STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 27: Models for Binary Data

Outline

- Previous
  - Exploring (normal) models for continuous data
    - Single mean
    - Two independent groups
    - Paired data
    - Multiple independent groups
    - Linear regression
- Today
  - Consider data that are not continuous
  - Explore models for binary data

Short Sighted

- Study from Australia following 1344 participants, aged 18-22.
- A total of 342 had myopia (-0.50D or worse defect)
- Assuming sample is representative, what can we learn at general prevalance of myopia in Australians in that age group?

Problem

- We have been working with models for continuous outcome variables
- This is not continuous data
- It is binary data
  - Each observation is yes/no, success/failure, 1/0
  - Each participant either myopic (‘success’), or not (‘failure’)
- Such data arises all the time
  - Will you support candidate X in the next election?
  - Did the chick successfully fledge?
  - Did the participant select option A (or B)?
  - Did the home team win the football match?
- We need a model for binary data
  - Probability distribution for binary data

Bernoulli distribution

- Recall: discrete probability distributions
- Random variable Y with two possible outcomes: success/failure
  - Represent success with 1
  - Represent failure with 0
- These two outcomes have associated probabilities
  - Earlier in semester: we assigned them actual numbers, e.g. 0.6 and 0.4
  - Now: represent the probability of success with an (unknown) parameter: p
- That gives the probability distribution

i	1	2	Total
y_i	0	1	
$\Pr(Y = y_i)$	$1 - p$	p	1

Bernoulli distribution: properties

- Recall: we found means and variances of discrete probability distributions

$$E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i)$$

$$\mathsf{Var}(Y) = \sum_{i=1}^{k} (y_i - E[Y])^2 \Pr(Y = y_i)$$

- Using these we can find the mean and variance of a Bernoulli distribution

$$E[Y] = p$$

$$\mathsf{Var}(Y) = p(1-p)$$

- Extension: Confirm these using the expectation and variance formulae above

Binary to binomial

- Typically interested in cases where there are many binary trials
  - Flip a coin 15 times
  - Record the myopia status of 1344 individuals
- The number of successes from multiple trials has a binomial distribution, if:

1. The trials are binary
   - The outcome can be represented as success / failure (or equivalent)
2. The number of trials n, is fixed
   - e.g. the number of trials does not depend on the number of successes (or failures) you see
3. The trials are independent
   - The outcome of one trial does not affect the outcome of another
4. The probability of success, p, is the same for each trial
   - The probability of success does not change from one trial to another

Binary to binomial

- Let's think about the simplest case
  - Y_1 and Y_2 are two (independent) random variables
  - Each of them has a Bernoulli distribution with probability of success p
- Our interest is in the random variable $X = Y_1 + Y_2$
  - Number of successes from two trials
- If we had a sample of 2 Australians (aged 18–22)
  - X is a random variable that represents how many of them are myopic

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1-p)^2$	$2p(1-p)$	p^2	1

$$
\begin{aligned}
Pr(X = 0) &= \Pr(Y_1 = 0 \text{ and } Y_2 = 0) \\
&= \Pr(Y_1 = 0)\Pr(Y_2 = 0) \qquad \text{multiplication rule: independence} \\
&= (1-p) \times (1-p)
\end{aligned}
$$

Binomial distribution: $n = 2$

- The probability distribution of $X = Y_1 + Y_2$ is

i	1	2	3	Total
x_i	0	1	2	
$\Pr(X = x_i)$	$(1-p)^2$	$2p(1-p)$	p^2	1

$$
\begin{aligned}
\Pr(X = 1) &= \Pr(Y_1 = 1 \text{ and } Y_2 = 0) + \Pr(Y_1 = 0 \text{ and } Y_2 = 1) \\
&= \Pr(Y_1 = 1)\Pr(Y_2 = 0) + \Pr(Y_1 = 0)\Pr(Y_2 = 1) \qquad \text{independence} \\
&= p(1-p) + (1-p)p
\end{aligned}
$$

Binomial distribution: general

- In general, the number of successes from n independent Bernoulli trials is:
  - $X = Y_1 + Y_2 + \ldots + Y_n$
- For moderate or large values of n
  - Possible, but extremely tedious, to write out full probability distribution
- We have a shortcut: we can find the probability of x successes from n independent Bernoulli trials

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways to obtain x successes from n trials[1]
- For each of these, the probability of observing those x successes is $p^x(1-p)^{n-x}$
  - E.g. there are two ways to see $x = 1$ success from $n = 2$ trials (see above)
    - Each of those has probability $p(1-p)$
  - E.g. there are 3003 ways to see $x = 5$ successes from $n = 15$ trials
    - Each of these has probability $p^5(1-p)^{10}$

[1] $x! = x \times (x-1) \times \ldots \times 3 \times 2 \times 1$, e.g. $3! = 3 \times 2 \times 1 = 6$. $x!$ is read as x factorial.

Binomial distribution: general

- The probability of x successes from n independent Bernoulli trials is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- We can use this to find the expectation and variance
  - The mean of a binomial distribution is $E[X] = np$
  - The variance of a binomial distribution $\mathsf{Var}(X) = np(1-p)$
- If there are $n = 100$ putts with probability of success $p = 0.2$, then
  - $E[X] = np = 100 \times 0.2 = 20$
  - $\mathsf{Var}(X) = np(1-p) = 100 \times 0.2 \times 0.8 = 16$
  - $\mathsf{sd}(X) = \sqrt{\mathsf{Var}(X)} = 4$

Binomial probabilities in R

- We don't have to calculate the long form of that equation
  - We can use the R function `dbinom`
- Example: what is $\Pr(X = 1)$ when $p = 0.2$ and $n = 2$

```
dbinom(x = 1, size = 2, prob = 0.2)

## [1] 0.32
```

- The arguments are:
  - `x = 1`: the number of successes x
  - `size = 2`: the number of trials n
  - `prob = 0.2`: the probability of success p
- Check that it gives the correct answer: we know it should be $2p(1-p)$

```
2*0.2*(1-0.2)

## [1] 0.32
```

More examples

- If we have sample of 15 individuals and probability myopia is 0.3 for each person:
- What is the probability that we see exactly 5 people in the sample with myopia?
- We have $x = 5$, $n = 15$, $p = 0.3$

```r
dbinom(x = 5, size = 15, prob = 0.3)

## [1] 0.2061
```

- What is the probability of getting 40 myopic individuals out of sample of size 100 if $p = 0.35$?

```r
dbinom(x = 40, size = 100, prob = 0.35)

## [1] 0.04739
```

Back to the data

- We want to estimate the probability an Australia aged 18–22 is myopic
- What is our statistical model?
  - Myopia diagnosis individual eacd individual is a Bernoulli trial with probability p
    - Assume independence between individual results
  - Equivalently, the total number of successful putts is binomially distributed
- Want to estimate a parameter (population) with a statistic (sample)
  - (Reasonably) obvious statistic: sample proportion x/n
- For myopia data:

$$\hat{p} = \frac{x}{n} = \frac{342}{1344} = 0.254$$

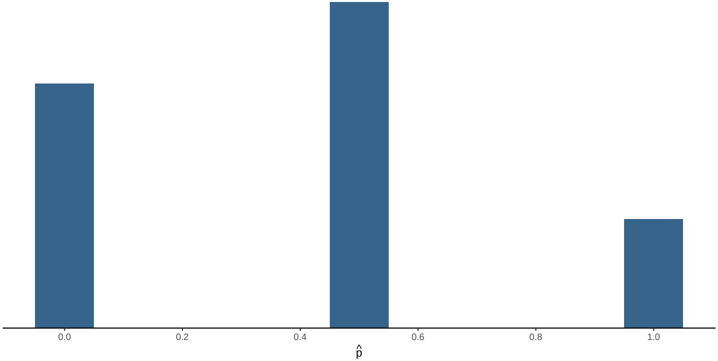
- Recall: $\hat{p}$ is the estimate of parameter p

Confidence interval

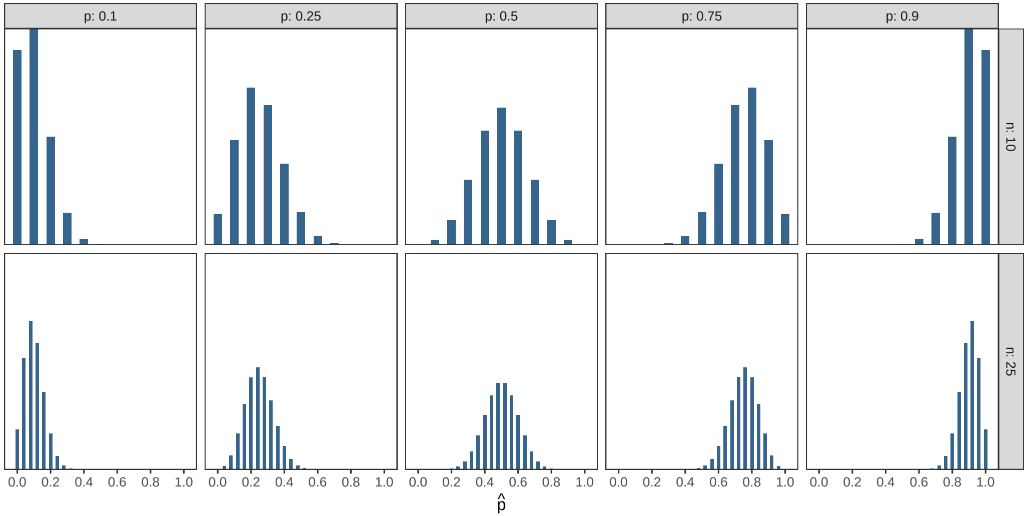
- How do we find a confidence interval?
- Recall: normal model
  - Found the sampling distribution
  - Obtained a confidence interval from the sampling distribution
- Can we do the same thing here?
  - The sampling distribution is the distribution of $\hat{p}$ if we take repeated samples
- Look at it graphically

Sampling distribution for $\hat{p}$: Start small with $n = 2$ and $p = 0.4$

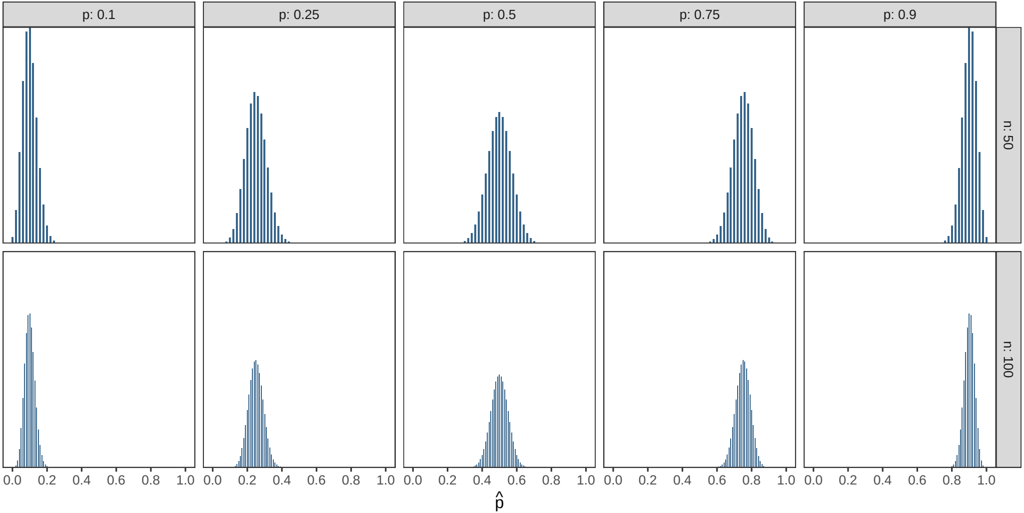
- There are three possibilities:
  - Observe $x = 0$ with probability 0.36: estimate $\hat{p} = 0$
  - Observe $x = 1$ with probability 0.48: estimate $\hat{p} = 0.5$
  - Observe $x = 2$ with probability 0.16: estimate $\hat{p} = 1$



Same principle, but increase the number of trials

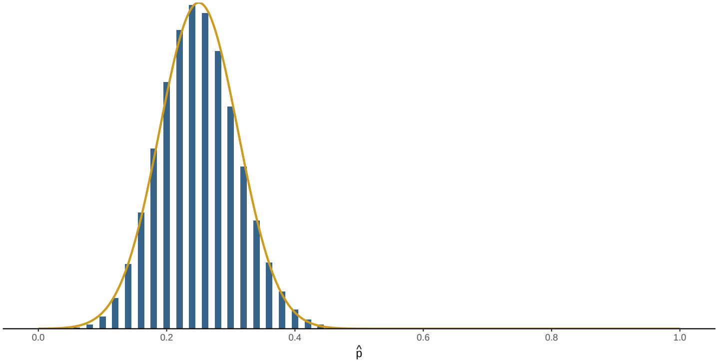


Increase the number of trials some more



Sampling distribution

- As the sample size gets larger, the sampling distribution looks increasingly normal
  - Normal pdf given in gold
- Example: $n = 50$, $p = 0.25$



Sampling distribution

- We can approximate the sampling distribution by a normal distribution
  - Provided n is large enough
- There are various rules of thumb used to determine if the normal approximation is appropriate
- One of these is
  - $np > 10$ and $n(1 - p) > 10$
- As we saw on the plots above, this reflects that
  - The sampling distribution is increasingly normal as n increases
  - When p is close to 0 or 1 it takes a larger n for it to approach normality
- In practice we use $n\hat{p}$ and $n(1 - \hat{p})$ to check if a normal approximation is reasonable

Sampling distribution

- We can approximate the sampling distribution by a normal distribution
  - Provided n is large enough
- The mean and variance are

$$E[\hat{p}] = p$$
$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

- So the standard error: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Extension: Derive $E[\hat{p}]$ and $\mathrm{Var}(\hat{p})$
  - We have $\hat{P} = \frac{X}{n}$ where $E[X] = np$ and $\mathrm{Var}(X) = np(1-p)$

Confidence interval in R

- We use the normal approximation to find a confidence interval: `prop.test`

```
n = 1344; x = 342
prop.test(x, n)

##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 323, df = 1, p-value <2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.23154 0.27881
## sample estimates:
##       p
## 0.25446
```

- We are 95% confident that the probability of myopia in a randomly sampled Australian aged 18-22 is between 0.232 and 0.279

Hypothesis test

- We can also test the hypothesis
  - $\mathsf{H}_0 : p = p_0$
  - $\mathsf{H}_A : p \neq p_0$
- `prop.test` defaults to $p_0 = 0.5$
  - It can be changed with option `p`, e.g. `p = 0.2`

Hypothesis test

R output

```
n = 1344; x = 342
prop.test(x, n, p=0.2)

##
##  1-sample proportions test with continuity correction
##
## data:  x out of n, null probability 0.2
## X-squared = 25, df = 1, p-value = 7e-07
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.232 0.279
## sample estimates:
##     p
## 0.254
```

Hypothesis test

continued

- Testing $p_0 = 0.2$ (reflects myopia in 20-year olds in UK): gives a p-value of 7×10^{-7}
- This quantifies the incompatibility between the data and null hypothesis
- Since p-value $< \alpha = 0.05$ there is (strong) evidence that the data are unusual given the null hypothesis is true
  - The data we have observed would be very unusual if the probability of myopia in Australians aged 18-22 was really 0.2.

Summary

- Introduced Bernoulli distribution for binary observations
- The number of successes from multiple binary trials have binomial distribution
  - Several conditions need to be satisfied
- Use a binomial model to find:
  - Confidence interval for p
  - Hypothesis test
    - We will look more into these in the next lecture