Chapter 5: Sampling distributions
Outline

1. Introduction

2. Sampling distribution of a proportion

3. Sampling distribution of the mean

4. Normal approximation to the binomial

5. The continuity correction
Introduction
What does it mean to take a sample of size $n$? $Y_1, \ldots, Y_n$ form a random sample if they are independent and have a common distribution.

- From a sample, we can calculate a sample statistic such as the sample mean $\bar{Y}$.
- $\bar{Y}$ is random too.
- The distribution of $\bar{Y}$ is called a sampling distribution.
Sampling Picture

- Chalk Board
- Randomness via sampling.
- \( \bar{Y}, \hat{p} \).
Sampling distribution of a proportion
Estimator for a Proportion

- A dichotomous trait occurs in a population with an unknown proportion $p$
- You wish to estimate $p$.
- Steps
  1. Collect a sample of size $n$
  2. Let $Y = \text{total observations which have the trait of interest}$.
- $Y \sim B(n, p)$ (Binomial)
- An estimator for the unknown $p$
  \[ \hat{p} = \frac{Y}{n} \]
Example

- What proportion ($p$) of the human population has type O blood (dichotomous because either have O or not).
- You randomly sample 100 people and observe 42 have type O.
- Here $n = 100$, $Y = 42$.
- An estimator for $p$?

$$\hat{p} = \frac{Y}{n} = .42$$

- What would happen if you took a different sample of 100 people? You’d get a different value of $\hat{p}$.
Sampling Distribution of $\hat{p}$

$Y \sim B(n, p)$, \hspace{1cm} \hat{p} = \frac{Y}{n}$

1. What is the distribution of $Y$: binomial
2. What is the distribution of $\hat{p}$: binomial divided by $n$. 
Chalkboard Example

- Write out the distribution of $\hat{\rho}$
- Will be a hassle if $n$ is big!
Mean and Variance of $\hat{p}$

\[ Y \sim B(n, p), \quad \hat{p} = \frac{Y}{n} \]

1. Moments of $Y$

\[ E(Y) = np, \quad \text{Var}(Y) = np(1 - p) \]

2. Moments of $\hat{p}$

\[ E(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{p(1 - p)}{n} \]
Motivate

\[ E(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{p(1 - p)}{n} \]
Example

Example: cross of two heterozygotes $Aa \times Aa$. Probability distribution of the offspring’s genotype:

<table>
<thead>
<tr>
<th>Offspring genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
</tr>
</tbody>
</table>

An offspring is dominant if it has genotype $AA$ or $Aa$.

Experiment: Get $n = 2$ offsprings, count the number $Y$ of dominant offspring, and calculate the sample proportion $\hat{p} = Y/2$.

- We would like $\hat{p}$ to be close to the “true” value $p = 0.75$
- $\hat{p}$ is random
- Distribution of $\hat{p}$ (from the binomial distribution):

<table>
<thead>
<tr>
<th>$Y$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{p}$</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>IP</td>
<td>0.0625</td>
<td>0.3750</td>
<td>0.5625</td>
</tr>
</tbody>
</table>
Example

Larger sample size: $Y =$ # of dominant offspring out of $n = 20$, $\hat{p} = Y/20$ the sample proportion.

- We still want $\hat{p}$ to be close to the “true” value $p = 0.75$
- $\hat{p}$ is still random
- What is the probability that $\hat{p}$ is within 0.05 of $p$? Translate into a binomial question

\[
\Pr\{0.70 \leq \hat{p} \leq 0.80\} = \Pr\{0.70 \leq Y/20 \leq 0.80\} \\
= \Pr\{14 \leq Y \leq 16\} \\
= \Pr\{Y = 14\} + \Pr\{Y = 15\} + \Pr\{Y = 16\} \\
= 0.56
\]

Sample size of 20 better than sample size of 2 !!
Sampling distribution of the mean
Categorical versus Numerical Variables

1. Dichotomous categorical variable (e.g. left-handedness)
   - Natural value to consider is proportion

2. Numerical variables (e.g. height)
   - Natural value to consider is the mean
A numerical trait occurs in a population with an unknown mean $\mu$.

You wish to estimate $\mu$.

Collect a random sample of size $n$

$$Y_1, Y_2, ..., Y_n$$

An estimator for the unknown $\mu$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$
Example

1. You want to study the serum cholesterol level of 17 year old males in the US.
2. Specifically, you are interested in the mean ($\mu$) of serum cholesterol levels of all 17 year old males in the US.
3. You collect a sample of size 25.
4. Observe $\bar{Y} = 30$.
5. What would happen if you collected a different sample of $n = 25$, 17 year old males.
Example

Example: weight of seeds of some variety of beans. Sample size $n = 4$

<table>
<thead>
<tr>
<th>Student #</th>
<th>Observations</th>
<th>sample mean $\bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>462 368 607 483</td>
<td>$\bar{y} = 480$</td>
</tr>
<tr>
<td>2</td>
<td>346 535 650 451</td>
<td>$\bar{y} = 495.5$</td>
</tr>
<tr>
<td>3</td>
<td>579 677 636 529</td>
<td>$\bar{y} = 605.25$</td>
</tr>
</tbody>
</table>

$\bar{Y}$ is random. How do we know its distribution? We will see 3 key facts.
Key fact # 1

If \( Y_1, \ldots, Y_n \) is a random sample, and if the \( Y_i \)'s have mean \( \mu \) and standard deviation \( \sigma \), then

\[
\bar{Y} \text{ has mean} \quad \mu_{\bar{Y}} = \mu
\]

and variance \( \text{var}(\bar{Y}) = \sigma^2 / n \), i.e. standard deviation

\[
\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}
\]

Seed weight example: Assume beans have mean \( \mu = 500 \text{ mg} \) and \( \sigma = 120 \text{ mg} \). In a sample of size \( n = 4 \), the sample mean \( \bar{Y} \) has mean \( \mu_{\bar{Y}} = 500 \text{ mg} \) and standard deviation \( \sigma_{\bar{Y}} = 120 / \sqrt{4} = 60 \text{ mg} \).
Key fact # 2

If $Y_1, \ldots, Y_n$ is a random sample, and if the $Y_i$'s are all from $\mathcal{N}(\mu, \sigma)$, then

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Actually, $Y_1 + \cdots + Y_n = n \bar{Y} \sim \mathcal{N}$ too.

Seed weight example: 100 students do the same experiment.
Key fact # 3

Central limit theorem
If $Y_1, \ldots, Y_n$ is a random sample from (almost) any distribution. Then, as $n$ gets large, $\bar{Y}$ is normally distributed.

Note: $Y_1 + \cdots + Y_n \sim$ normally too.
How big must $n$ be?

Usually, $n = 30$ is big enough, unless the distribution is strongly skewed.

Remarkable result! It explains why the normal distribution is so common, so “normal”. It is what we get when we average over lots of pieces. Ex: human height. Results from ...
Ex: beans are filtered, discarded if too small.
Example: Mixture of 2 bean varieties.
Exercise

Snowfall $Y \sim \mathcal{N}(0.53, 0.21)$ on winter days (inches). Take the sample mean $\bar{Y}$ of a random sample of 30 winter days, over the 10 previous years. What is the probability that $\bar{Y} \leq 0.50$ in?

- $\bar{Y}$ has mean 0.53 inches
- $\bar{Y}$ has standard deviation $0.21/\sqrt{30} = 0.0383$ inches
- $\bar{Y}$’s distribution is approximately normal, because the sample size is large enough ($n = 30$)

\[
\Pr\{\bar{Y} \leq 0.50\} = \Pr\left\{\frac{\bar{Y} - 0.53}{0.0383} \leq \frac{0.50 - 0.53}{0.0383}\right\} \\
\approx \Pr\{Z \leq -0.782\} = 0.217
\]
Normal approximation to the binomial
The normal approximation to the binomial

Example: $X = \#$ of children with side effects after a vaccine, out of $n = 200$ children. Probability of side effect: $p = 0.05$. So $X \sim B(200, 0.05)$.

What is $P\{X \leq 15\}$?

- Direct calculation:

$$P\{X = 0\} + P\{X = 1\} + \cdots + P\{X = 15\} = \binom{200}{0}0.05^00.95^{200} + \cdots + \binom{200}{15}0.05^{15}0.95^{185}$$

- Or we can use a trick: the binomial might be close to a normal distribution. Pretend $X$ is normally distributed!
The normal approximation to the binomial

- \( X = Y_1 + \cdots + Y_{200} \) where
  \[ Y_1 = \begin{cases} 
  1 & \text{if child #1 has side effects,} \\
  0 & \text{otherwise.} 
  \end{cases} \]
  \[ Y_{200} = \begin{cases} 
  1 & \text{if child #200 has side effects,} \\
  0 & \text{otherwise.} 
  \end{cases} \]

- Apply key result #3: if \( n \) (# of children) is large enough, then \( Y_1 + \cdots + Y_n \) has a normal distribution.
- Use the normal distribution with \( X \)'s mean and variance:
  \[ \mu = np = 10, \quad \sigma = \sqrt{np(1-p)} = 3.08 \]

If \( X \sim \mathcal{B}(n, p) \) and if \( n \) is large enough:

- if \( np \geq 5 \) and \( n(1-p) \geq 5 \)

(rule of thumb), then \( X \)'s distribution is approximately

\[ \mathcal{N}(np, \sqrt{np(1-p)}) \]
The normal approximation to the binomial

Back to our question: $\Pr\{X \leq 15\}$.

$n p = 10$ and $n (1 - p) = 190$ are both $\geq 5$, so $X \approx \mathcal{N}(10, 3.08)$.

\[
\Pr\{X \leq 15\} = \Pr\left\{ \frac{X - 10}{3.08} \leq \frac{15 - 10}{3.08} \right\} \\
\approx \Pr\{Z \leq 1.62\} \\
= 0.9474
\]

True value:

> sum( dbinom(0:15, size=200, prob=0.05))

[1] 0.9556444
- Continuity Correction
The continuity correction

# of children with side effect

0 5 10 15 20
The continuity correction
The continuity correction
The continuity correction

$X$ binomial $\mathcal{B}(200, 0.05)$, and $Y$ normal $\mathcal{N}(10, 3.08)$.

No continuity correction:

$$\mathbb{P}\{X \leq 15\} \approx \mathbb{P}\{Y \leq 15\} = \mathbb{P}\left\{ \frac{Y - 10}{3.08} \leq \frac{15 - 10}{3.08} \right\}$$

$$= \mathbb{P}\{Z \leq 1.62\}$$

$$= 0.9474$$

The continuity correction gives a better approximation.

$$\mathbb{P}\{X \leq 15\} \approx \mathbb{P}\{Y \leq 15.5\} = \mathbb{P}\left\{ \frac{Y - 10}{3.08} \leq \frac{15.5 - 10}{3.08} \right\}$$

$$= \mathbb{P}\{Z \leq 1.78\}$$

$$= 0.9624$$

(true value was 0.9556)
The continuity correction

$X$ binomial $\mathcal{B}(200, 0.05)$, and $Y$ normal $\mathcal{N}(10, 3.08)$.

What is the probability that between 8 and 15 children get side effects?

$$\mathbb{P}\{8 \leq X \leq 15\} \approx \mathbb{P}\{7.5 \leq X \leq 15.5\}$$

$$= \mathbb{P}\left\{\frac{7.5 - 10}{3.08} \leq \frac{Y - 10}{3.08} \leq \frac{15.5 - 10}{3.08}\right\}$$

$$= \mathbb{P}\{-0.81 \leq Z \leq 1.78\}$$

$$= \mathbb{P}\{Z \leq 1.78\} - \mathbb{P}\{Z \leq -0.81\}$$

$$= 0.7535$$

True value:

```r
> sum( dbinom(8:15, size=200, prob=0.05) )
[1] 0.7423397```