Sampling distributions

Cécile Ané

Stat 371

Spring 2006

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

Sampling distributions

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! It can differ from sample to sample. The textbook refers to a *meta-experiment*.
- The distribution of $\bar{Y}$ is called a sampling distribution.

Sampling distribution of a proportion

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

Offspring genotype		
AA	*Aa*	*aa*
0.25	0.50	0.25

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):

Y	0	1	2
$\hat{p}$	0.0	0.5	1.0
$\mathbb{P}$	0.0625	0.3750	0.5625

Sampling distribution of a proportion

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

$$\begin{aligned}\mathbb{P}\{0.70 \le \hat{p} \le 0.80\} &= \mathbb{P}\{0.70 \le Y/20 \le 0.80\} \\ &= \mathbb{P}\{14 \le Y \le 16\} \\ &= \mathbb{P}\{Y = 14\} + \mathbb{P}\{Y = 15\} + \mathbb{P}\{Y = 16\} \\ &= 0.56\end{aligned}$$

Sample size of 20 better than sample size of 2 !!

Sampling distribution of the mean

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

Student #	Observations				sample mean $\bar{y}$
1	462	368	607	483	$\bar{y} = 480$
2	346	535	650	451	$\bar{y} = 495.5$
3	579	677	636	529	$\bar{y} = 605.25$

$\bar{Y}$ 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 μ and standard deviation σ, then

$\bar{Y}$ has mean

$$\mu_{\bar{Y}} = \mu$$

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

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

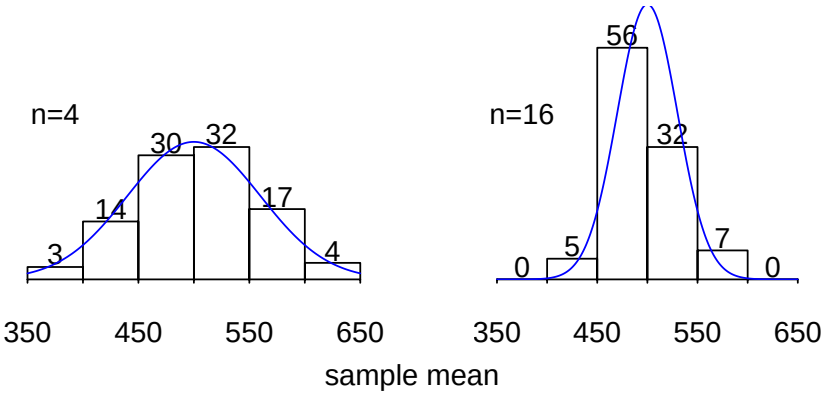
Seed weight example: Assume beans have mean $\mu = 500$ mg and $\sigma = 120$ mg. In a sample of size $n = 4$, the sample mean $\bar{Y}$ has mean $\mu_{\bar{Y}} = 500$ mg and standard deviation $\sigma_{\bar{Y}} = 120/\sqrt{4} = 60$ mg.

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

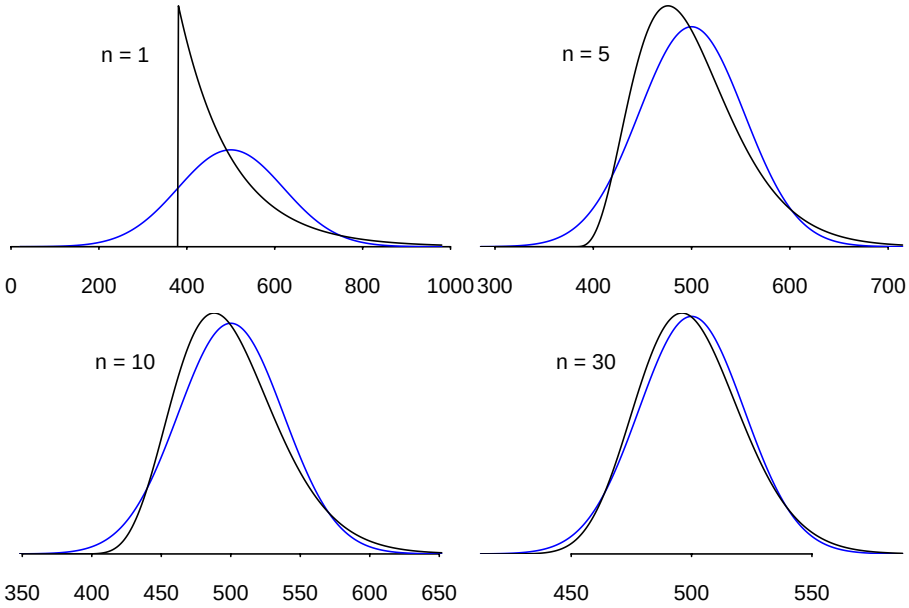
If $Y_1, \ldots, Y_n$ is a random sample from (almost) any distribution. Then, as n gets large, $\bar{Y}$ is normally distributed.

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

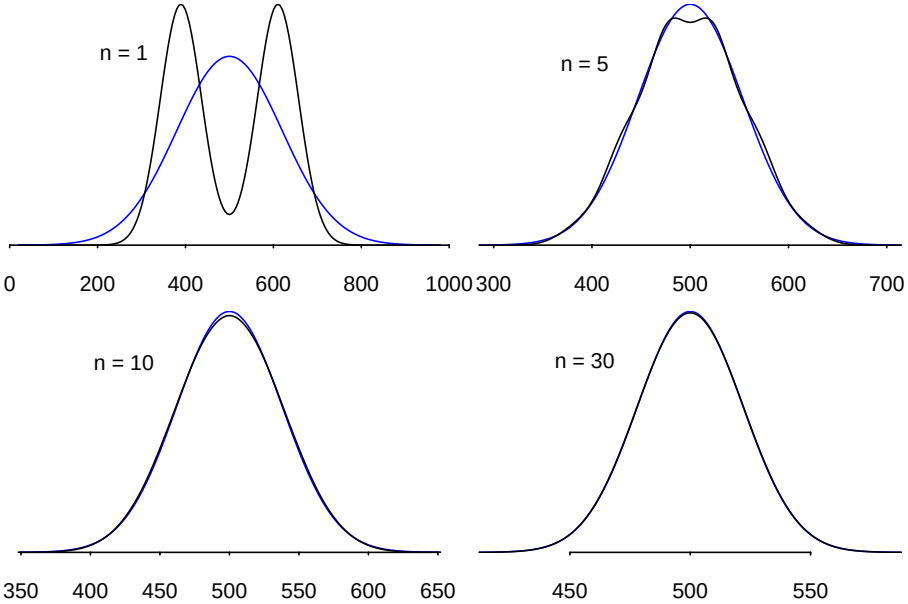
Usually, $n = 30$ is big enough, *unless* the distribution is strongly skewed.

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 $Y \sim \mathcal{N}(.53, .21)$ 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 $\bar{Y} \leq .50$ 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 ($n = 30$)

$$\begin{aligned} \mathbb{P}\{\bar{Y} \leq .50\} &= \mathbb{P}\left\{\frac{\bar{Y} - 0.53}{.0383} \leq \frac{0.50 - 0.53}{.0383}\right\} \\ &\simeq \mathbb{P}\{Z \leq -0.782\} = 0.217 \end{aligned}$$

The normal approximation to the binomial

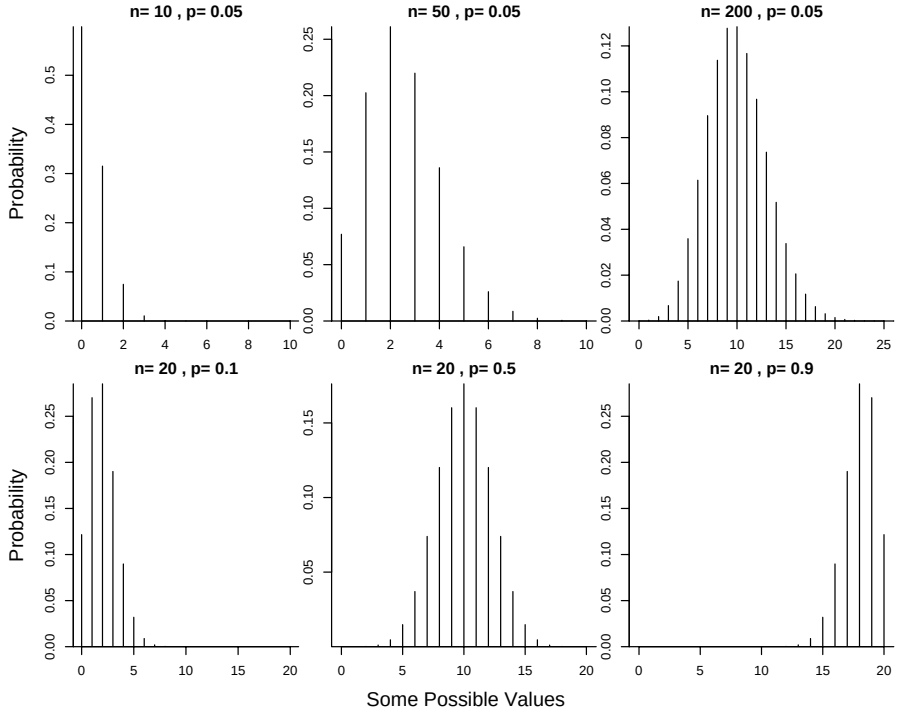
Example: $X = \#$ of children with side effects after a vaccine, out of $n = 200$ children. Probability of side effect: $p = 0.05$. So $X \sim \mathcal{B}(200, 0.05)$.

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

- Direct calculation:

$$\mathbb{P}\{X = 0\} + \mathbb{P}\{X = 1\} + \cdots + \mathbb{P}\{X = 15\} = {}_{200}C_0 .05^0 .95^{200} + \cdots + {}_{200}C_{15} .05^{15} .95^{185}$$

  Heavy!
- 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:

$$\mu = np = 10\,, \qquad \sigma = \sqrt{np(1-p)} = 3.08$$

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

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

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

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

The normal approximation to the binomial

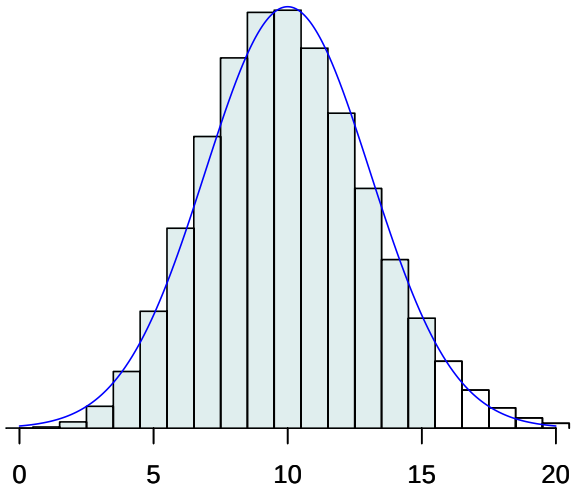
Back to our question: $\mathbb{P}\{X \leq 15\}$.
$np = 10$ and $n(1-p) = 190$ are both ≥ 5, so $X \approx \mathcal{N}(10, 3.08)$.

$$\begin{aligned} \mathbb{P}\{X \leq 15\} &= \mathbb{P}\left\{\frac{X - 10}{3.08} \leq \frac{15 - 10}{3.08}\right\} \\ &\simeq \mathbb{P}\{Z \leq 1.62\} \\ &= 0.9474 \end{aligned}$$

True value:

```
> sum( dbinom(0:15, size=200, prob=0.05))
[1] 0.9556444
```

The continuity correction



# of children with side effect

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?

$$\begin{aligned}
\mathbb{P}\{8 \leq X \leq 15\} &\simeq \mathbb{P}\{7.5 \leq X \leq 15.5\} \\
&= \mathbb{P}\left\{\frac{7.5-10}{3.08} \leq \frac{Y-10}{3.08} \leq \frac{15.5-10}{3.08}\right\} \\
&= \mathbb{P}\{-0.81 \leq Z \leq 1.78\} \\
&= \mathbb{P}\{Z \leq 1.78\} - \mathbb{P}\{Z \leq -0.81\} \\
&= 0.7535
\end{aligned}$$

True value:

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

The continuity correction

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

No continuity correction:

$$\begin{aligned}
\mathbb{P}\{X \leq 15\} &\simeq \mathbb{P}\{Y \leq 15\} = \mathbb{P}\left\{\frac{Y-10}{3.08} \leq \frac{15-10}{3.08}\right\} \\
&= \mathbb{P}\{Z \leq 1.62\} \\
&= 0.9474
\end{aligned}$$

The continuity correction gives a better approximation.

$$\begin{aligned}
\mathbb{P}\{X \leq 15\} &\simeq \mathbb{P}\{Y \leq 15.5\} = \mathbb{P}\left\{\frac{Y-10}{3.08} \leq \frac{15.5-10}{3.08}\right\} \\
&= \mathbb{P}\{Z \leq 1.78\} \\
&= 0.9624
\end{aligned}$$

(true value was 0.9556)