Proportions
A *proportion* is the fraction of individuals having a particular attribute.
Example:
2092 adult passengers on the Titanic;

654 survived

Proportion of survivors = \( \frac{654}{2092} \approx 0.3 \)
Probability that two out of three randomly chosen passengers survived the Titanic
Binomial distribution

The *binomial distribution* describes the probability of a given number of "successes" from a fixed number of independent trials.
\( n \) trials; \( p \) probability of success

\[
\Pr[X] = \binom{n}{X} p^X (1 - p)^{n-X}
\]

\[
\binom{n}{X} = \frac{n!}{X!(n-X)!}
\]
n! = n × n-1 × n-2 × ... 3 × 2 × 1

6! = 6 × 5 × 4 × 3 × 2 × 1 = 720

0! = 1
1! = 1
Probability that two out of three randomly chosen passengers survived the Titanic

\[ Pr[2] = \binom{3}{2}(0.3)^2(1 - 0.3)^{3-2} \]

\[ = \frac{3!}{2! \times 1!}(0.3)^2(0.7)^1 \]

\[ = 3(0.3)^2(0.7) = 0.189 \]
Probability that two out of three randomly chosen passengers survived the Titanic

\[ \Pr[2] = \binom{3}{2}(0.3)^2(1 - 0.3)^{3-2} \]

Number of ways to get 2 survivors out of 3 passengers

Probability of 2 survivors

Probability of 1 death
Example: Paradise flycatchers

A population of paradise flycatchers has 80% brown males and 20% white. Your field assistant captures 5 male flycatchers at random.

What is the chance that 3 of those are brown and 2 are white?
\[ p = 0.8 \quad n = 5 \quad X = 3 \]

\[
\Pr[3] = \binom{5}{3} 0.8^3 (1 - 0.8)^{5-3} = \frac{120}{6 \times 2} 0.8^3 (0.2)^2 = 0.205
\]
Try at home:

What is the probability that 3 or more are brown?
Try at home:

What is the probability that 3 or more are brown?

\[
\Pr[3 \text{ or more are brown}] = \Pr[3] + \Pr[4] + \Pr[5] \\
= 0.205 + 0.410 + 0.328 \\
= 0.943
\]
Properties of the binomial distribution:
Number of successes

\[ \mu = np \]

\[ \sigma^2 = np(1 - p) \]
Binomial distribution for $p = 0.8$, $n = 5$

$\mu = np$
Proportion of successes in a sample

\[ \hat{p} = \frac{X}{n} \]

The hat (^) shows that this is an estimate of \( p \).
Properties of sample proportions

Mean: $p$

Variance: $\frac{p(1-p)}{n}$
Estimating a proportion

\[ \hat{p} = \frac{X}{n} \]
Standard error of the estimate of a proportion

\[ SE[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \]
A larger sample has a lower standard error
The law of large numbers

The greater the sample size, the closer an estimate of a proportion is likely to be to its true value.
95% confidence interval for a proportion

\[ p' = \frac{X + 2}{n + 4} \]

\[ \left( p' - 1.96 \sqrt{\frac{p'(1-p')}{n+4}} \right) \leq p \leq \left( p' + 1.96 \sqrt{\frac{p'(1-p')}{n+4}} \right) \]

This is the Agresti-Coull confidence interval
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists?
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists?

\[ \hat{p} = \frac{30}{87}, \text{ or } 0.345 \]
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists? What is the 95% confidence interval for this estimate?

\[ \hat{p} = \frac{30}{87}, \text{ or } 0.345 \]
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists? What is the 95% confidence interval for this estimate?

\[ p' = \frac{X + 2}{n + 4} = \frac{30 + 2}{87 + 4} = 0.352 \]
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists? What is the 95% confidence interval for this estimate?

\[
p' = \frac{X + 2}{n + 4} = \frac{30 + 2}{87 + 4} = 0.352
\]

\[
p' \pm Z\sqrt{\frac{p'(1 - p')}{n + 4}} = 0.352 \pm 1.96\sqrt{\frac{0.352(1 - 0.352)}{87 + 4}}
\]

\[
= 0.352 \pm 0.098
\]
Example: The daughters of radiologists

30 out of 87 offspring of male radiologists are males, and the rest female. What is the best estimate of the proportion of sons among radiologists? What is the 95% confidence interval for this estimate?

\[ p' = \frac{X + 2}{n + 4} = \frac{30 + 2}{87 + 4} = 0.352 \]

\[ p' \pm Z \sqrt{\frac{p'(1 - p')}{n + 4}} = 0.352 \pm 1.96 \sqrt{\frac{0.352(1 - 0.352)}{87 + 4}} \]

\[ = 0.352 \pm 0.098 \]

\[ 0.254 < p < 0.450 \]
Hypothesis testing on proportions

The binomial test
The *binomial test* uses data to test whether a population proportion \( p \) matches a null expectation for the proportion.

- \( H_0 \): The relative frequency of successes in the population is \( p_0 \).
- \( H_A \): The relative frequency of successes in the population is not \( p_0 \).
Example

An example: Imagine a student takes a multiple choice test before starting a statistics class. Each of the 10 questions on the test have 5 possible answers, only one of which is correct. This student gets 4 answers right. Can we deduce from this that this student knows anything at all about statistics?
Hypotheses

$H_0$: Student got correct answers randomly.

$H_A$: Student got more answers correct than random.

This is properly a one-tailed test.
Hypotheses

\( H_0: \) Student got correct answers randomly.

\( H_0: p_0 = 0.2. \)

\( H_A: \) Student got more answers correct than random.

\( H_A: p_0 > 0.2 \)
\[ N = 10, \ p_0 = 0.2 \]

\[ P = \text{Pr}[4] + \text{Pr}[5] + \text{Pr}[6] + \ldots + \text{Pr}[10] \]
\[ = \binom{10}{4}(0.2)^4(0.8)^6 + \binom{10}{5}(0.2)^5(0.8)^5 + \binom{10}{6}(0.2)^6(0.8)^4 + \ldots \]
\[ = 0.12 \]

Note: The capital \( P \) here is used for the \( P \)-value, in contrast to the population proportion with a small \( p \).
$P = 0.12$

This is greater than the $\alpha$ value of 0.05, so we would *not* reject the null hypothesis.

*It is plausible that the student had four answers correct just by guessing randomly.*