Outline

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:

y	0	1	2	3	4	5	6	7	8	9	10
phat	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
prob	0.001	0.01	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.01	0.001

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 $n = 10$:

$$\begin{aligned}
\mathbb{P}\{0.40 \leq \hat{p} \leq 0.60\} &= \mathbb{P}\{0.40 \leq Y/10 \leq 0.60\} \\
&= \mathbb{P}\{4 \leq Y \leq 6\} \\
&= \mathbb{P}\{Y = 4\} + \mathbb{P}\{Y = 5\} + \mathbb{P}\{Y = 6\} \\
&= 0.66
\end{aligned}$$

When $n = 20$:

$$\begin{aligned}
\mathbb{P}\{0.40 \leq \hat{p} \leq 0.60\} &= \mathbb{P}\{0.40 \leq Y/20 \leq 0.60\} \\
&= \mathbb{P}\{8 \leq Y \leq 12\} \\
&= \mathbb{P}\{Y = 8\} + \cdots + \mathbb{P}\{Y = 12\} \\
&= 0.74
\end{aligned}$$

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 $n = 4$

Experimenter #	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	$\bar{y} = 605.25$

$\mu =$ 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 μ?
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.

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

Standard error of the mean = $\sigma/\sqrt{n}$

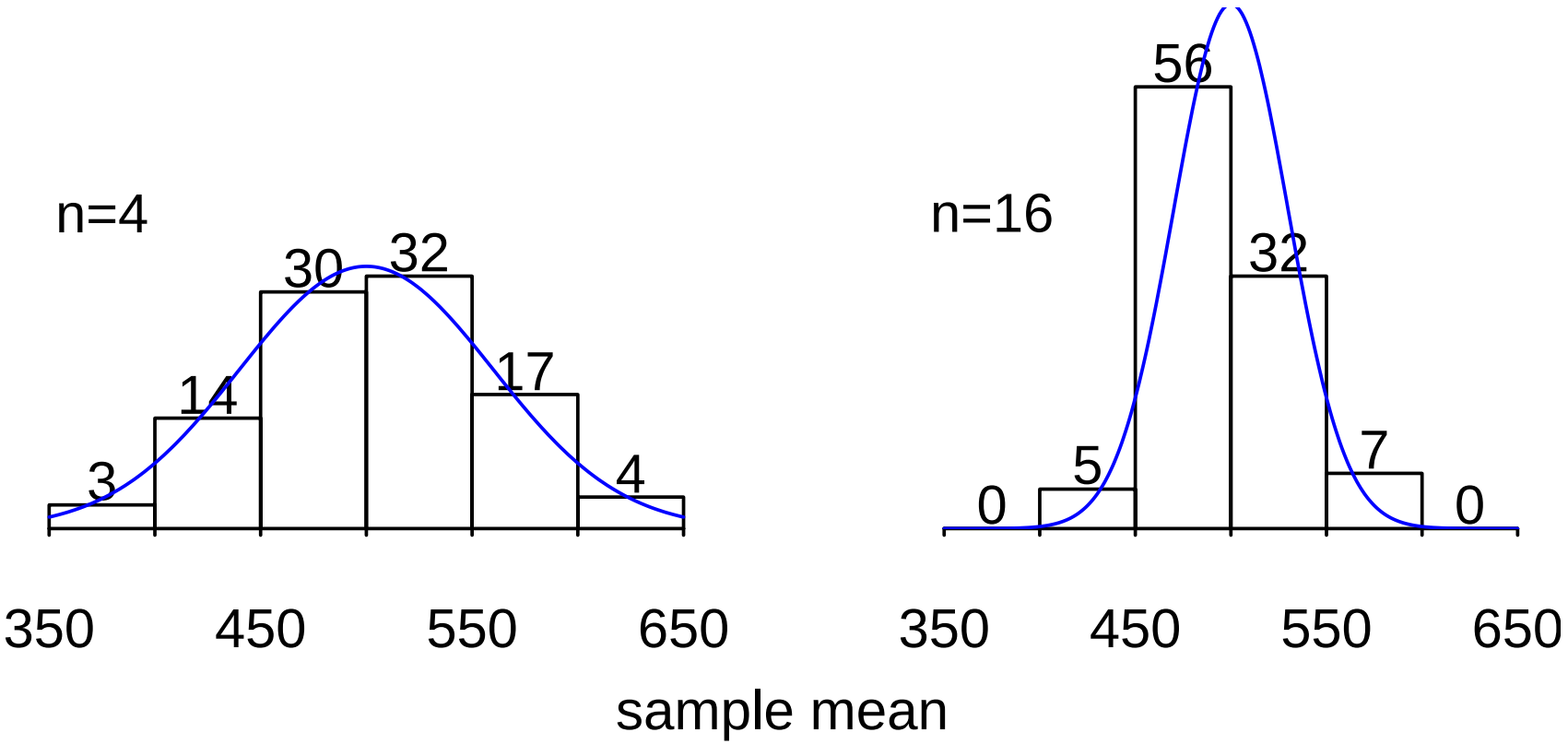
Key fact # 2

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.



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.

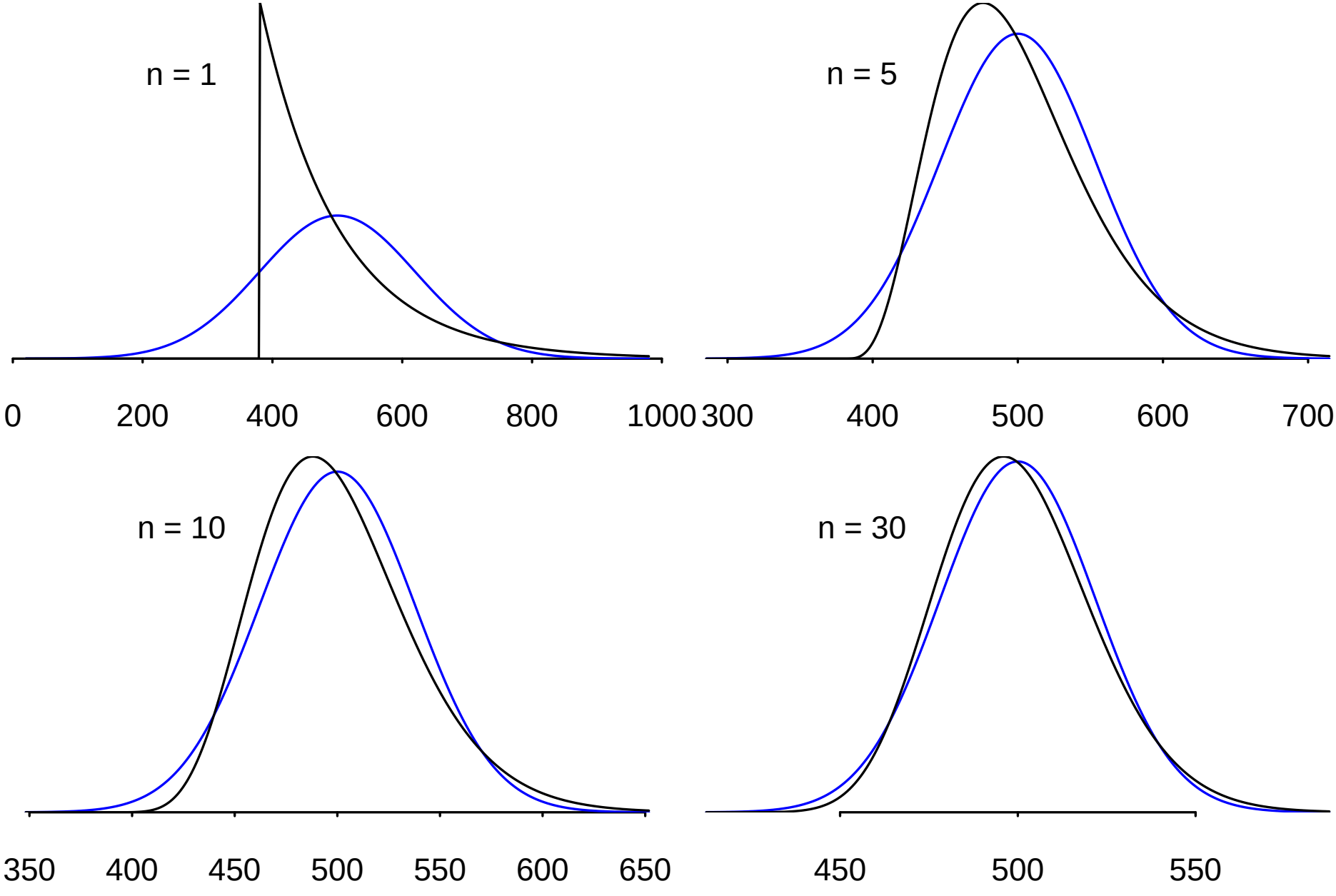
Note: $Y_1 + \cdots + Y_n$ has a normal distribution approximately, too.

How big must n be?

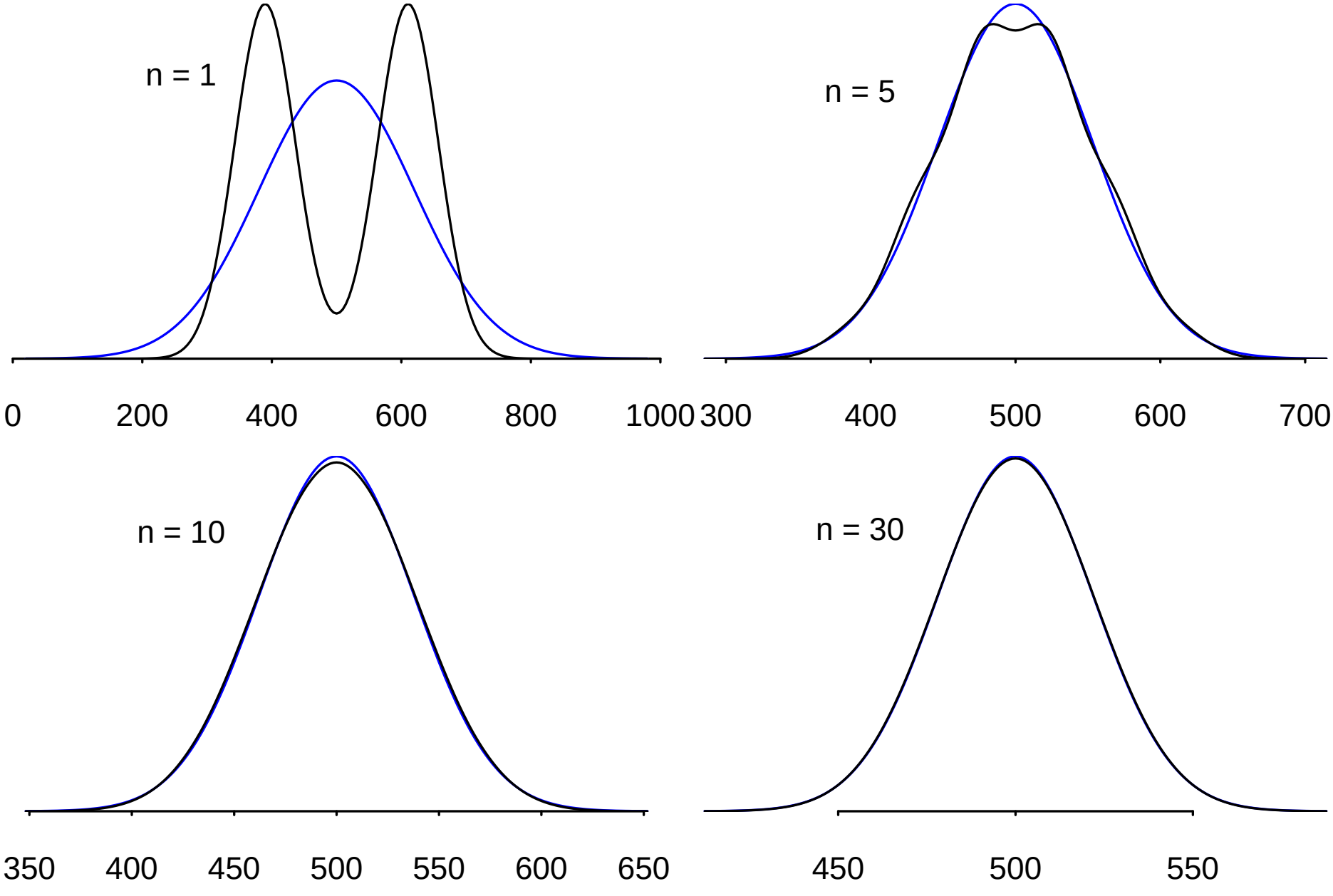
Usually, $n = 30$ is big enough, *unless* the distribution is strongly skewed.

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?

- $\bar{Y}$ has mean 2 bp
- $\bar{Y}$ has standard deviation $1.4/\sqrt{150} = 0.114$ bp
- $\bar{Y}$'s distribution is approximately normal, because the sample size is large ($n = 150$).

$$\mathbb{P}\{\bar{Y} \leq .50\} = \qquad \simeq \mathbb{P}\{Z \leq -3.50\} = 0.00023$$

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 $X \sim \mathcal{B}(200, 0.05)$.

What is $\mathbb{P}\{\hat{p} \leq 0.075\} =$?? i.e. $\mathbb{P}\{X \leq 15\}$?

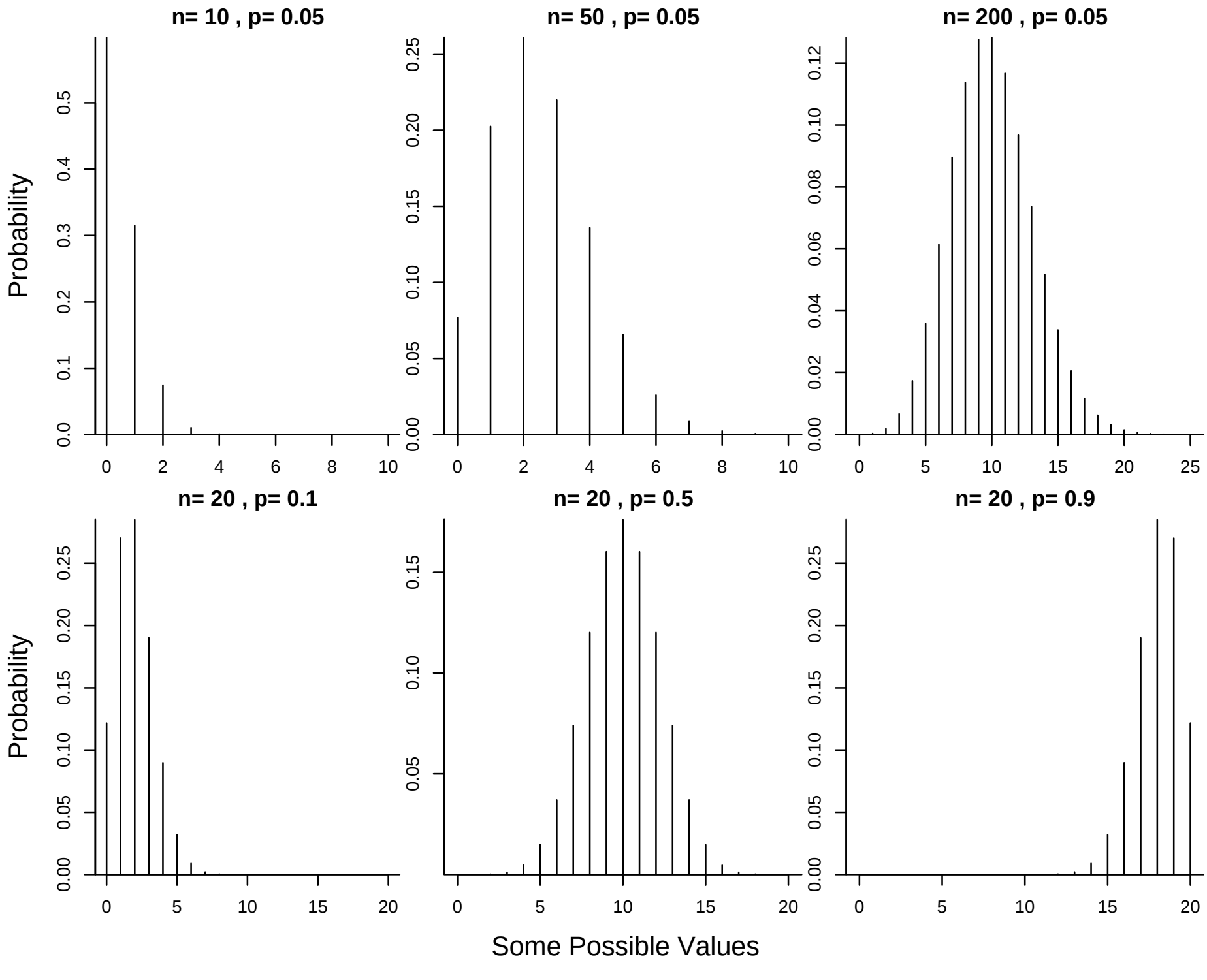
Direct calculation:

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

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?)


n= 10 , p= 0.05
n= 50 , p= 0.05
n= 200 , p= 0.05
n= 20 , p= 0.1
n= 20 , p= 0.5
n= 20 , p= 0.9
Probability
Probability
Some Possible Values

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:

$$X \sim \mathcal{N}(np, \sqrt{np(1-p)})$$
$$\hat{p} \sim \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}}) \qquad \text{approximately}$$

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 ≥ 5.

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

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

True value:

```
>       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

> 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:

	$Y_1, \ldots, Y_n$	$\bar{Y}$
mean	μ	μ
std deviation	σ	$\sigma/\sqrt{n}$
Normal dist approximately?	Not necessarily	yes if n large (~ 30 works in most cases)

Binary (success/failure) data:

	Y	$\hat{p}$
mean	np	p
std deviation	$\sqrt{np(1-p)}$	$\sqrt{p(1-p)/n}$
Normal dist approximately?	yes if $np > 5$ and $n(1-p) > 5$	yes if $np > 5$ and $n(1-p) > 5$