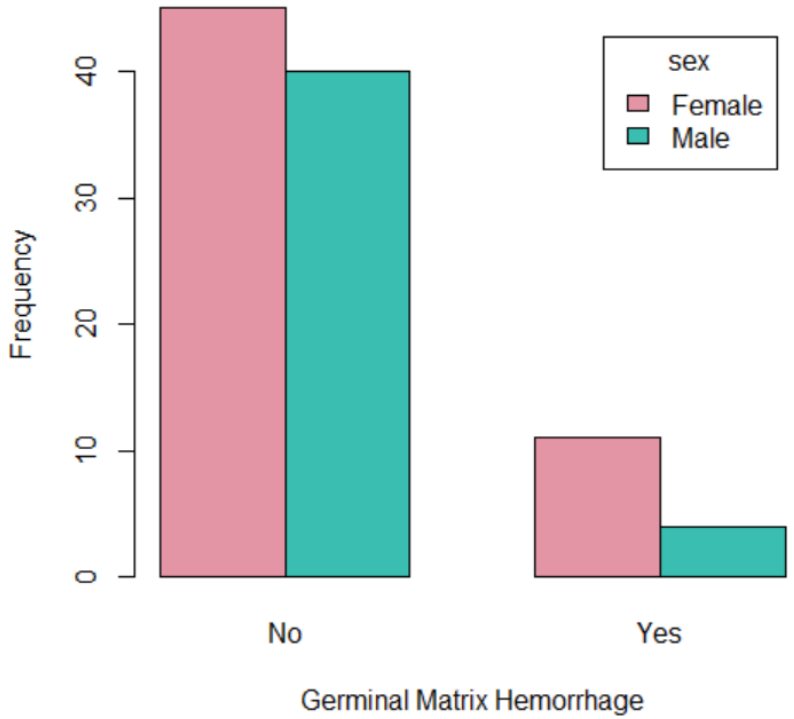
- Separate bars created by categories on horizontal axis
  - Bars should not be connected
- Height of bar represents the number of observations (frequency) in that category



Side-by-side Bar Graph

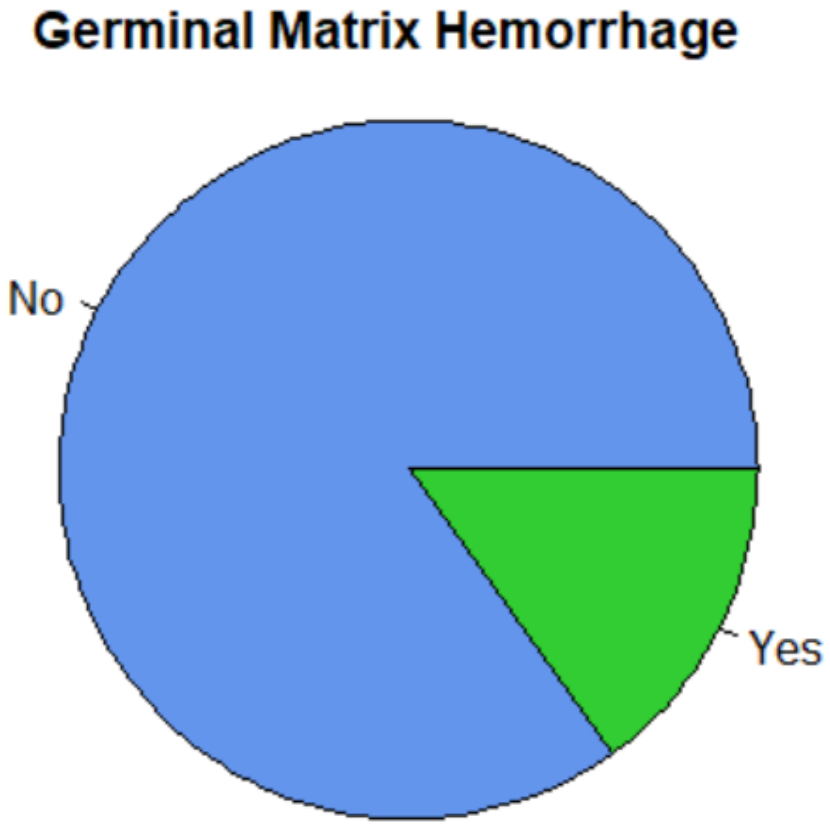
- Can show a binary variable within levels of another binary variable
- Bars can be stacked or side-by-side



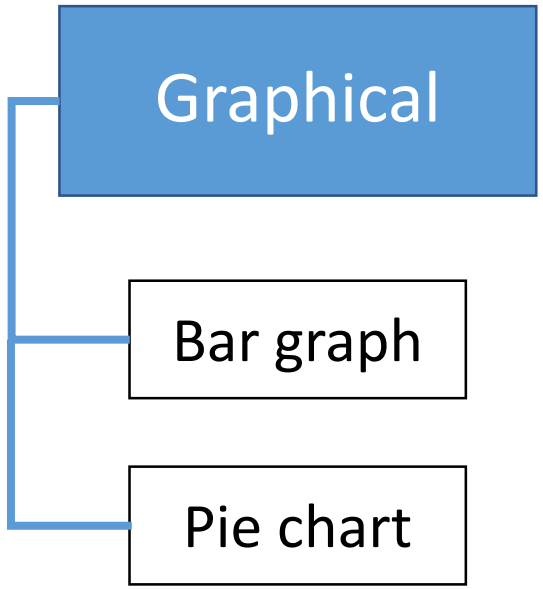


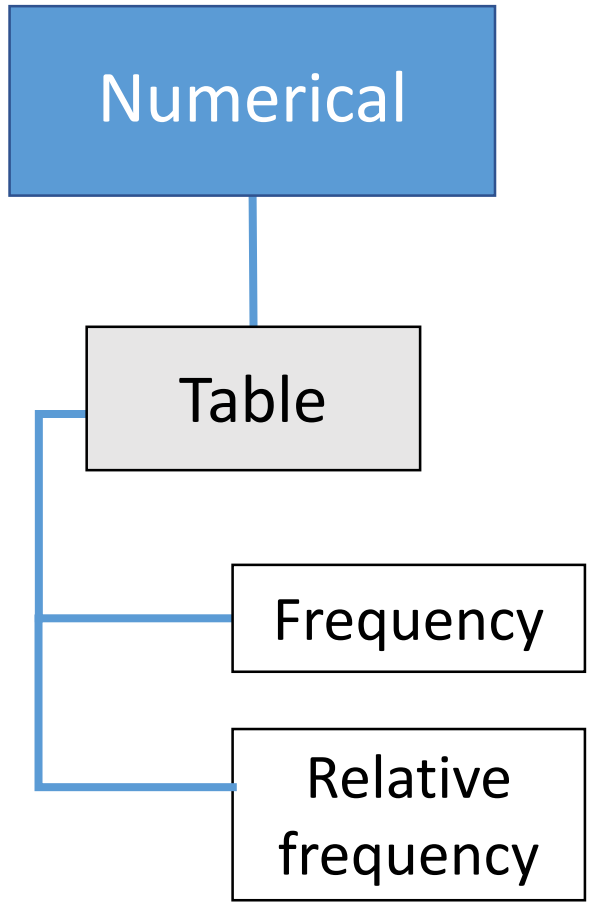
Pie Chart

- Frequency in each category is represented by percentage of a circle
- Especially useful when you have a very small amount of space to put a graph in (no axes, axis labels, etc.)



Summaries of Binary Data





Tables

- Tables are a very common way to summarize a binary variable
- Typically includes frequencies and relative frequencies
  - **Frequency** is a count or the number in each category
  - **Relative frequency** is the percentage in each category

Law		Frequency	Relative Frequency
Yes	1	28	54.9%
No	0	23	45.1%
		51	100.0%

Table 1: Summary of 51 states (including Washington, DC) by whether they have a law prohibiting smoking in restaurants, 2009.

Tables

- A two-way or 2x2 table is used to summarize two binary variables simultaneously
- In a 2x2 table, relative frequencies (percents) can be calculated by row, by column, or overall

Frequencies (counts):

toxemia	Female	Male
No	45	34
Yes	11	10

Overall percentages:

toxemia	Female	Male
No	0.45	0.34
Yes	0.11	0.10

Row percentages:

toxemia	Female	Male
No	0.5696203	0.4303797
Yes	0.5238095	0.4761905

Column percentages:

toxemia	Female	Male
No	0.8035714	0.7727273
Yes	0.1964286	0.2272727

Example: Toxemia Diagnosis

- What percentage of male infants have a mother with a toxemia diagnosis?
  22.7%
- What percentage of infants who have a mother with a toxemia diagnosis are male?
  47.6%
- What percentage of infants are male and have mother with a toxemia diagnosis?
  10.0%

Overall percentages:

toxemia	sex Female	Male
No	0.45	0.34
Yes	0.11	0.10

Row percentages:

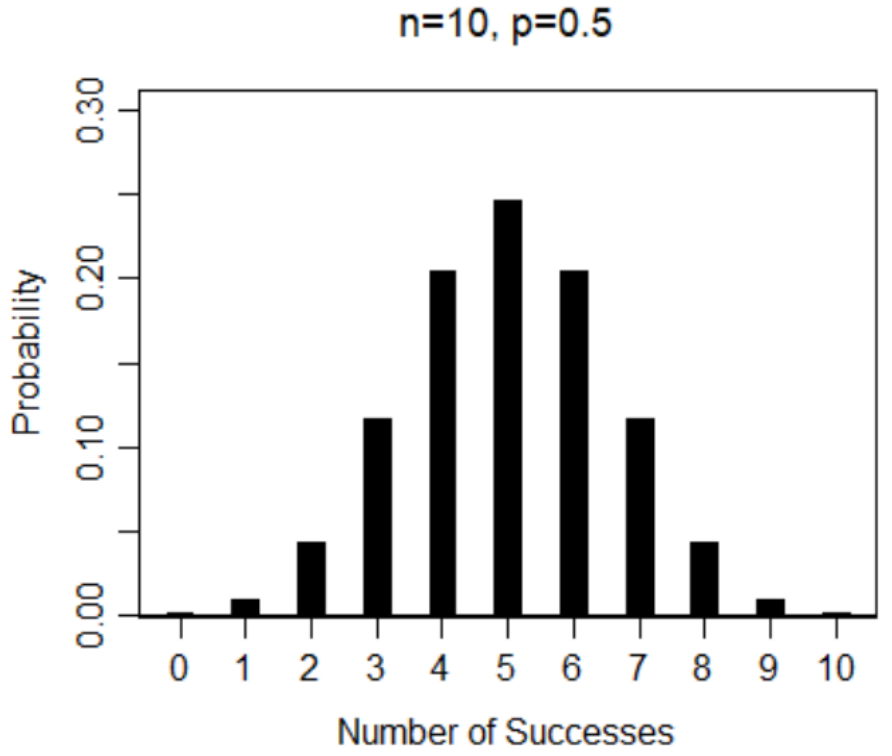
toxemia	sex Female	Male
No	0.5696203	0.4303797
Yes	0.5238095	0.4761905

Column percentages:

toxemia	sex Female	Male
No	0.8035714	0.7727273
Yes	0.1964286	0.2272727

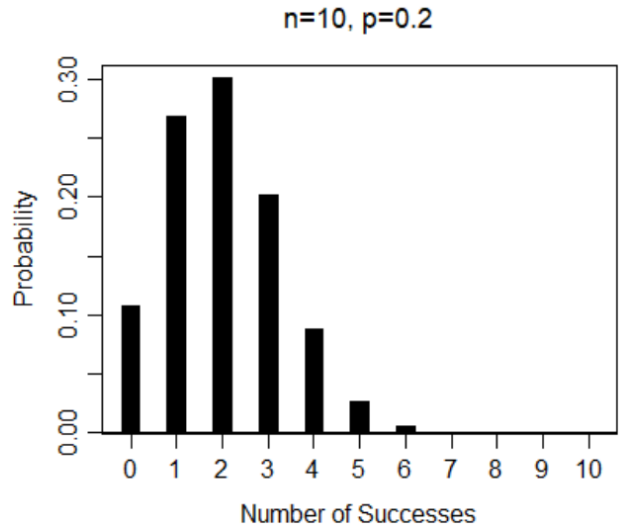
Binomial Distribution

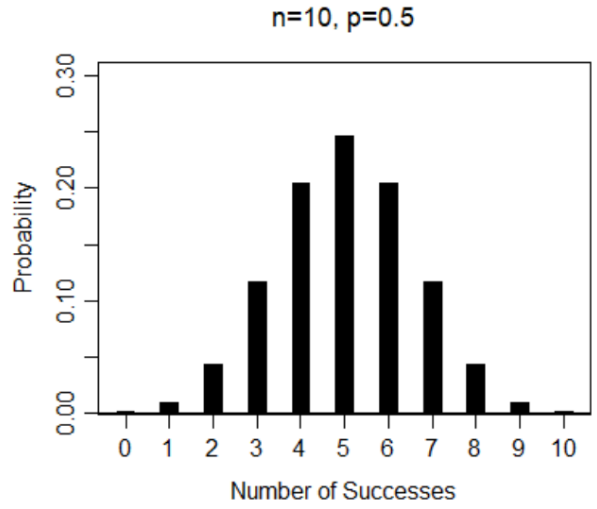
- Situation of interest:
  - Binary variable (two categories, let's call them "success" and "failure")
  - n = 10 subjects
  - Probability of "success" = 0.5
  - How many subjects do we expect to have a "success" (out of 10)?
- **Binomial distribution** describes the probability of each possible number of "successes"

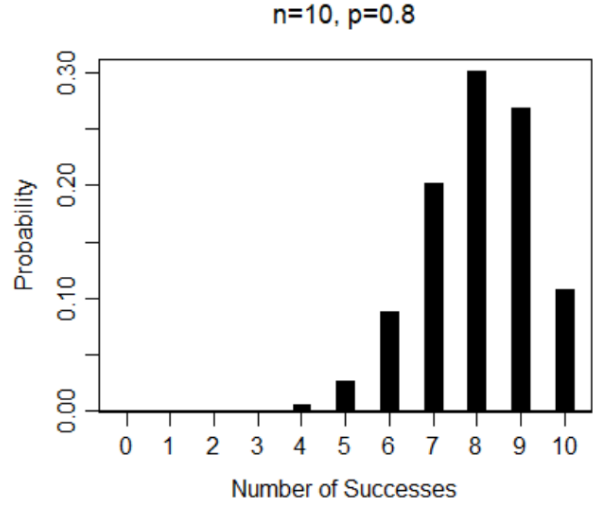


Binomial Distribution

- Shape of binomial distribution changes based on n (number of subjects) and p (probability of being in the category of interest)
- If n is big enough and p isn't too small or too large, we can use the normal distribution as an approximation
  - Rule of thumb: n*p*(1-p) ≥ 5







Binomial Distribution

- Interest lies in p, the probability of being in the category of interest
- We'll talk about estimation and inference for p
  - Relies on binomial distribution or normal approximation of the binomial distribution

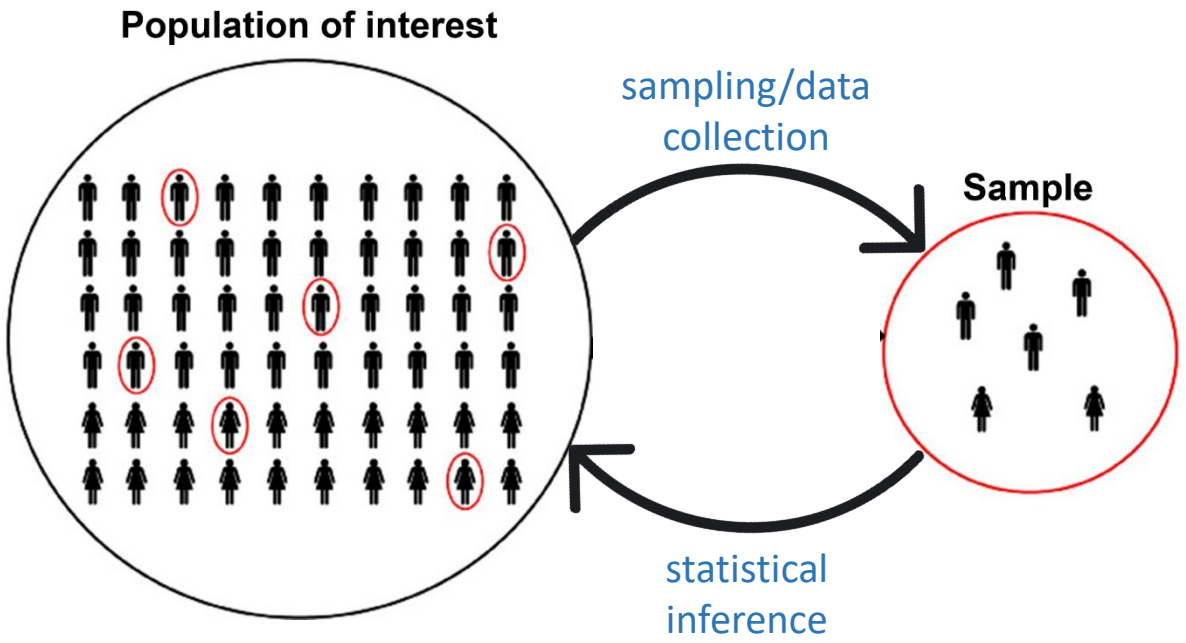
Binary Variable

- Consider a binary variable
  - Categories = “Category 1”, “Category 2”
- How can we best describe that variable with a single number?
  - With a continuous variable, we used the mean
  - We could report the number of subjects in each category, but that’s 2 numbers
  - What about the proportion of subjects in the first category? That gives us information about the second category “for free”.
- Let p be the probability of being in Category 1 (usually called “success”)
  - Or you can think of p as the proportion of subjects in Category 1
  - $1 - p$ is the probability of being in Category 2

Notation

- p = population proportion
  (of subjects in the category of interest)
- $\hat{p}$ = sample proportion
  (of subjects in the category of interest)

p is fixed



$\hat{p}$ varies (would be different with a different sample of the population)

* p is not to be confused with p-value!

Defining and Estimating p

- Example: In a study of 50 adults with leukemia, 61.4% survived at least five years after diagnosis.

  Let p be the probability of surviving at least 5 years after diagnosis in adults with leukemia (population)

  p hat=0.614

- Example: In a sample of 400 children whose mothers smoke (but they do not smoke themselves), 42 reported episodes of wheezing.

  Let p be the proportion of children experiencing wheezing among those whose mothers smoke (population)

  p hat = 42/400 = 0.105

Defining and Estimating p

- Example: A study of 30 veterans found that 10 have been diagnosed with PTSD. We are interested in estimating the probability that a veteran will not be diagnosed with PTSD.

  Let p be the probability of not being diagnosed with PTSD in veterans

  p hat = 20/30 = 0.6667

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: Germinal Matrix Hemorrhage

- Estimate the probability of germinal matrix hemorrhage among all low birth weight infants.

  Let p be the probability of germinal matrix hemorrhage among all low birth weight infants (population)

Example: Germinal Matrix Hemorrhage

Counts for hemorrhage variable:

No	Yes
85	15

Percentages for hemorrhage variable:

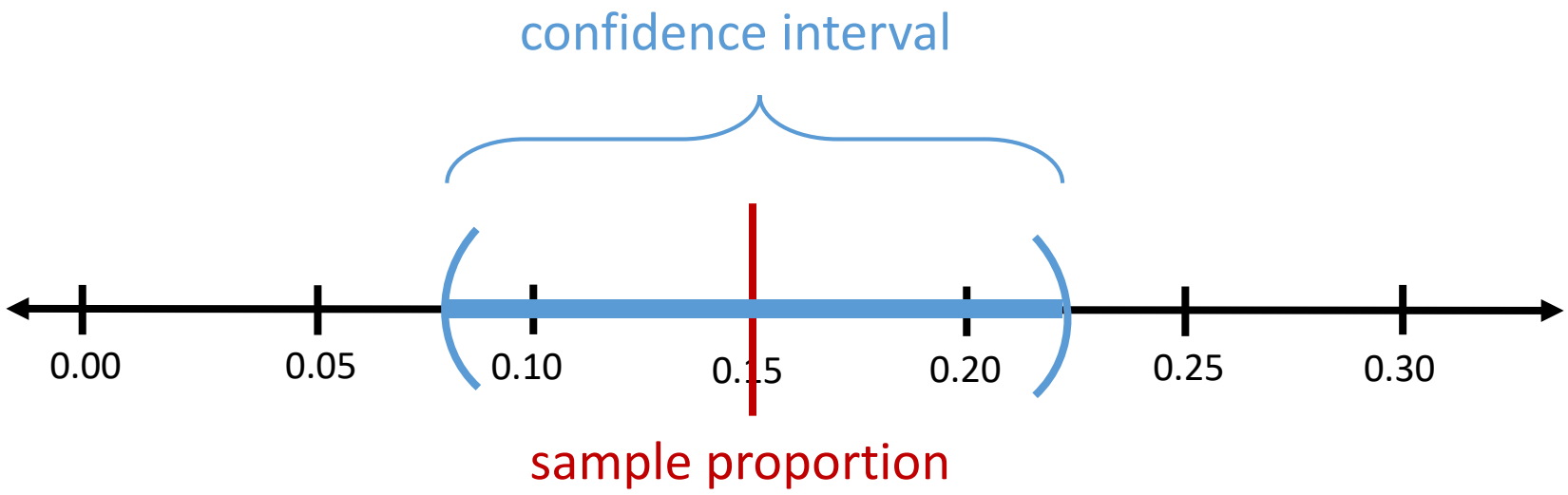
No	Yes
0.85	0.15

$$\hat{p} = \frac{15}{100} = 0.15$$

Our best estimate of p (the probability of germinal matrix hemorrhage among all low birth weight infants) is $\hat{p}$ (the probability of germinal matrix hemorrhage in our sample). We have that $\hat{p} = 0.15$ or 15%.

Confidence Interval

- Instead of just giving an estimate of the population proportion, we can give an interval that we're pretty sure includes the population proportion



Confidence Intervals for p

- There are two methods that can be used to make a confidence interval for p:
  - Exact binomial
  - Normal approximation to the binomial

- Normal approximation for the confidence interval is only appropriate when $n\hat{p}(1-\hat{p}) \geq 5$
- Exact binomial confidence interval is always correct/appropriate
- Normal approximation used to be popular when high computational time prohibited use of exact binomial. Since this typically isn't an issue now, just use the exact binomial method in your statistical software.

Example: Germinal Matrix Hemorrhage

- Calculate and interpret a 95% confidence interval for the probability of germinal matrix hemorrhage among all low birth weight infants.

Let p be the probability of germinal matrix hemorrhage among all low birth weight infants.

Example: Germinal Matrix Hemorrhage

95% exact binomial confidence interval for probability of germinal matrix hemorrhage:

```
[1] 0.08645439 0.23530750
attr(,"conf.level")
[1] 0.95
```

95% CI for p: (0.086, 0.235)

We are 95% confident that the probability of germinal matrix hemorrhage among all low birth weight infants is between 8.6% and 23.5%.

Example: Germinal Matrix Hemorrhage

Would it have been appropriate to use the normal approximation of the binomial distribution to calculate the CI?

$$n\hat{p}(1-\hat{p}) = 100(0.15)(0.85) = 12.75$$

Since 12.75 > 5, yes, it is appropriate to use the normal approximation.

Example: Germinal Matrix Hemorrhage

95% confidence interval (using normal approximation) for probability of germinal matrix hemorrhage:

```
[1] 0.0891491 0.2385308
attr(,"conf.level")
[1] 0.95
```

95% CI for p: (0.089, 0.238)

We are 95% confident that the probability of germinal matrix hemorrhage among all low birth weight infants is between 8.9% and 23.8%.

Plausible Values

- The CI contains what we consider to be plausible values for the population proportion
- Gives us a standardized way to answer the question, “Is the proportion in the population equal to ___?”
- If the value of interest is not in the CI, we have evidence that the answer is “No”.

Example: Toxemia Diagnosis

- Is the proportion of toxemia diagnoses among mothers of low birth rate infants equal to 7%, which is the rate of toxemia diagnoses among mothers of normal birth weight infants?

  Let p be the proportion of toxemia diagnoses among mothers of low birth weight infants

Example: Toxemia Diagnosis

95% exact binomial confidence interval for probability of toxemia diagnosis among mothers of LBW infants:

```
[1] 0.1349437 0.3029154
attr(,"conf.level")
[1] 0.95
```

95% CI for p: (0.135, 0.303)

Since the 95% confidence interval for the probability of having a toxemia diagnosis among mothers of LBW infants does NOT include 0.07, we have evidence to suggest that the probability of toxemia diagnosis in this group is different from that among mothers of normal birth weight infants.

Important Points

- Graphical summaries for a binary variable
- Numerical summaries for a binary variable
- Binomial distribution is the foundation of our estimation and inference for p
- Sample proportion is used to estimate population proportion
- Two methods for calculating confidence interval for a proportion
  - What are they?
  - When is each method appropriate?
- Interpretation of confidence interval for a proportion