# Chapter 9

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

# Dichotomous observations

- In Chapters 5, 6, 7, and 8 we have dealt with continuous observations $Y$.
- With continuous observations it is natural to look at the mean or median.
- Now we'll consider *categorical data*.
- The simplest categorical data simply tells us which of two categories a subject is in, e.g. Male or Female, Diseased or Non-diseased, $\geq 50$ years or $< 50$ years, etc.
- This type of data is called *dichotomous* or *binary*.
- Now we'll concentrate on the proportion of successes $p$.

# Testing $H_0 : p = p_0$

- Pick one of the two categories to be "success." Then have a population with an unknown proportion $p$ of success.
- We often want to test $H_0 : p = p_0$ where $p_0$ is a known value.
- If we take a random sample on $n$ individuals and count $Y$, the number of "successes" then $Y \sim \text{binom}(n, p_0)$ if $H_0 : p = p_0$ is true.
- The sample proportion $\hat{p} = Y/n$ estimates the true population proportion. If $\hat{p}$ is far away from $p_0$ then we have evidence of $p \neq p_0$.

# Example 9.4.9 Harvest Moon Festival

- Can people close to death postpone dying until after a meaningful event?
- Researchers studied death from natural causes among Chinese women over age 75 living in California. During Harvest Moon oldest woman in the family has important role.
- Researchers found $y = 33$ deaths in the week preceding Harvest Moon & 70 after, so $n = 33 + 70 = 103$.
- If people *cannot* postpone, then probability of dying before is same as dying after, 50%.

# P-value for Harvest Moon Festival data

- Let $p$ be proportion that die before Harvest moon. Want to test $H_0 : p = 0.5$ vs. $H_A : p < 0.5$.
- Here $\hat{p} = \frac{3}{103} = 0.32$, smaller than the hypothesized value 0.5, so there's some evidence toward $H_A$.
- The P-value is computed

$$\text{P-value} = \Pr\{Y/n \leq \hat{p} | p = p_0\} = \Pr\{Y \leq 33 | p = 0.5\},$$

the probability of seeing a $\hat{p}$ even less than 0.32, given $H_0$ is true.

- Since $Y \sim \text{binom}(103, 0.5)$ under $H_0$, we can find

$$\text{P-value} = \sum_{j=0}^{33} \Pr\{Y = j\} = 0.0001705.$$

# P-value from `binom.test`

R's function `binom.test` computes P-values for tests of $H_0 : p = p_0$ vs. one of (a) $H_A : p \neq p_0$, (b) $H_A : p < p_0$, or (c) $H_A : p > p_0$.

```
> binom.test(33,103,p=0.5,alternative="less")

        Exact binomial test

data:  33 and 103
number of successes = 33, number of trials = 103, p-value = 0.0001705
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4041263
sample estimates:
probability of success
             0.3203883
```

Since P-value $= 0.00017 < 0.05$ we reject $H_0 : p = 0.5$ in favor of $H_A : p < 0.5$. There is statistically significant evidence that people can postpone dying.

# 9.2 Confidence intervals

- Your book discusses one method for obtaining confidence intervals for an unknown $p$; a somewhat cruder, but similar method is carried out in R by `prop.test(y,n)`.
- There are packages available that will compute the book's interval.
- `binom.test(y,n)` implements the method that I would recommend, so use `binom.test(y,n)` instead.
- `binom.test(y,n)` uses the relationship between confidence intervals and hypothesis tests to compute an exact confidence interval; other confidence intervals (including your book's) are approximate.

# How `binom.test` computes confidence intervals

- For a given $y$ and $n$, R finds all $p_0$ such that $H_0 : p = p_0$ *is not rejected* vs. $H_A : p \neq p_0$ at the 5% level. The set of these $p_0$'s is the 95% confidence interval for $p$.
- It's much more work than your book's method, but it's extremely accurate.

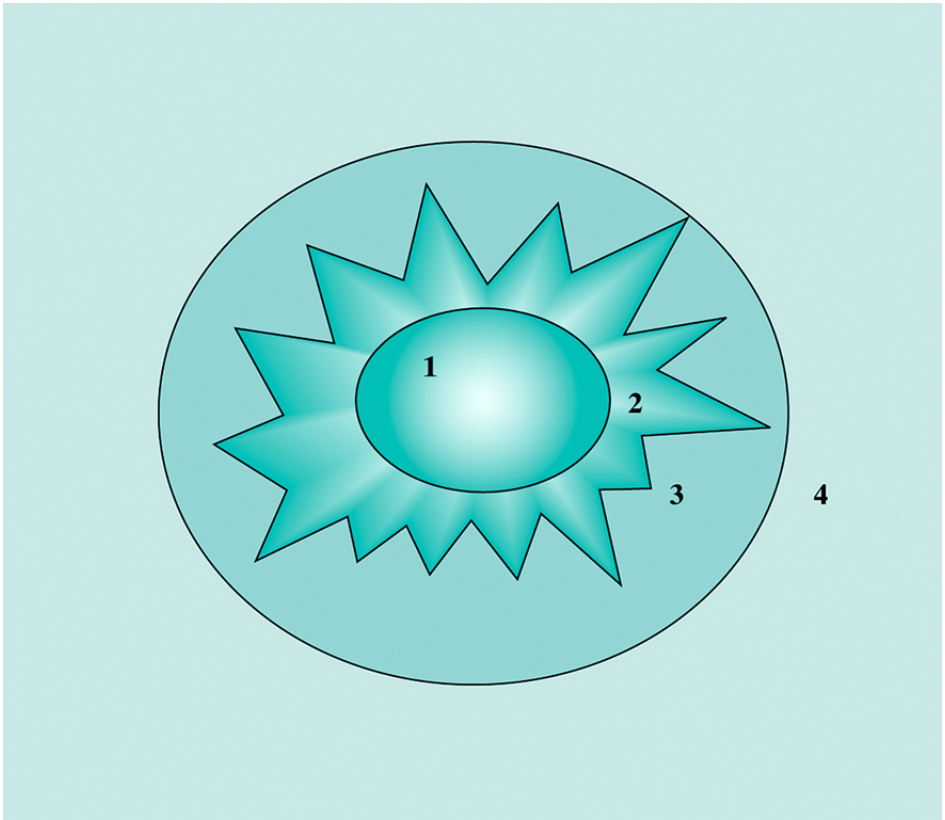
```
> binom.test(33,103)
95 percent confidence interval:
 0.2318410 0.4195741
```

We are 95% confident that the true probability of dying right before Harvest Moon (vs. right after) is between 23% and 42%. There is evidence that people "wait to die."

# 9.4 $\chi^2$ goodness-of-fit test

- So far we have considered binary variables with unknown population $p = \Pr\{\text{success}\}$.
- Specifically, we tested $H_0 : p = p_0$ where $p_0$ is known.
- This is now generalized categorical variables with *more than two categories*.
- Example 9.4.1 Deer habitat and fire. Does fire affect deer behavior? After a fire burned 730 acres, researchers surveyed the 3,000 acres surrounding the area, divided into four regions.

# Schematic of 3000-acre parcel w/ interior 730-acre fire




(1) the region near the heat of the burn, (2) the inside edge of the burn, (3) the outside edge of the burn, (4) the area outside of the burned area.

# Null and alternative hypotheses

- $H_0$ : deer show no preference to any burned/unburned habitat, i.e. they are randomly distributed over the 3,000 acres.
- $H_A$ : deer have a preference for some of the regions.
- Under the null hypothesis, the probabilities of deer in the regions are proportional to the sizes of the regions.

**Table 9.4.1** Deer distribution

| Region | Acres | Proportion |
|---|---|---|
| 1. Inner burn | 520 | 0.173 |
| 2. Inner edge | 210 | 0.070 |
| 3. Outer edge | 240 | 0.080 |
| 4. Outer unburned | 2,030 | 0.677 |
| | 3,000 | 1.000 |

# Observed vs. expected

- The researchers found 75 deer in the whole region. Before we find out how they were distributed across the 4 regions, what do we *expect* to see?
- If 17.3% of the deer are in the inner burn, then we expect $0.173(75) = 13.00$ of the 75 to be found there.
- We expect $0.070(75) = 5.25$ of the 75 in the inner edge.
- We expect $0.080(75) = 6.00$ of the 75 in the outer edge.
- We expect $0.677(75) = 50.75$ of the 75 in the outer unburned.
- The numbers *actually observed* are 2, 12, 18, and 43.

# Observed vs. expected

| Region | $H_0$ proportion | Observed proportion | Expected under $H_0$ | Observed number |
|---|---|---|---|---|
| Inner burn | 0.173 | 0.027 | 13.00 | 2 |
| Inner edge | 0.070 | 0.16 | 5.25 | 12 |
| Outer edge | 0.080 | 0.24 | 6.00 | 18 |
| Outer unburned | 0.677 | 0.57 | 50.75 | 43 |

# Chi-square statistic

Let $e_i$ be the expected number in each category and let $o_i$ be the observed number. The chi-square goodness of fit test statistic is

$$\chi^2_S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

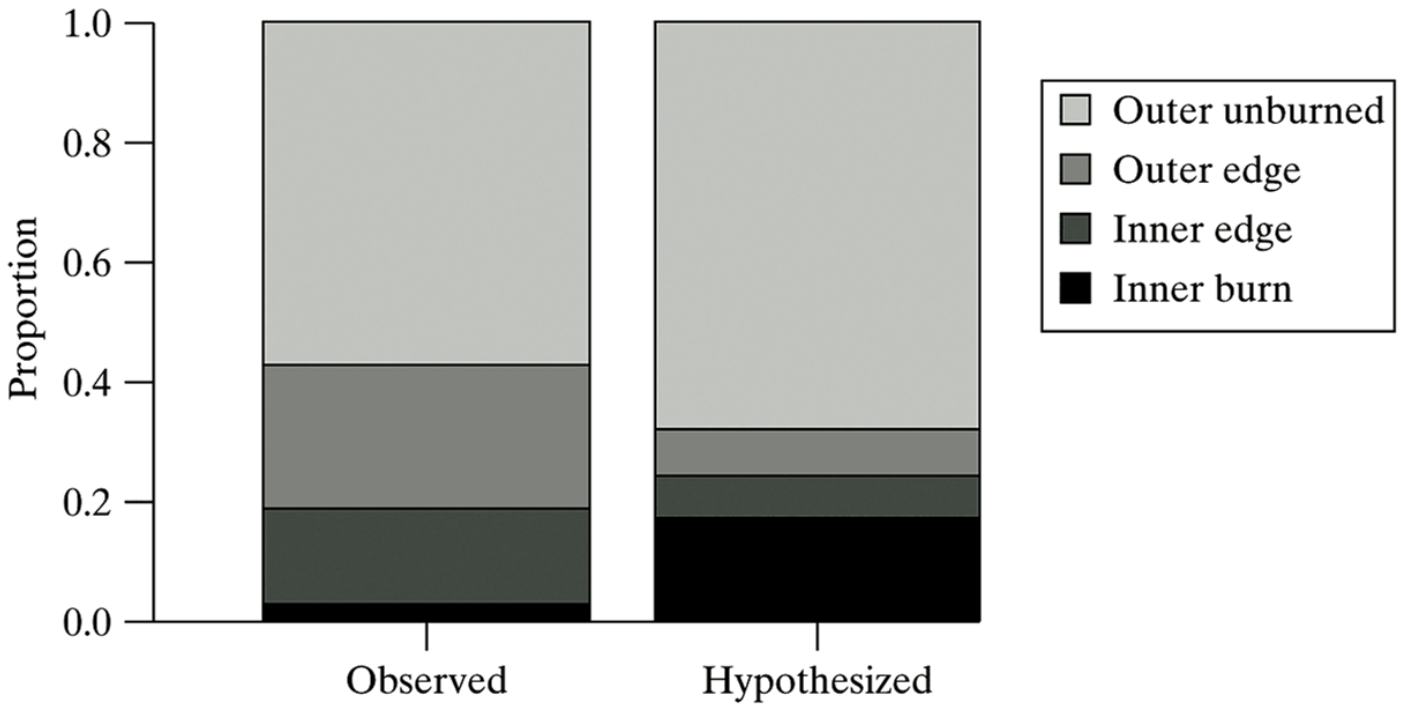
where the summation is over all $k$ categories.

---

For the deer data

$$\chi^2_S = \frac{(2-13)^2}{13} + \frac{(12-5.25)^2}{5.25} + \frac{(18-6)^2}{6} + \frac{(43-50.75)^2}{50.75} = 43.2.$$

# Deer proportions




If the gray-shaded areas have markedly different sizes across hypothesized and observed, then the $\chi^2_S$ will be large. If they are perfectly equal then $\chi^2_S = 0$.

# $\chi^2$ distribution

- If $H_0$ is true then $\chi^2_S$ has a special distribution called the chi-square distribution (p. 352).
- Like the t distribution, the $\chi^2$ has degrees of freedom $df$ attached to it; the $df = k - 1$, the number of categories minus one.
- The P-value for testing $H_0$ is the tail probability P-value $= \Pr\{\chi^2_{df} > \chi^2_S\}$.
- R takes care of details for us with the `chisq.test(counts,p=probabilities)` command.
- `counts` is a list of the observed numbers falling into each of the $k$ categories and `probabilities` are the hypothesized probabilities under $H_0$.
- P-value is probability of seeing observed *even further away from expected*, under $H_0$, than we saw.

# R command `chisq.test`

Need to define two lists: (1) the observed counts in each category, and (2) a list of the hypothesized $H_0$ probabilities that add up to one.

```
> deer=c(2,12,18,43)
> prob=c(0.173,0.070,0.080,0.677)
> chisq.test(deer,p=prob)
        Chi-squared test for given probabilities

data:  deer
X-squared = 43.1524, df = 3, p-value = 2.284e-09
```

Since P-value $= 0.0000000023 < 0.05$, we $H_0$ at the 5% level. There is statistically significant evidence that the deer prefer some areas to others, i.e. they are not randomly scattered around the burn area. The data tell us that deer prefer the inner and outer edges to the inner burn and outer unburned areas.

# Sign test revisited

- Used on paired data; "before" and "after" measurements differenced.
- The sign test counts the number of positive differences $N_+$.
- If $H_0 : \eta_D = 0$ is true, then the probability of a positive difference is $H_0 : p = 0.5$.
- So the sign test simply tests that the proportion of "+" is 0.5, i.e. `binom.test(`$N_+$`,`$n$`)`.

# Homework problems from Chapter 9

- Use either `binom.test` for dichotomous or `chisq.test` for more than two categories.
- Dichotomous: 9.2.4, 9.2.5(a,b), 9.2.6, 9.4.4, 9.4.7.
- More than two categories 9.4.1, 9.4.3, 9.4.10.