Outline

1. Sampling distributions
   - Random Samples
   - 3 key facts
   - Normal approximation to the binomial
Random samples

$Y_1, \ldots, Y_n$ form a **random sample** if they are independent and have a common distribution.

   The sample must be **representative** of the targeted population for the $Y$’s common distribution to be unbiased and undistorted.

From a sample, we can calculate a **sample statistic** such as the sample mean $\bar{Y}$.

   $\bar{Y}$ is random too! It can differ from sample to sample.

   The distribution of $\bar{Y}$ is called a **sampling distribution**.
Discrete data: Sampling distribution of a proportion

Number of fruit flies with miniature wings: if allele $m$ is not detrimental then $p$ should be 0.5. Sample of $n = 10$ male offsprings, count the number $Y$ with miniature wings, calculate the sample proportion $\hat{p} = Y/n$.

We would like $\hat{p}$ to be close to the “true” value $p$.

$p$ is fixed and unknown; $\hat{p}$ is random and observed.

Distribution of $\hat{p} =$ its sampling distribution.

If we assume $p = 0.50$, from the binomial:

<table>
<thead>
<tr>
<th>y</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>phat</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>prob</td>
<td>0.001</td>
<td>0.01</td>
<td>0.044</td>
<td>0.117</td>
<td>0.205</td>
<td>0.246</td>
<td>0.205</td>
<td>0.117</td>
<td>0.044</td>
<td>0.01</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Discrete data: Sampling distribution of a proportion

$p$: fixed and unknown, $\hat{p}$: observed but random.
How close is $\hat{p}$ from $p$? How often is $\hat{p}$ is within 0.10 of $p$?
Translate into a binomial question. If true $p = 0.5$:

When $n = 10$:
\[
\Pr\{0.40 \leq \hat{p} \leq 0.60\} = \Pr\{0.40 \leq Y/10 \leq 0.60\} \\
= \Pr\{4 \leq Y \leq 6\} \\
= \Pr\{Y = 4\} + \Pr\{Y = 5\} + \Pr\{Y = 6\} \\
= 0.66
\]

When $n = 20$:
\[
\Pr\{0.40 \leq \hat{p} \leq 0.60\} = \Pr\{0.40 \leq Y/20 \leq 0.60\} \\
= \Pr\{8 \leq Y \leq 12\} \\
= \Pr\{Y = 8\} + \cdots + \Pr\{Y = 12\} \\
= 0.74
\]

Conclusion: sample size of 20 better than sample size of 10!
Continuous data: Sampling distribution of the mean

Example: weight of seeds of some variety of beans. Sample size \( n = 4 \)

<table>
<thead>
<tr>
<th>Experimenter #</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>

\( \mu \) = population mean of all seeds: of interest but unknown. 
\( \bar{Y} \): observed but random.

How do we know the distribution of \( \bar{Y} \)? How close to \( \mu \)? 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

\[
\bar{Y} \text{ has mean } \mu_{\bar{Y}} = \mu \text{ and variance } \text{var}(\bar{Y}) = \frac{\sigma^2}{n}, \text{ i.e. standard deviation }
\]

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

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 \( \sigma_{\bar{Y}} = \frac{120}{\sqrt{4}} = 60 \text{ mg} \).

**Standard error** of an estimate = standard deviation of its sampling distribution. Measures **precision** of the estimate.

Standard error of the mean = \( \frac{\sigma}{\sqrt{n}} \)
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 $\bar{Y}$ also has a normal distribution.

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

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

Seed weight example: 100 experimenters do the same expt.

Sample mean

\begin{align*}
\text{n=4} & \quad \begin{array}{c}
350 \quad 450 \quad 550 \quad 650 \\
3 \quad 14 \quad 30 \quad 32 \quad 17 \quad 4
\end{array} \\
\text{n=16} & \quad \begin{array}{c}
350 \quad 450 \quad 550 \quad 650 \\
0 \quad 5 \quad 56 \quad 32 \quad 7 \quad 0
\end{array}
\end{align*}
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 approximately normally distributed.

Note: \( Y_1 + \cdots + Y_n \) has a normal distribution approximately, 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.
Number of DNA mutations

Between human and chimp genes: \( Y = \# \) nucleotide differences across stretch of 100 base pairs has mean 2 bp and SD 1.4 bp. Distribution skewed

Take mean \( \bar{Y} \) of a random sample of 150 stretches of 100-bp. Probability that \( \bar{Y} \leq 1.6 \) bp?

\[
\begin{align*}
\bar{Y} \text{ has mean } & 2 \text{ bp} \\
\bar{Y} \text{ has standard deviation } & 1.4/\sqrt{150} = 0.114 \text{ bp} \\
\bar{Y}'s \text{ distribution is approximately normal, because the} & \text{ sample size is large (} n = 150). \\
\mathbb{P}\{\bar{Y} \leq .50\} & \approx \mathbb{P}\{Z \leq -3.50\} = 0.00023
\end{align*}
\]
The normal approximation to the binomial

\( X = \# \) of children with side effects (mild fever) after vaccine A, out of \( n = 200 \) children.
If probability of side effect \( p = 0.05 \), then \( X \sim B(200, 0.05) \).

What is \( \Pr\{\hat{p} \leq 0.075\} = \)?? i.e. \( \Pr\{X \leq 15\} \)?

Direct calculation:

\[
\Pr\{X = 0\} + \Pr\{X = 1\} + \cdots + \Pr\{X = 15\} = \\
\left(\begin{array}{c} 200 \\ 0 \end{array}\right) \cdot 0.05^0 \cdot 0.95^{200} + \cdots + \left(\begin{array}{c} 200 \\ 15 \end{array}\right) \cdot 0.05^{15} \cdot 0.95^{185}
\]

Heavy!

Or use fact 3: the binomial is close to a normal distribution, if \( n \) large. Pretend \( X \) is normally distributed!

(why use fact 3?)
We can use fact 3 because $X = Y_1 + \cdots + Y_{200}$ where

$$Y_1 = \begin{cases} 1 & \text{if child #1 has fever} \\ 0 & \text{otherwise} \end{cases}, \ldots, Y_{200} = \begin{cases} 1 & \text{if child #200 has fever} \\ 0 & \text{otherwise.} \end{cases}$$

**Normal approximation to the binomial**

If $X \sim \mathcal{B}(n, p)$ and if $n$ is large enough so that both

$$np \geq 5 \quad \text{and} \quad n(1 - p) \geq 5$$

(rule of thumb), then $X$ and the sample proportion $\hat{p} = X/n$ are both approximately normally distributed:

$$X \sim \mathcal{N}(np, \sqrt{np(1 - p)})$$

$$\hat{p} \sim \mathcal{N}(p, \sqrt{\frac{p(1 - p)}{n}})$$

approximately
The normal approximation to the binomial

\( n = 200 \) children, \( p = 0.05 \) of mild fever, \( \mathbb{P}\{\hat{p} \leq 0.075\} = ? \) i.e. \( \mathbb{P}\{X \leq 15\} = ? \)

Mean of \( X \): \( \mu = np = 10 \), std dev: \( \sigma = \sqrt{np(1-p)} = 3.08 \).

Is \( n \) large enough? \( np = 10 \) and \( n(1-p) = 190 \) both \( \geq 5 \).

so \( X \approx \mathcal{N}(10, 3.08) \).

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

True value:

```r
> dbinom(0:15, size=200, prob=0.05)
[1] 0.000 0.000 0.002 0.007 0.017 0.036 0.061 0.090 0.114
[10] 0.128 0.128 0.117 0.097 0.074 0.052 0.034
> sum(dbinom(0:15, size=200, prob=0.05))
[1] 0.9556444
```

```r
> xvalues = 0:30 # how to plot the binomial distribution
> yvalues = dbinom(0:30, size=200, prob=0.05)
> plot(xvalues, yvalues, type="h")
```
## Recap

### Continuous data:

<table>
<thead>
<tr>
<th></th>
<th>$Y_1, \ldots, Y_n$</th>
<th>$\bar{Y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$\mu$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>std deviation</td>
<td>$\sigma$</td>
<td>$\sigma/\sqrt{n}$</td>
</tr>
<tr>
<td>Normal dist approximately?</td>
<td>Not necessarily</td>
<td>yes if $n$ large</td>
</tr>
<tr>
<td></td>
<td></td>
<td>($\sim 30$ works in most cases)</td>
</tr>
</tbody>
</table>

### Binary (success/failure) data:

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
<th>$\hat{p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$np$</td>
<td>$p$</td>
</tr>
<tr>
<td>std deviation</td>
<td>$\sqrt{np(1-p)}$</td>
<td>$\sqrt{p(1-p)/n}$</td>
</tr>
<tr>
<td>Normal dist approximately?</td>
<td>yes if $np &gt; 5$ and $n(1-p) &gt; 5$</td>
<td>yes if $np &gt; 5$ and $n(1-p) &gt; 5$</td>
</tr>
</tbody>
</table>