Chapter 7:
Analyzing proportions

1

Proportion

- Proportion of observations in a given category

$$\hat{p} = \frac{num.\,in\,category}{n}$$

- Ranges from 0 to 1

- Examples:
  - Proportion of ALS patients that survive at least 10 years
  - Proportion of smokers that develop lung cancer
  - Proportion of a population that is female

2

Binomial distribution

- In last chapter the null distribution of a proportion was obtained with a vast number of random samples
- More efficient method is to use the **binomial distribution**
- Individuals/observations fall into **two mutually exclusive** categories: successes and failures
  - Ex: is toad right-handed; is coin flip heads; does patient have resting heart rate > 60 bpm
- The **binomial distribution** provides the probability distribution for the number of "successes" in a fixed number of independent trials, when the probability of the success is the same in each trial

3

Binomial distribution

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

n choose *X: number of unique sequences that result in X successes*

$$\binom{n}{X} = \frac{n!}{X!\,(n-X)!}$$

$$n! = n \times (n-1) \times (n-2) \times (n-3) \times \cdots \times 2 \times 1$$

4

Binomial distribution

$$\binom{n}{X} = \frac{n!}{X!\,(n-X)!}$$

$$\binom{5}{3} = \frac{5\times4\times3\times2\times1}{(3\times2\times1)(2\times1)} = \frac{120}{(6)(2)} = 10$$

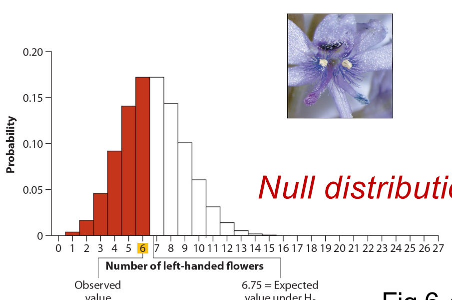
5 choose *3: number of unique sequences that result in 3 successes*

```
10 possible outcomes?

11100
11010
10110
01110
11001
10101
01101
10011
01011
00111
```

5

Chap 6: From vast number of computer draws



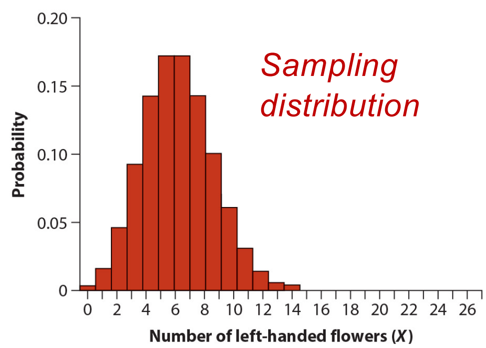
Null distribution

Fig 6-4-1

TABLE 7.1-1 The probability of obtaining *X* left-handed flowers out of *n* = 27 randomly sampled, if the proportion of left-handed plants in the population is 0.25.

Number of left-handed flowers (*X*)	Pr[*X*]
0	4.2×10^{-4}
1	0.0038
2	0.0165
3	0.0459
4	0.0917
5	0.1406
6	0.1719
7	0.1719
8	0.1432
9	0.1008
10	0.0605
11	0.0312
12	0.0138
13	0.0053
14	0.0018
15	5.1×10^{-4}
16	1.3×10^{-4}
17	2.8×10^{-5}
18	5.1×10^{-6}
19	8.1×10^{-7}
20	1.1×10^{-7}
21	1.2×10^{-8}
22	1.1×10^{-9}
23	7.9×10^{-11}
24	4.4×10^{-12}
25	1.8×10^{-13}
26	4.5×10^{-15}
27	5.5×10^{-17}

Chap 7: From binomial distribution



Sampling distribution

Fig 7-1.1

6

Sampling the distribution of a proportion

- p is the "real" proportion of the population (parameter)
- $\hat{p}$ is the estimated proportion from a sample (estimate/statistic)
- Sample size matters
- Suppose $p = 0.25$, the shape of the sampling distribution of $\hat{p}$ depends on n
- Larger n provides more precise estimates

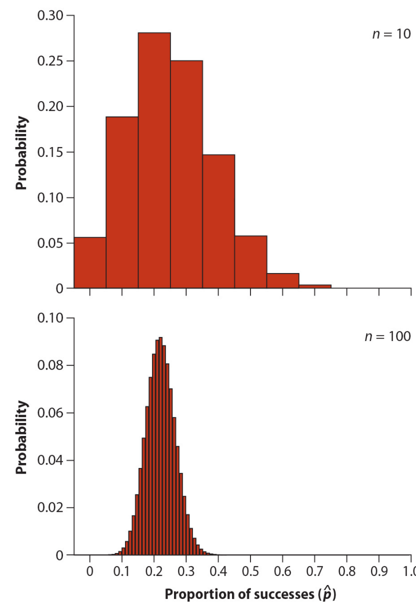


Fig 7-1.2

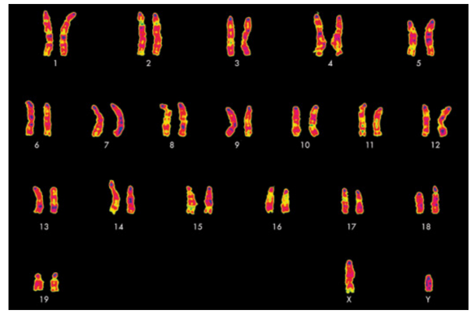
7

Binomial test

- The **binomial test** uses data to test whether a population proportion (p) matches a null expectation (p_0) for the population
- 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

8

Ex 7.2: Sex and the X



- Theory that spermatogenesis (sperm formation) genes should occur more often on the X chrom
- Identified 25 spermatogenesis genes, found that 10 (40%) were on the X chromosome
- If genes were distributed randomly, you'd expect 6.1%
  - Because X chrom contains 6.1% of all genes in genome

9

Ex 7.2: Sex and the X

- H_0: probability that spermatogenesis gene occurs on X chromosome is $p = 0.061$
- H_A: probability that spermatogenesis gene occurs on X chromosome is not 0.061 ($p \neq 0.061$)
- Under null and sample of 25 spermatogenesis gene:
  - Expect 25 x 0.061 = 1.525
- Is observed number of 10 different enough to reject the null hypothesis?

10

Binomial test

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

- Use binomial distribution to calculate probability of getting ≥ 10 successes in 25 samples with *p* = 0.061
- Sum these probabilities and multiply by 2 (for two-sided test)
- *P* = 2 x Pr[*number of successes ≥ 10*]
- *P* = 1.98 x 10^{-6}

TABLE 7.2-1 Probabilities in the right-hand tail of the binomial distribution with $n = 25$ and $p = 0.061$.

Number of genes on X	Probability under the null hypothesis
10	9.1×10^{-7}
11	8.0×10^{-8}
12	6.1×10^{-9}
13	4.0×10^{-10}
14	2.2×10^{-11}
15	1.0×10^{-12}
16	4.3×10^{-14}
17	1.5×10^{-15}
18	4.2×10^{-17}
19	1.0×10^{-18}
20	2.0×10^{-20}
21	3.1×10^{-22}
22	3.6×10^{-24}
23	3.1×10^{-26}
24	1.7×10^{-28}
25	4.3×10^{-31}

11

12

Estimating proportions with uncertainty

- Recall that $\hat{p}$ is the sample estimate and *p* is the (true) population proportion
- The **standard error of a proportion** tells you the precision (uncertainty) of the estimate
- The 95% confidence interval of $\hat{p}$: 95% confident that *p* is between lower and upper limits
  - Offers another method to analyze the data (other than hypothesis testing)

13

Ex 7.3: She-turtles

- Sex determination in sea turtles is determined by the temperature of incubated eggs
  - Warmer = female, cooler = male
- Historic sex ratio is ~ 50:50, but increasing temps could be causing skew toward females
- In a sample of 169 juvenile green sea turtles, 38 were male
  - $\hat{p} = 131/169 = 0.775$ (proportion females in sample)

14

Standard error of proportion

- Calculation for the standard error of a proportion

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

But p is the population parameter, which is typically unknown

- So the standard error of a proportion needs to be estimated using the sample proportion

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

15

Standard error of proportion

- In a sample of 169 juvenile green sea turtles, 38 were male and 131 were female
  - $\hat{p} = 131/169 = 0.775$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.775(1-0.775)}{169}} = 0.001$$

16

Confidence interval of proportion

- Different methods of calculation exist
- Textbook recommends Agresti-Coull method

intermediate calculation

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

$$p' - 1.96\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + 1.96\sqrt{\frac{p'(1-p')}{n+4}}$$

X: number successes
n: sample size

$$0.706 < p < 0.832$$

17

Confidence interval of proportion

- 95% confidence interval for proportion of females:
  - $0.706 < p < 0.832$
- Interval does **not** include the null proportion of 0.5, and in fact is well above the null
- Data are *inconsistent* with the null
  - Can be confident that the population proportion of females is much higher than 0.5
- Confidence intervals provide an alternative to hypothesis testing

18

Notes

- Skipping section 7.4

19