# Sampling Distributions

## Basic Ideas

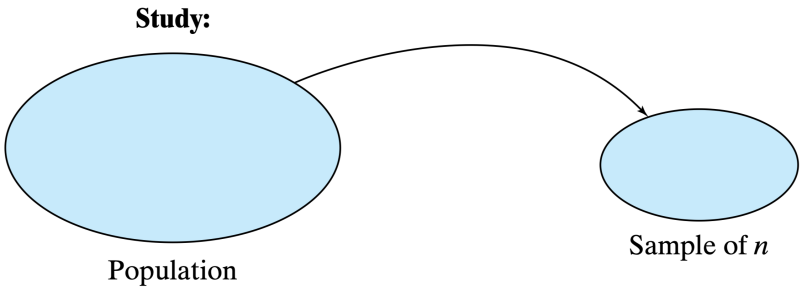
- An important goal of data analysis is to distinguish between features of the data that reflect real biological facts and features that may reflect only chance effects.
- The random sampling model provides a framework for making this distinction: Chance effects are regarded as sampling error. That is, discrepancy between the sample and the population.
- In this chapter we develop the theoretical background that will enable us to place specific limits on the degree of sampling error to be expected in a study.

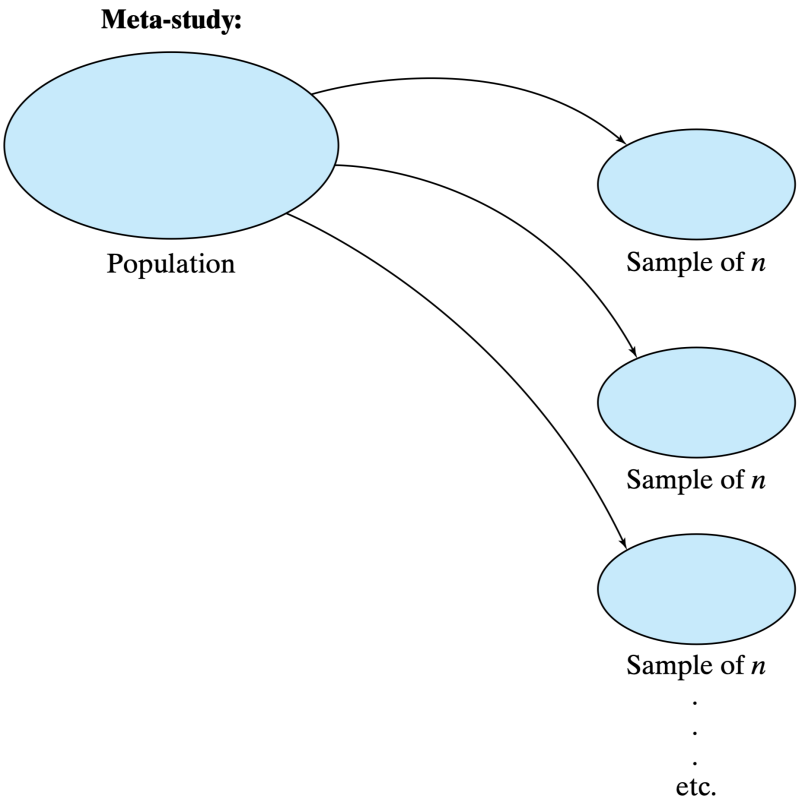
### Sampling variability

- The variability among random samples from the same population is called **sampling variability**.
- A probability distribution that characterizes some aspect of sampling variability is termed a **sampling distribution**.
- We have to expect a certain amount of discrepancy between the sample and the population due to the sampling error.

### The meta-study

A **meta-study** consists of indefinitely many repetitions, or replications, of the same study. If the study consists of drawing a random sample of size $n$ from some population, the corresponding meta-study involves drawing repeated random samples of size $n$ from the same population.



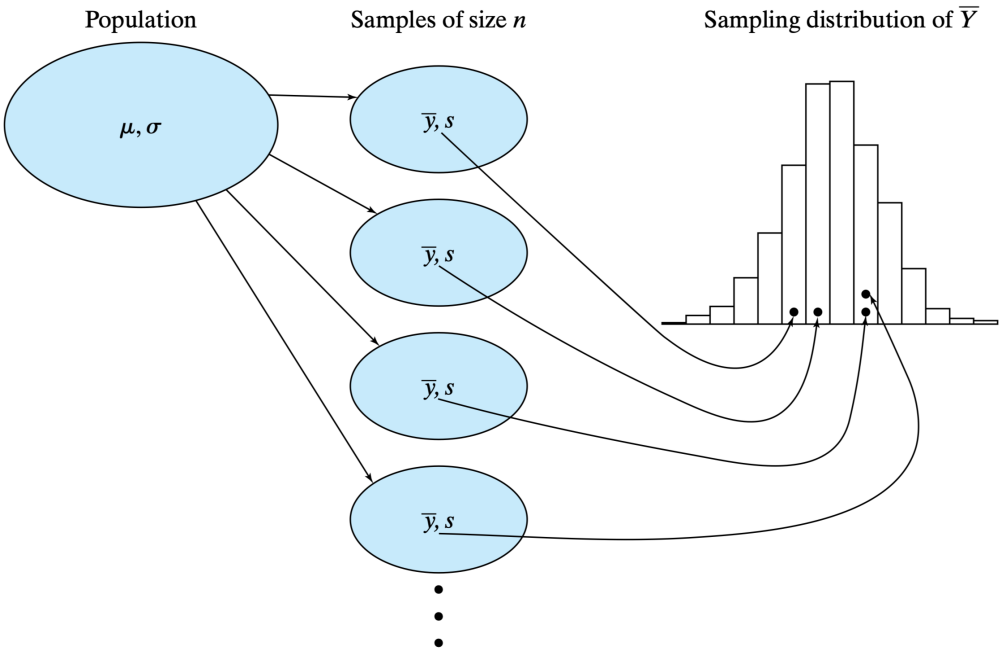



## The Sample Mean

- The sample mean $\bar{y}$ can be used, not only as a description of the data in the sample, but also as an estimate of the population mean $\mu$. It is natural to ask, "How close to $\mu$ is $\bar{y}$?"
- We cannot answer this question for the mean $\bar{y}$ of a particular sample due to the **randomness** of the sample. Regarding the sample mean as a random variable $\bar{Y}$, the question then becomes: "How close to $\mu$ is $\bar{Y}$ *likely* to be?"
- To characterize such randomness, we resort to the **sampling distribution** of the sample mean $\bar{Y}$, the probability distribution that describes sampling variability in $\bar{Y}$.

To visualize the sampling distribution of $\bar{Y}$, imagine the meta-study as follows:

- Random samples of size $n$ are repeatedly drawn from a fixed population with mean $\mu$ and standard deviation $\sigma$; each sample has its own mean $\bar{y}$.
- The variation of the $\bar{y}$'s among the samples is specified by the sampling distribution of $\bar{Y}$.




When we think of $\bar{Y}$ as a random variable, we need to be aware of two basic facts

- On average, the sample mean equals to the population mean. That is, the average of the sampling distribution of $\bar{Y}$ is $\mu$.
- As the sample size increases, the standard deviation of $\bar{Y}$ decreases. That is, for larger samples, $\bar{Y}$ will tend to be closer to the population mean.

## Theorem

### Theorem 5.2.1: The Sampling Distribution of $\overline{Y}$

1. Mean The mean of the sampling distribution of $\overline{Y}$ is equal to the population mean. In symbols,

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

2. Standard deviation The standard deviation of the sampling distribution of $\overline{Y}$ is equal to the population standard deviation divided by the square root of the sample size. In symbols,

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

3. Shape
   (a) If the population distribution of $Y$ is normal, then the sampling distribution of $\overline{Y}$ is normal, regardless of the sample size $n$.
   (b) *Central Limit Theorem* If $n$ is large, then the sampling distribution of $\overline{Y}$ is approximately normal, even if the population distribution of $Y$ is not normal.

- Consider the random sample $Y_1, \ldots, Y_n$, drawn from a population with mean $\mu$ and standard deviation $\sigma$. The sample mean is denoted as $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Try to derive

Parts 1 and 2 of the above theorem.

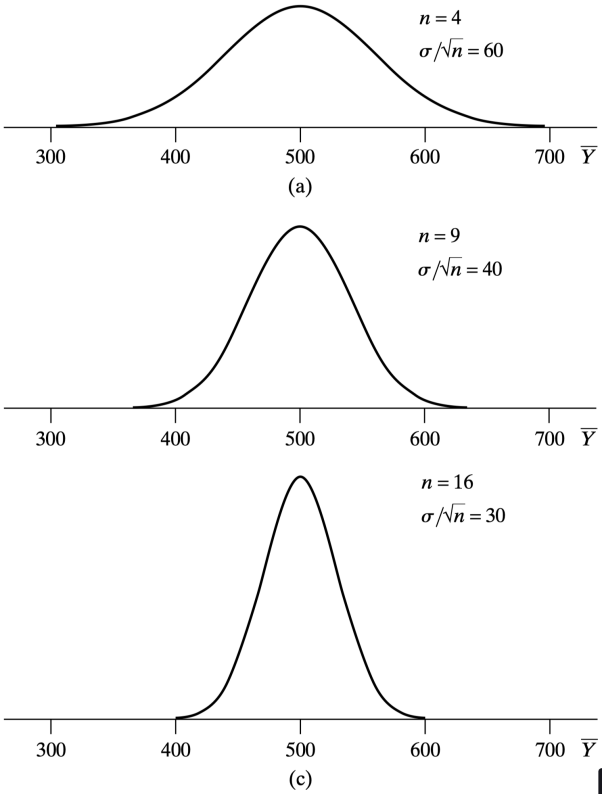
- The **Central Limit Theorem** states that, no matter what distribution $Y$ may have in the population, if the sample size is large enough, then the sampling distribution of $\bar{Y}$ will be approximately a normal distribution.
- It is because of the Central Limit Theorem (and other similar theorems) that the normal distribution plays such a central role in statistics.
- It is natural to ask how "large" a sample size is required by the Central Limit Theorem.
  - If the shape is normal, any n will do.
  - If the shape is moderately nonnormal, a moderate n is adequate.
  - If the shape is highly nonnormal, then a rather large n will be required.

## Example: weights of seeds

A large population of seeds of the princess bean *Phaseotus vulgaris* is to be sampled. The weights of the seeds in the population follow a normal distribution with mean $\mu = 500$ mg and standard deviation $\sigma = 120$ mg. Suppose now that a random sample of four seeds is to be weighed, and let $\bar{Y}$ represent the mean weight of the four seeds. What is the sampling distribution of $\bar{Y}$? $N(500, 3600)$

## Dependence of sample size

- Larger $n$ gives a smaller value of $\sigma_{\bar{Y}}$ and consequently a smaller expected sampling error if $\bar{y}$ is used as an estimate of $\mu$.
- If the population distribution is not normal, then the shape of the sampling distribution of $\bar{Y}$ depends on $n$, being more nearly normal for larger $n$.
- The mean of a larger sample is not necessarily closer to $\mu$ than the mean of a smaller sample, but it has a **greater probability** of being close. It is in this sense that a larger sample provides more information about the population mean than a smaller sample.




## Populations, samples, and sampling distributions

It is important to distinguish clearly among three different distributions related to a quantitative variable $Y$:

- the distribution of $Y$ in the population;
- the distribution of $Y$ in a sample of data, and
- the sampling distribution of $\bar{Y}$.

| Distribution | Mean | Standard deviation |
|---|---|---|
| $Y$ in population | $\mu$ | $\sigma$ |
| $Y$ in sample | $\bar{y}$ | $s$ |
| $\bar{Y}$ (in meta-study) | $\mu_{\bar{Y}} = \mu$ | $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ |

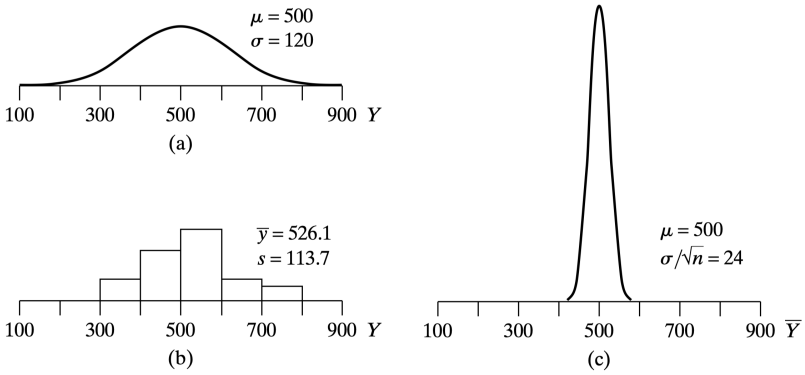
## Example

Recall the weights of seeds example, the population mean and standard deviation are $\mu = 500$ mg and $\sigma = 120$ mg. Suppose we weigh a random sample of $n = 25$ seeds from the population and obtain the data in the table below

**Table 5.2.3** Weights of 25 princess bean seeds

| Weight (mg) | | | | | | |
|---|---|---|---|---|---|---|
| 343 | 755 | 431 | 480 | 516 | 469 | 694 |
| 659 | 441 | 562 | 597 | 502 | 612 | 549 |
| 348 | 469 | 545 | 728 | 416 | 536 | 581 |
| 433 | 583 | 570 | 334 | | | |

- The population distribution of $Y =$ weights is represented in (a)
- the sample mean is $\bar{y} = 526.1$ mg and the sample standard deviation is $s = 113.7$ mg. (b) shows a histogram of the data; this histogram represents the distribution of $Y$ in the sample.
- The sampling distribution of $\bar{Y}$ as shown in (c) is a theoretical distribution which relates, not to the particular sample shown in the histogram, but rather to the meta-study of infinitely repeated samples of size $n = 25$. The mean and standard deviation of the sampling distribution are $\mu_Y = 500$ mg and $\sigma_Y = 120/\sqrt{25} = 24$ mg.




Notice that the distributions in (a) and (b) are more or less similar; in fact, the distribution in (b) is an estimate of the distribution in (a). By contrast, the distribution in (c) is much narrower, because it represents a distribution of **means** rather than of individual observations.

# The Normal Approximation to the Binomial Distribution

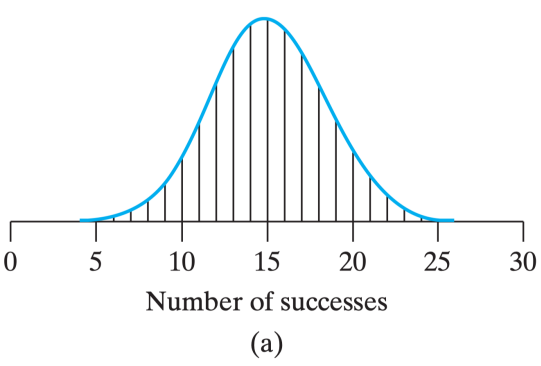
- The binomial random variable $X \sim B(n, p)$ is the sum of $n$ identical Bernoulli random variables, each with expected value $p$ and variance $p(1-p)$. In other words, if $X_1, \ldots, X_n$ are identical (and independent) Bernoulli random variables with parameter $p$, then $X = X_1 + \cdots + X_n$.
- Think of $X_1, \ldots, X_n$ as a random sample. Then the sample mean $\hat{P} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is governed by the Central Limit Theorem.

## Theorem

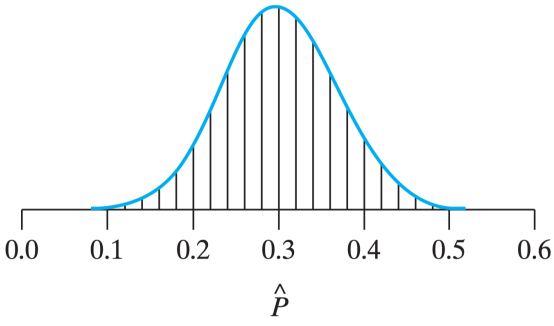
- If $n$ is large, then the binomial distribution of the probability of success, $\hat{P}$, can be approximated by a normal distribution with mean $= p$ and standard deviation $= \sqrt{p(1-p)/n}$.
- If $n$ is large, then the binomial distribution of the number of successes, $Y$, can be approximated by a normal distribution with mean $= np$ and standard deviation $= \sqrt{np(1-p)}$.

## Example: normal approximation to binomial

We consider a binomial distribution with $n = 50$ and $p = 0.3$. (a) shows this binomial distribution, using spikes to represent probabilities; superimposed is a normal curve with mean $= np = 15$ and standard deviation $= \sqrt{np(1-p)} = 3.24$. (b) shows the sampling distribution of $\hat{P}$; superimposed is a normal curve with mean $= p = 0.3$ and standard deviation $= \sqrt{p(1-p)/n} = 0.0648$.

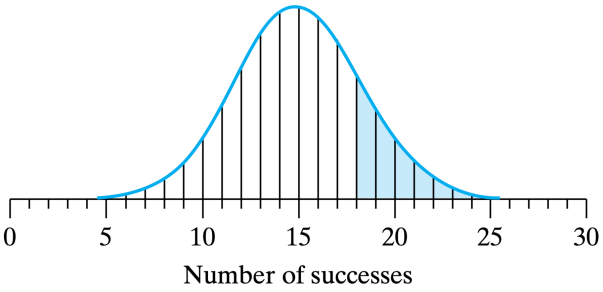



(a)




(b)

To illustrate the use of the normal approximation, let us find the probability that $50$ independent trials result in at least $18$ successes, i.e., $P(Y \geq 18)$. The exact calculation using the binomial formula is very tedious, which involves $50 - 18 + 1 = 33$ terms ($0.2178$ ). If instead the normal approximation is adopted, we only need to find the corresponding area under the normal curve.




The Z score that corresponds to $18$ is

$$z = \frac{18 - 15}{3.2404} = 0.93.$$

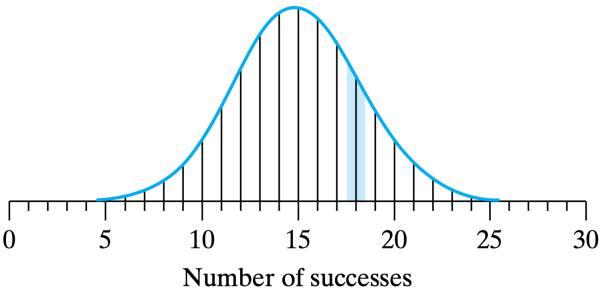
We find that the area is $1 - 0.8238 = 0.1762$ using Z table.

## The continuity correction

- What would happen if we want to compute $P(Y = 18)$ using the normal approximation, the probability of 18 successes?
- We think of "$18$" as covering the space from $17.5$ to $18.5$ and thus we consider the area under the normal curve between $17.5$ and $18.5$.
- Compute $P(Y \geq 18)$ using the continuity correction.
  - The Z score is

$$z = \frac{17.5 - 15}{3.2404} = 0.77$$

  - From the Z table, we find that the area above $0.77$ is $1 - 0.7794 = 0.2206$.
- What about $P(12 \leq Y \leq 18)$ and $P(12 < Y < 18)$?




## Summary of continuity correction

- If $P(Y = n)$ use

$$P(n - 0.5 < Y < n + 0.5).$$

- If $P(Y > n)$ use

$$P(Y > n + 0.5).$$

- If $P(Y \leq n)$ use

$$P(Y < n + 0.5).$$

- If $P(Y < n)$ use

$$P(Y < n - 0.5).$$

- If $P(Y \geq n)$ use

$$P(Y > n - 0.5).$$

## How large must $n$ be?

The required $n$ depends on the value of $p$.

- If $p = 0.5$, then the binomial distribution is *symmetric* and the normal approximation is quite good even for $n$ as small as $10$.
- However, if $p = 0.1$, the binomial distribution for $n = 10$ is quite skewed and is poorly fitted by a normal curve; for larger $n$ the skewness is diminished and the normal approximation is better.
- A simple rule of thumb is the following:
  - *The normal approximation to the binomial distribution is fairly good if both $np$ and $n(1 - p)$ are at least equal to $5$.*