Sampling Distributions

For Counts and Proportions

IPS Chapter 5.1

Objectives (IPS Chapter 5.1)

Sampling distributions for counts and proportions

- Binomial distributions for sample counts
- Binomial distributions in statistical sampling
- Binomial mean and standard deviation
- Sample proportions
- Normal approximation
- Binomial formulas

Reminder: the two types of data

- ***Quantitative***
  - Something that can be counted or measured and then averaged across individuals in the population (e.g., your height, your age, your IQ score)

- ***Categorical***
  - Something that falls into one of several categories. What can be counted is the proportion of individuals in each category (e.g., your gender, your hair color, your blood type—A, B, AB, O).

How do you figure it out? Ask:

- What are the *n* individuals/units in the sample (of size "*n*")?
- What is being recorded about those *n* individuals/units?
- Is that a number (→ quantitative) or a statement (→ categorical)?

Binomial distributions for sample counts

Binomial distributions are models for some categorical variables, typically representing the number of successes in a series of *n* trials.

The observations must meet these requirements:

- The total number of observations *n* is fixed in advance.
- Each observation falls into just 1 of 2 categories: success and failure.
- The outcomes of all *n* observations are statistically independent.
- All *n* observations have the same probability of "success," *p*.

We record the next 50 births at a local hospital. Each newborn is either a boy or a girl; each baby is either born on a Sunday or not.

We express a binomial distribution for the count X of successes among n observations as a function of the parameters n and p: $B(n,p)$.

- The parameter n is the total number of observations.
- The parameter p is the probability of success on each observation.
- The count of successes X can be any whole number between 0 and n.

A coin is flipped 10 times. Each outcome is either a head or a tail. The variable X is the number of heads among those 10 flips, our count of "successes."

On each flip, the probability of success, "head," is 0.5. The number X of heads among 10 flips has the binomial distribution $B(n = 10, p = 0.5)$.

Applications for binomial distributions

Binomial distributions describe the possible number of times that a particular event will occur in a sequence of observations.

They are used when we want to know about the occurrence of an event, not its magnitude.

- In a clinical trial, a patient's condition may improve or not. We study the number of patients who improved, not how much better they feel.
- Is a person ambitious or not? The binomial distribution describes the number of ambitious persons, not how ambitious they are.
- In quality control we assess the number of defective items in a lot of goods, irrespective of the type of defect.

Imagine that coins are spread out so that half of them are heads up, and half tails up. Close your eyes and pick one. The probability that this coin is heads up is 0.5.

However, if you don't put the coin back in the pile, the probability of picking up another coin and having it be heads up is now less than 0.5. The successive observations are not independent.

Likewise, choosing a simple random sample (SRS) from any population is not quite a binomial setting. However, when the population is large, removing a few items has a very small effect on the composition of the remaining population: successive observations are very nearly independent.

Binomial distribution in statistical sampling

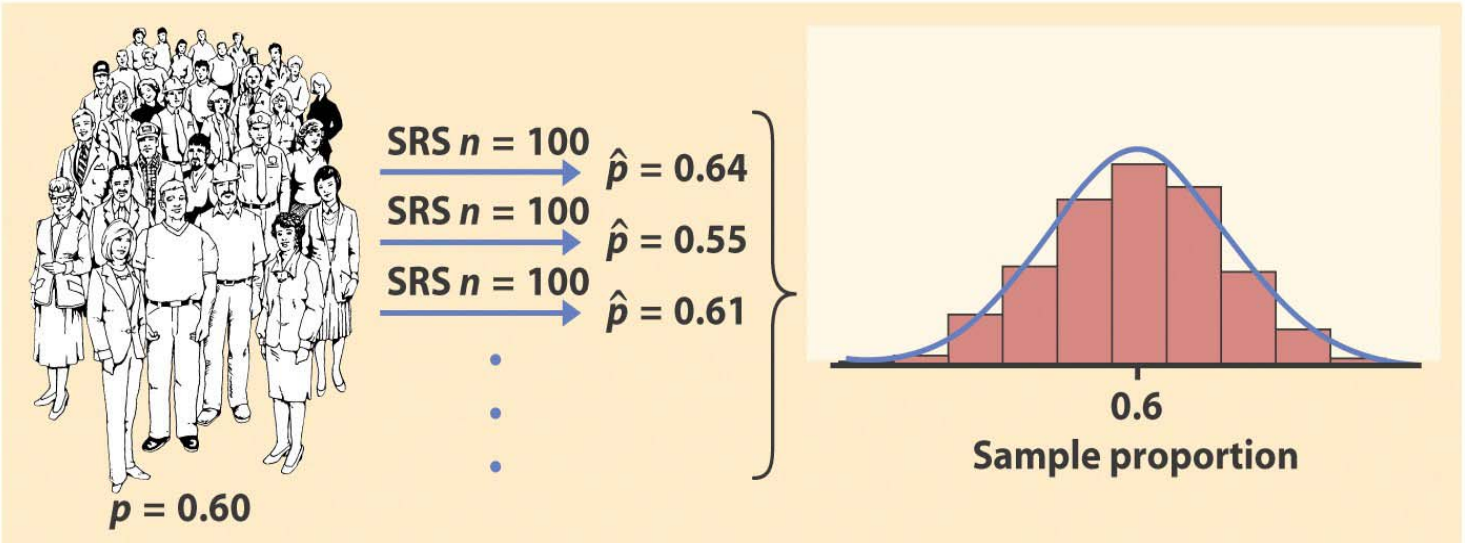
A population contains a proportion *p* of successes. If the population is much larger than the sample, the count *X* of successes in an SRS of size n has approximately the binomial distribution *B*(*n, p*).

The *n* observations will be nearly independent when the size of the population is much larger than the size of the sample. As a rule of thumb, the **binomial sampling distribution for counts** can be used when the population is at least 20 times as large as the sample.

Reminder: Sampling variability

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called sampling variability.

If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern.



Calculations

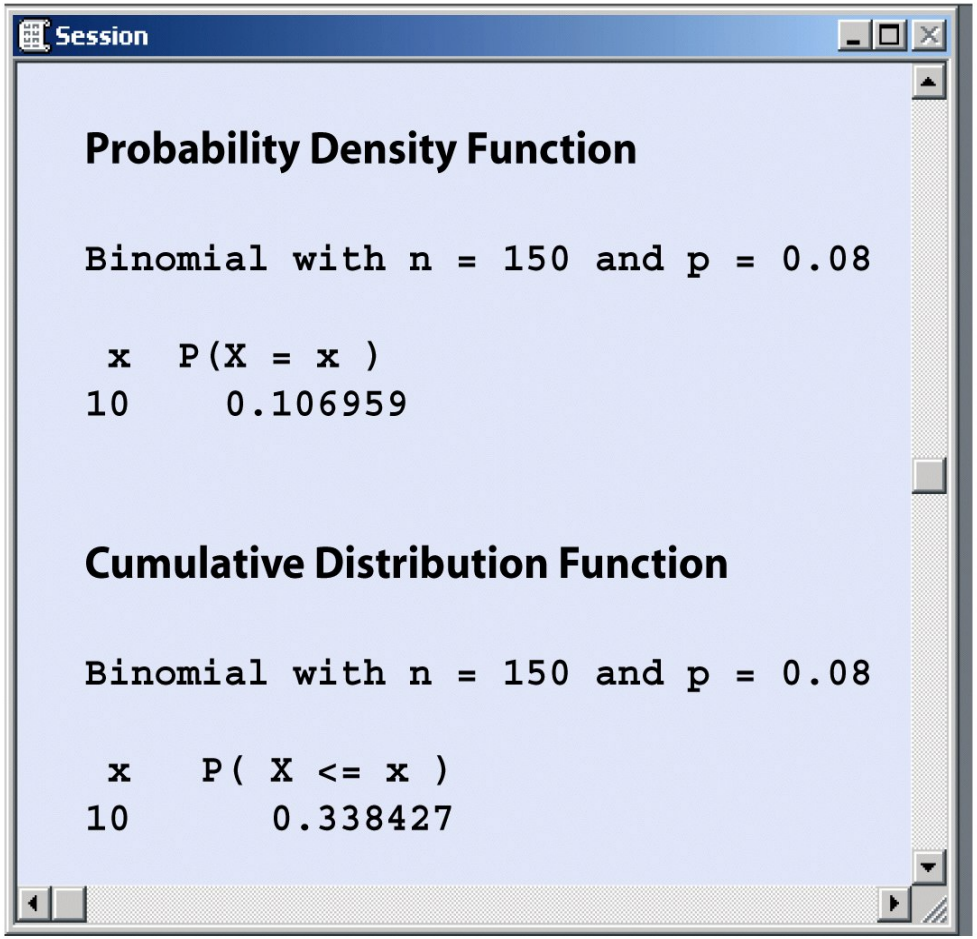
The probabilities for a Binomial distribution can be calculated by using software.

In **Minitab,**

Menu/Calc/

Probability Distributions/Binomial

- Choose “Probability” for the probability of a given number of successes $P(X = x)$
- Or “Cumulative probability” for the density function $P(X \leq x)$



Software commands: **Excel:**

=BINOMDIST (number_*s*, trials, probability_*s*, cumulative)

Number_*s:*

number of successes in trials.

Trials:

number of independent trials.

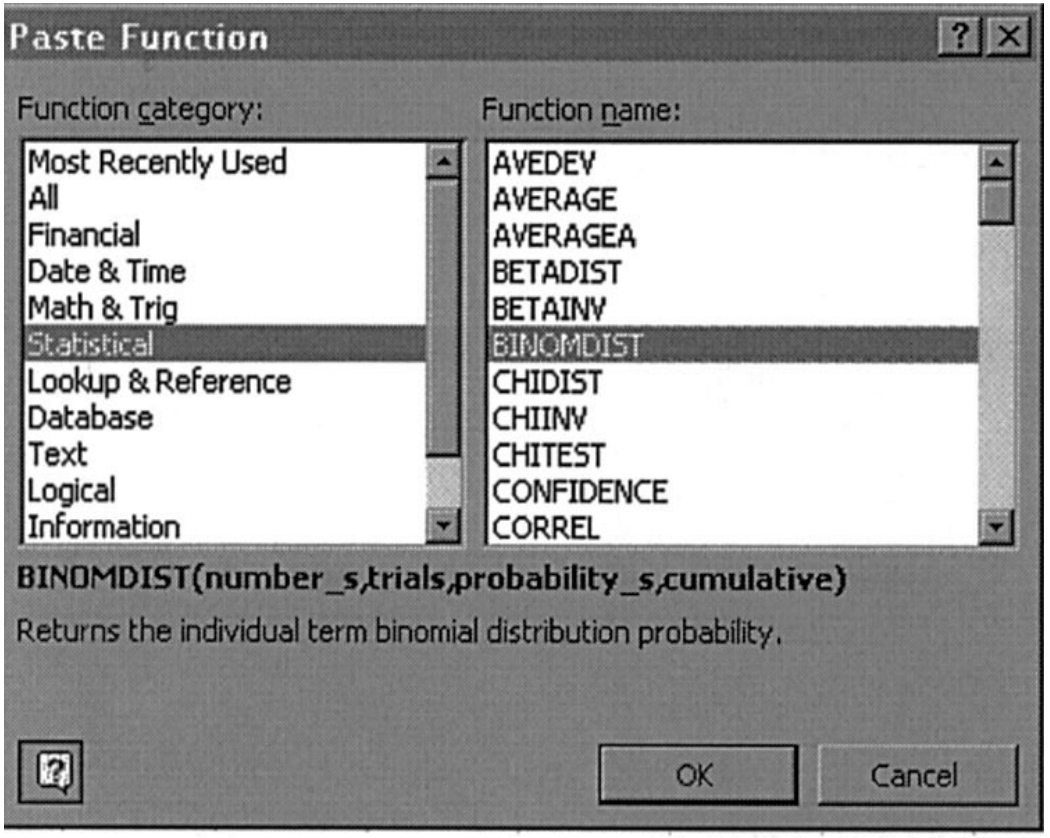
Probability_*s:*

probability of success on each trial.

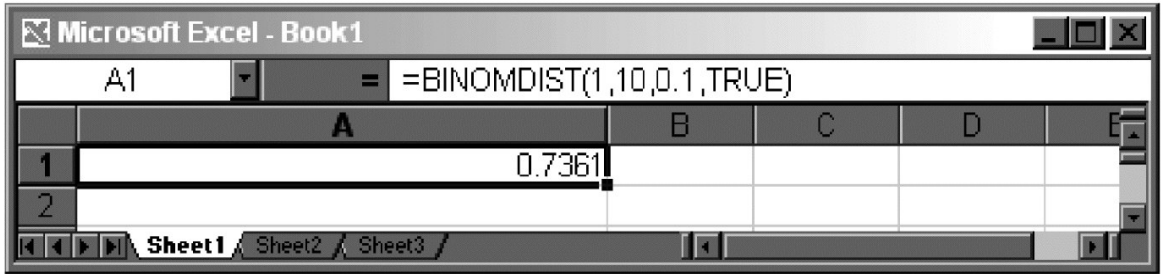
Cumulative:

a logical value that determines the form of the function.

- TRUE, or 1, for the cumulative $P(X \le \text{Number\_}s)$
- FALSE, or 0, for the probability function $P(X = \text{Number\_}s)$.



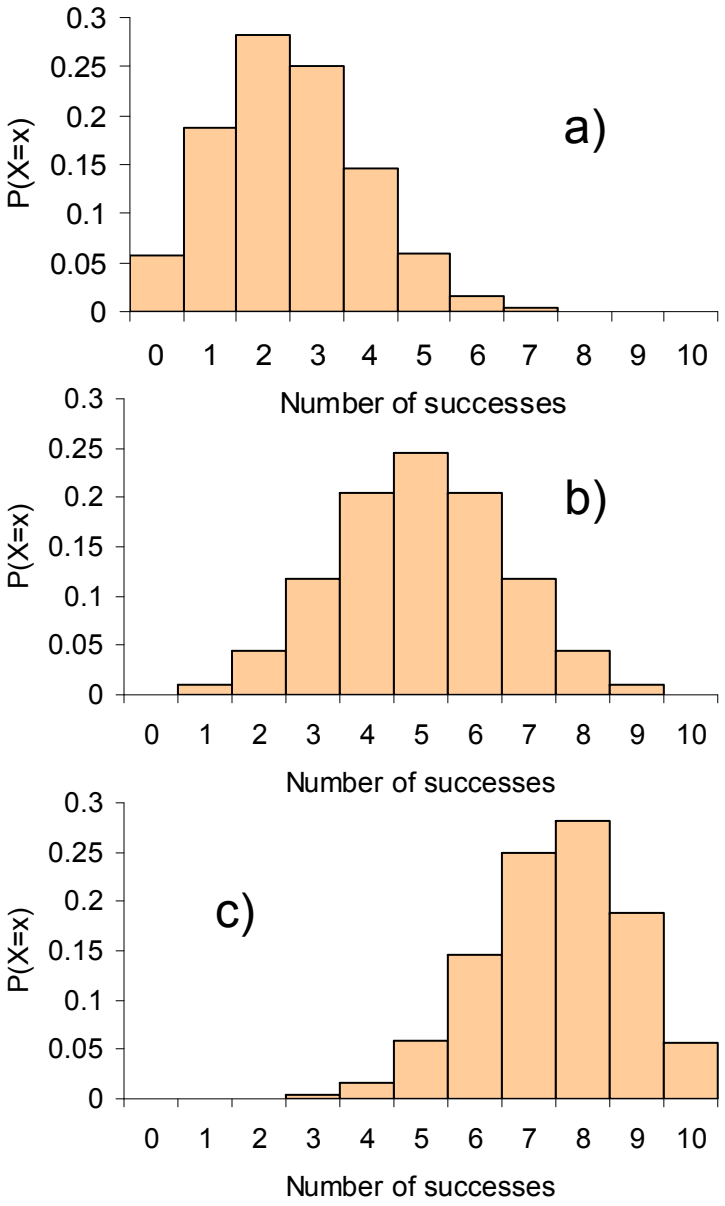
Microsoft Excel



Binomial mean and standard deviation

The center and spread of the binomial distribution for a count *X* are defined by the mean μ and standard deviation σ:

$$\mu = np \qquad \sigma = \sqrt{npq} = \sqrt{np(1-p)}$$



Effect of changing *p* when *n* is fixed.

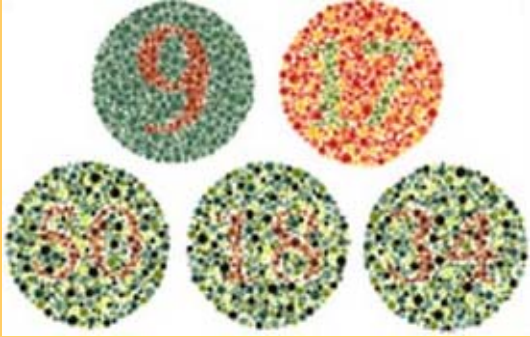
a) $n = 10$, $p = 0.25$

b) $n = 10$, $p = 0.5$

c) $n = 10$, $p = 0.75$

For small samples, binomial distributions are skewed when *p* is different from 0.5.

Color blindness



The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a random sample of size 25 from this population.

The population is definitely larger than 20 times the sample size, thus we can approximate the sampling distribution by $B(n = 25, p = 0.08)$.

- What is the probability that five individuals or fewer in the sample are color blind?

  Use Excel's "=BINOMDIST(number_*s*,trials,probability_*s*,cumulative)"

  $P(x \le 5)$ = BINOMDIST(5, 25, .08, 1) = 0.9877

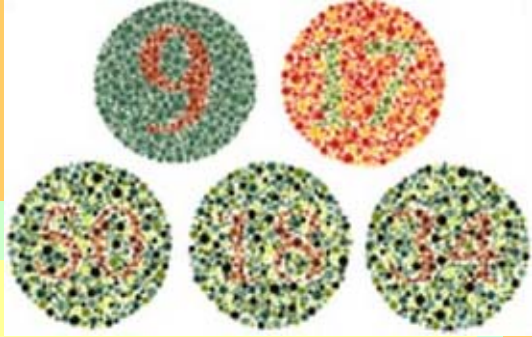
- What is the probability that more than five will be color blind?

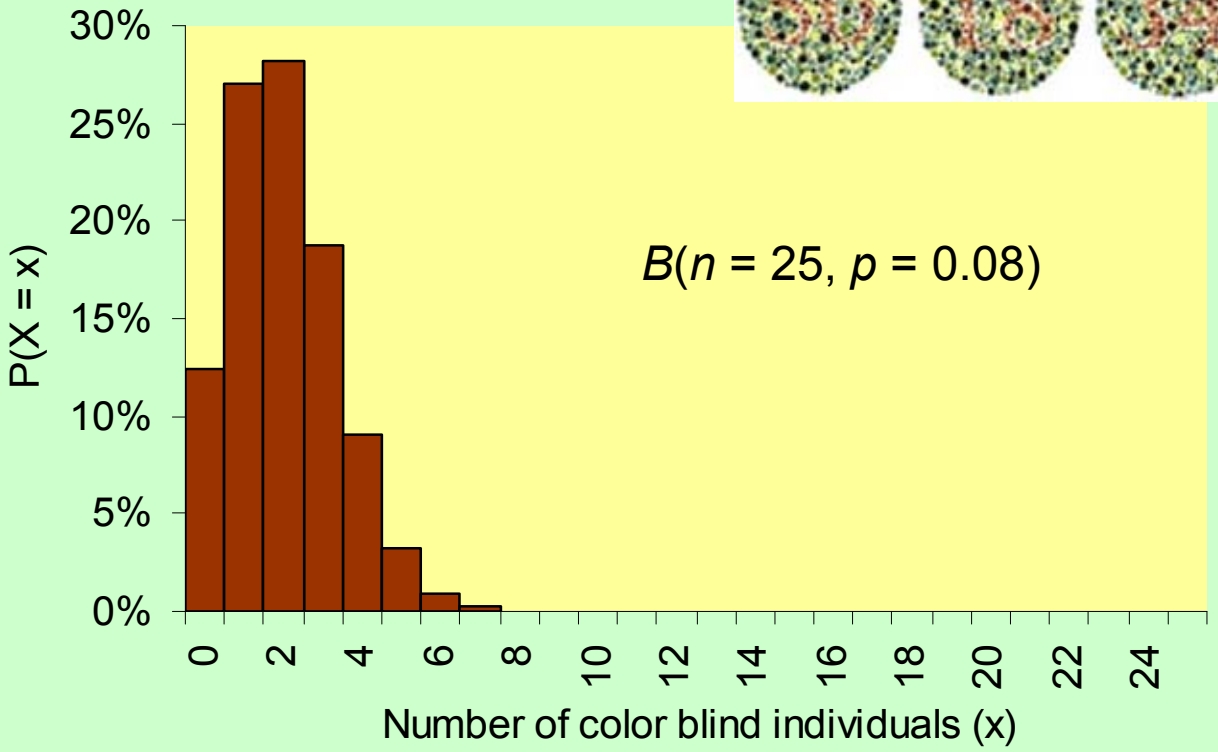
  $P(x > 5) = 1 - P(x \le 5) = 1 - 0.9666 = 0.0123$

- What is the probability that exactly five will be color blind?

  $P(x \le 5)$ = BINOMDIST(5, 25, .08, 0) = 0.0329

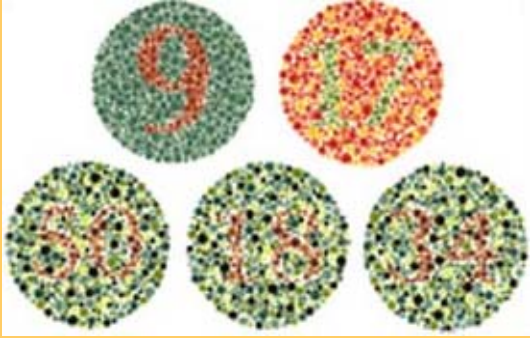
x	P(X = x)	P(X <= x)
0	12.44%	12.44%
1	27.04%	39.47%
2	28.21%	67.68%
3	18.81%	86.49%
4	9.00%	95.49%
5	3.29%	98.77%
6	0.95%	99.72%
7	0.23%	99.95%
8	0.04%	99.99%
9	0.01%	100.00%
10	0.00%	100.00%
11	0.00%	100.00%
12	0.00%	100.00%
13	0.00%	100.00%
14	0.00%	100.00%
15	0.00%	100.00%
16	0.00%	100.00%
17	0.00%	100.00%
18	0.00%	100.00%
19	0.00%	100.00%
20	0.00%	100.00%
21	0.00%	100.00%
22	0.00%	100.00%
23	0.00%	100.00%
24	0.00%	100.00%
25	0.00%	100.00%





Probability distribution and histogram for the number of color blind individuals among 25 Caucasian males.

What are the mean and standard deviation of the count of color blind individuals in the SRS of 25 Caucasian American males?



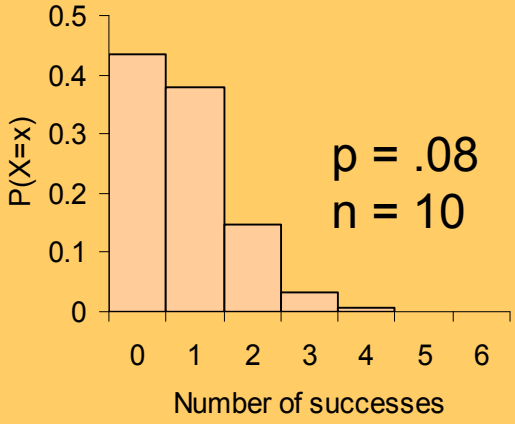
$\mu = np = 25*0.08 = 2$

$\sigma = \sqrt{np(1-p)} = \sqrt{(25*0.08*0.92)} = 1.36$

What if we take an SRS of size 10? Of size 75?

$\mu = 10*0.08 = 0.8$

$\sigma = \sqrt{(10*0.08*0.92)} = 0.86$



$\mu = 75*0.08 = 6$

$\sigma = \sqrt{(75*0.08*0.92)} = 3.35$

Sample proportions

The proportion of "successes" can be more informative than the count. In statistical sampling the sample proportion of successes, $\hat{p}$, is used to estimate the proportion *p* of successes in a population.

For any SRS of size *n,* the sample proportion of successes is:

$$\hat{p} = \frac{\text{count of successes in the sample}}{n} = \frac{X}{n}$$

- In an SRS of 50 students in an undergrad class, 10 are Hispanic:

  $\hat{p}$ = (10)/(50) = 0.2 *(proportion of Hispanics in sample)*

- The 30 subjects in an SRS are asked to taste an unmarked brand of coffee and rate it "would buy" or "would not buy." Eighteen subjects rated the coffee "would buy."

  $\hat{p}$ = (18)/(30) = 0.6 *(proportion of "would buy")*

If the sample size is much smaller than the size of a population with proportion *p* of successes, then the mean and standard deviation of $\hat{p}$ are:

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Because the mean is *p,* we say that the sample proportion in an SRS is an **unbiased estimator** of the population proportion *p*.

- The variability decreases as the sample size increases. So larger samples usually give closer estimates of the population proportion *p*.

Normal approximation

If n is large, and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution $N(\mu = np, \sigma^2 = np(1 - p))$. Practically, the Normal approximation can be used when both $np \geq 10$ and $n(1 - p) \geq 10$.

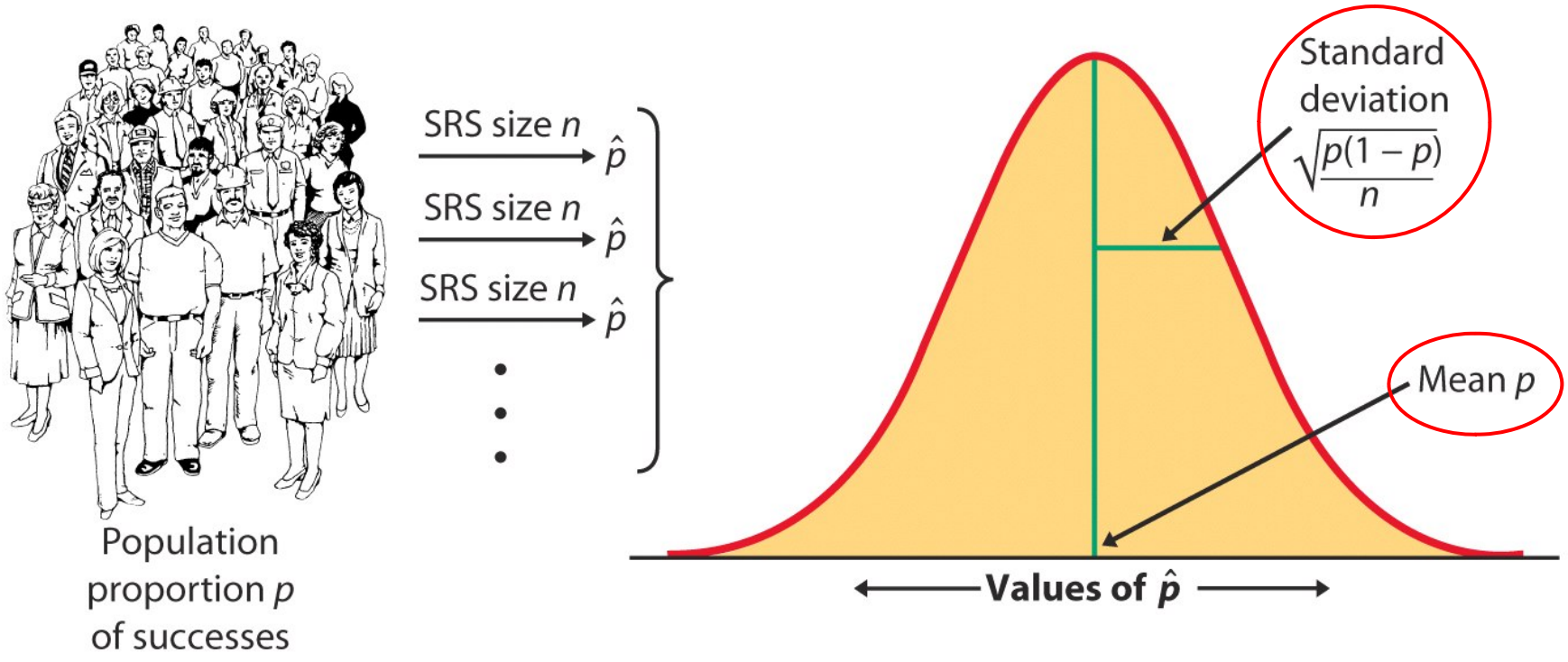
If X is the count of successes in the sample and $\hat{p} = X/n$, the sample proportion of successes, their sampling distributions for large n, are:

- **X approximately $N(\mu = np, \sigma^2 = np(1 - p))$**
- **$\hat{p}$ is approximately $N(\mu = p, \sigma^2 = p(1 - p)/n)$**

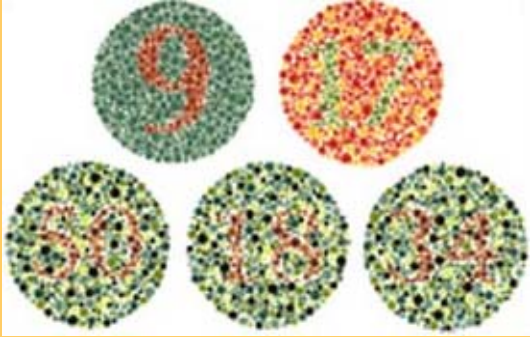
Sampling distribution of the sample proportion

The sampling distribution of $\hat{p}$ is never exactly normal. But as the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.

The normal approximation is most accurate for any fixed *n* when *p* is close to 0.5, and least accurate when *p* is near 0 or near 1.



Color blindness



The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is about 8%.

We take a random sample of size 125 from this population. What is the probability that six individuals or fewer in the sample are color blind?

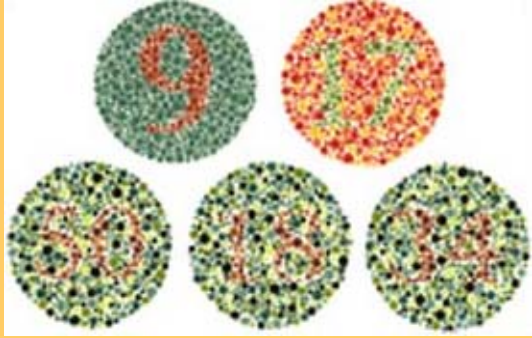
- Sampling distribution of the count X: $B(n = 125, p = 0.08) \rightarrow np = 10$
  $P(X \le 6)$ = BINOMDIST(6, 125, .08, 1) = 0.1198 or about 12%

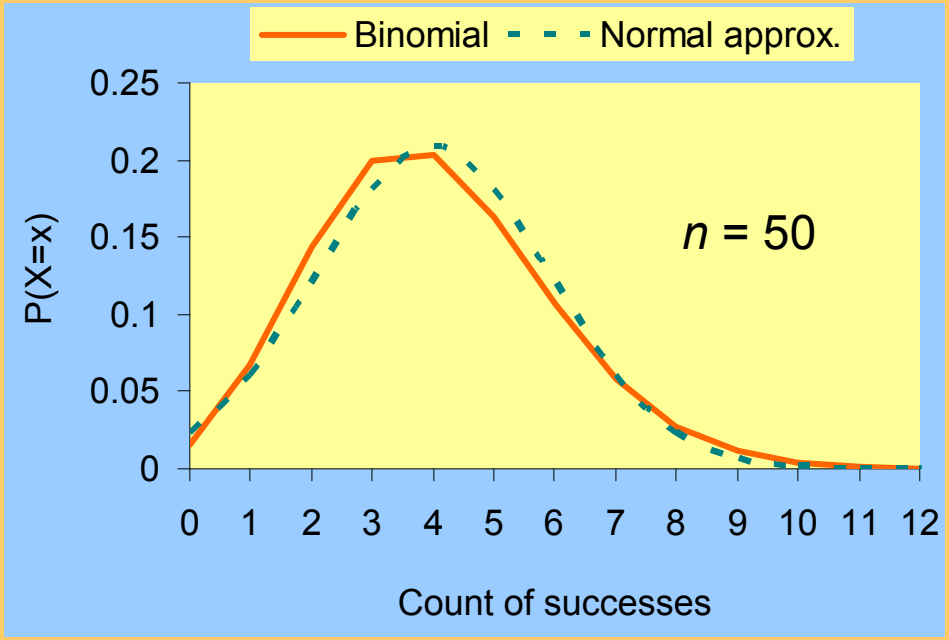
- Normal approximation for the count X: $N(\text{np} = 10, \sqrt{np(1-p)} = 3.033)$
  $P(X \le 6)$ = NORMDIST(6, 10, 3.033, 1) = 0.0936 or 9%
  Or $z = (x - \mu)/\sigma = (6 - 10)/3.033 = -1.32 \rightarrow P(X \le 6) = 0.0934$ from Table A

The normal approximation is reasonable, though not perfect. Here $p = 0.08$ is not close to 0.5 when the normal approximation is at its best.

A sample size of 125 is the smallest sample size that can allow use of the normal approximation ($np = 10$ and $n(1 - p) = 115$).

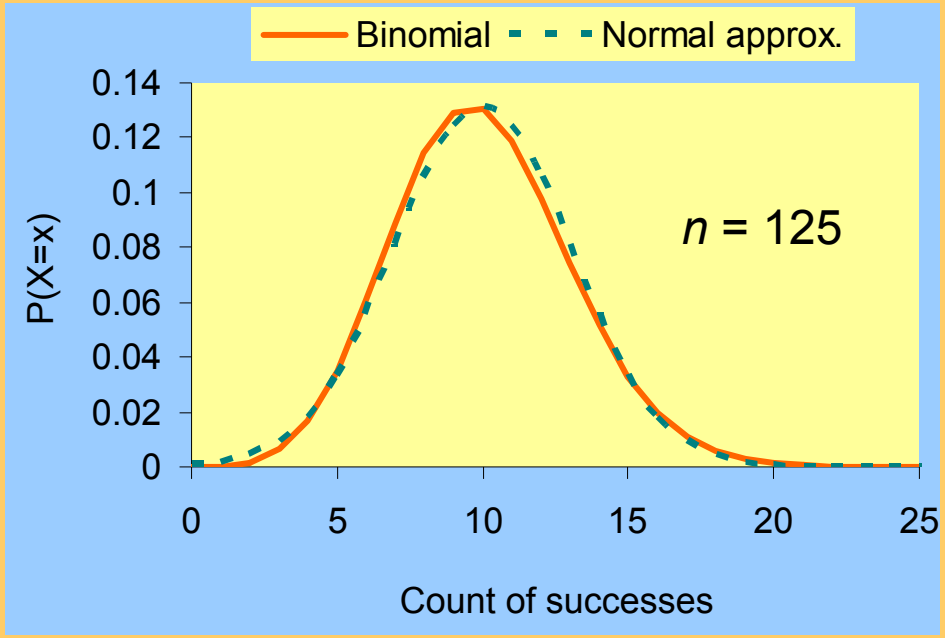
Sampling distributions for the color blindness example.

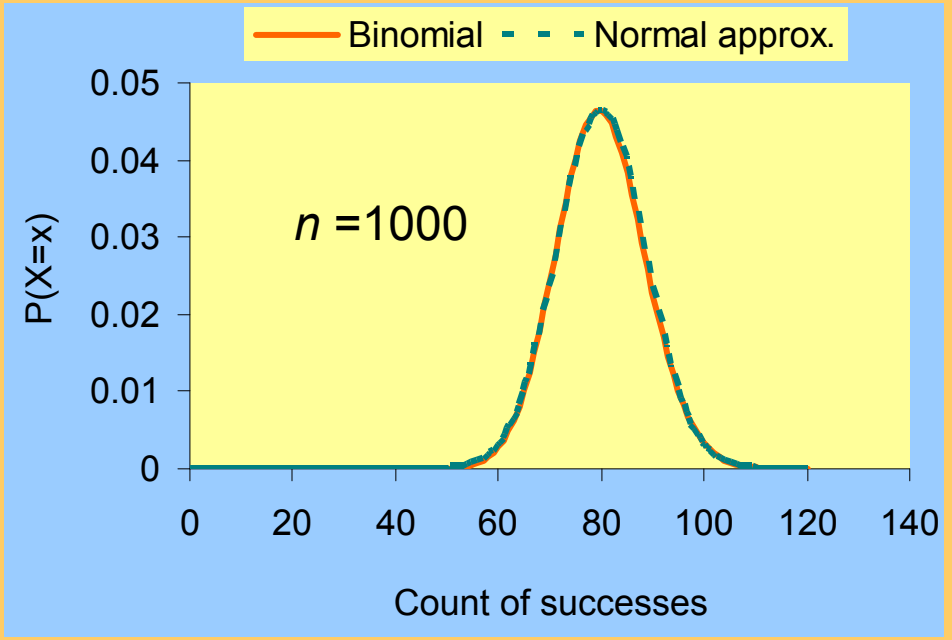




The larger the sample size, the better the normal approximation suits the binomial distribution.

Avoid sample sizes too small for np or $n(1 - p)$ to reach at least 10 (e.g., n = 50).





Normal approximation: continuity correction

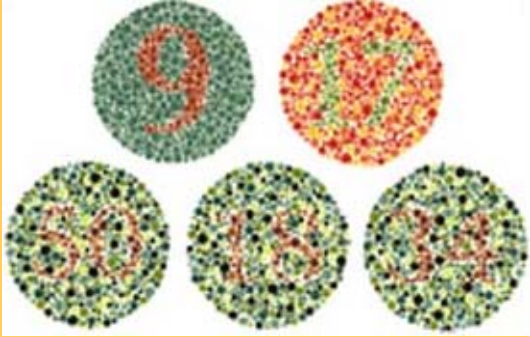
The normal distribution is a better approximation of the binomial distribution, if we perform a continuity correction where $x' = x + 0.5$ is substituted for x, and $P(X \leq x)$ is replaced by $P(X \leq x + 0.5)$.

Why? A binomial random variable is a discrete variable that can only take whole numerical values. In contrast, a normal random variable is a continuous variable that can take any numerical value.

$P(X \leq 10)$ for a binomial variable is $P(X \leq 10.5)$ using a normal approximation.

$P(X < 10)$ for a binomial variable excludes the outcome $X = 10$, so we exclude the entire interval from 9.5 to 10.5 and calculate $P(X \leq 9.5)$ when using a normal approximation.

Color blindness



The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is about 8%. We take a random sample of size 125 from this population.

- Sampling distribution of the count X: $B(n = 125, p = 0.08) \rightarrow np = 10$
  $P(X \le 6.5) = P(X \le 6)$ = BINOMDIST(6, 125, .08, 1) = 0.1198
  $P(X < 6) = P(X \le 5)$ = BINOMDIST(5, 125, .08, 1) = 0.0595

- Normal approximation for the count X: $N(np = 10, \sqrt{np(1-p)} = 3.033)$
  $P(X \le 6.5)$ = NORMDIST(6.5, 10, 3.033, 1) = 0.1243
  $P(X \le 6)$ = NORMDIST(6, 10, 3.033, 1) = 0.0936 $\neq P(X \le 6.5)$
  $P(X < 6) = P(X \le 6)$ = NORMDIST(6, 10, 3.033, 1) = 0.0936

The continuity correction provides a more accurate estimate:

- Binomial $P(X \le 6) = 0.1198$ → *this is the exact probability*
- Normal $P(X \le 6) = 0.0936$, while $P(X \le 6.5) = 0.1243$ → *estimates*

Binomial formulas

The number of ways of arranging *k* successes in a series of *n* observations (with constant probability *p* of success) is the number of possible combinations (unordered sequences).

This can be calculated with the **binomial coefficient**:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Where k = 0, 1, 2, ..., or n.

Binomial formulas

- The binomial coefficient "*n*_choose_*k*" uses the **factorial** notation "**!**".
- The factorial *n*! for any strictly positive whole number *n* is:

$$n! = n \times (n - 1) \times (n - 2) \times \cdots \times 3 \times 2 \times 1$$

- For example: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
- Note that $0! = 1$.

Calculations for binomial probabilities

The binomial coefficient counts the number of ways in which k successes can be arranged among n observations.

The **binomial probability *P*(*X* = *k*)** is this count multiplied by the probability of any specific arrangement of the *k* successes:

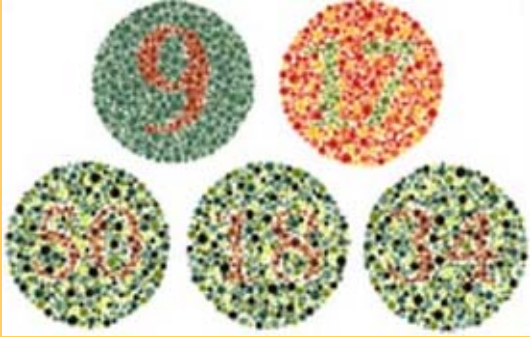
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

X	*P*(*X*)
0	${}_nC_0\, p^0 q^n = q^n$
1	${}_nC_1\, p^1 q^{n-1}$
2	${}_nC_2\, p^2 q^{n-2}$
…	…
k	${}_nC_x\, p^k q^{n-k}$
…	…
n	${}_nC_n\, p^n q^0 = p^n$
Total	*1*

The probability that a binomial random variable takes any range of values is the sum of each probability for getting exactly that many successes in *n* observations.

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

Color blindness



The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a random sample of size 25 from this population.

What is the probability that exactly five individuals in the sample are color blind?

- Use Excel's "=BINOMDIST(number_*s*,trials,probability_*s*,cumulative)"

  $P(x = 5)$ = BINOMDIST(5, 25, 0.08, 0) = 0.03285

- $P(x = 5) = (n! / k!(n - k)!)p^k(1 - p)^{n-k} = (25! / 5!(20)!)\ 0.08^5 0.92^5$

  $P(x = 5) = (21*22*23*24*24*25 / 1*2*3*4*5)\ 0.08^5 0.92^{20}$

  $P(x = 5) = 53{,}130 * 0.0000033 * 0.1887 = 0.03285$

Sampling Distributions
for Sample Means

IPS Chapter 5.2

Objectives (IPS Chapter 5.2)

Sampling distribution of a sample mean

- The mean and standard deviation of $\bar{x}$
- For normally distributed populations
- The central limit theorem
- Weibull distributions

Reminder: What is a sampling distribution?

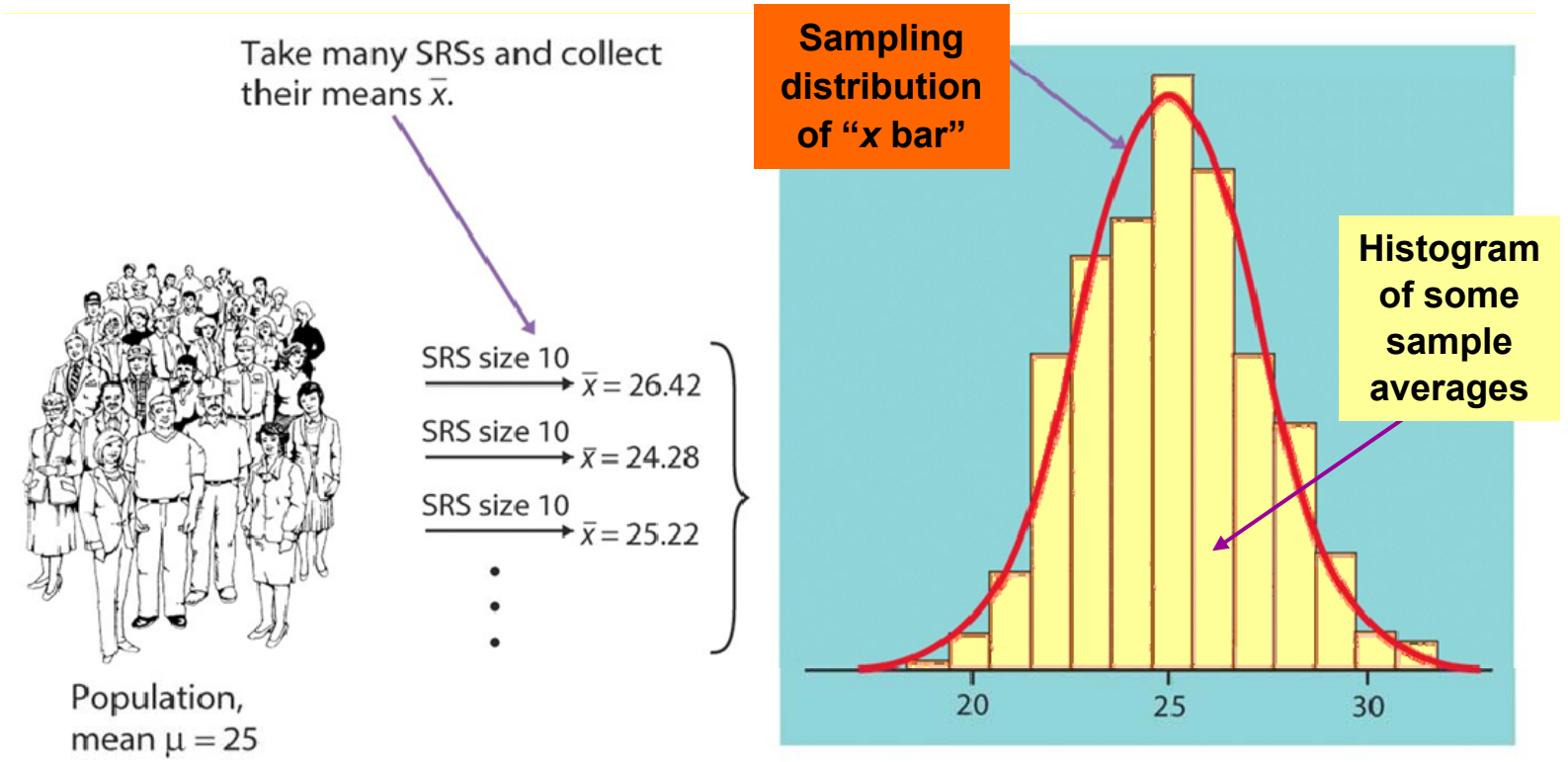
The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population. It is a theoretical idea — we do not actually build it.

The sampling distribution of a statistic is the **probability distribution** of that statistic.

Sampling distribution of the sample mean

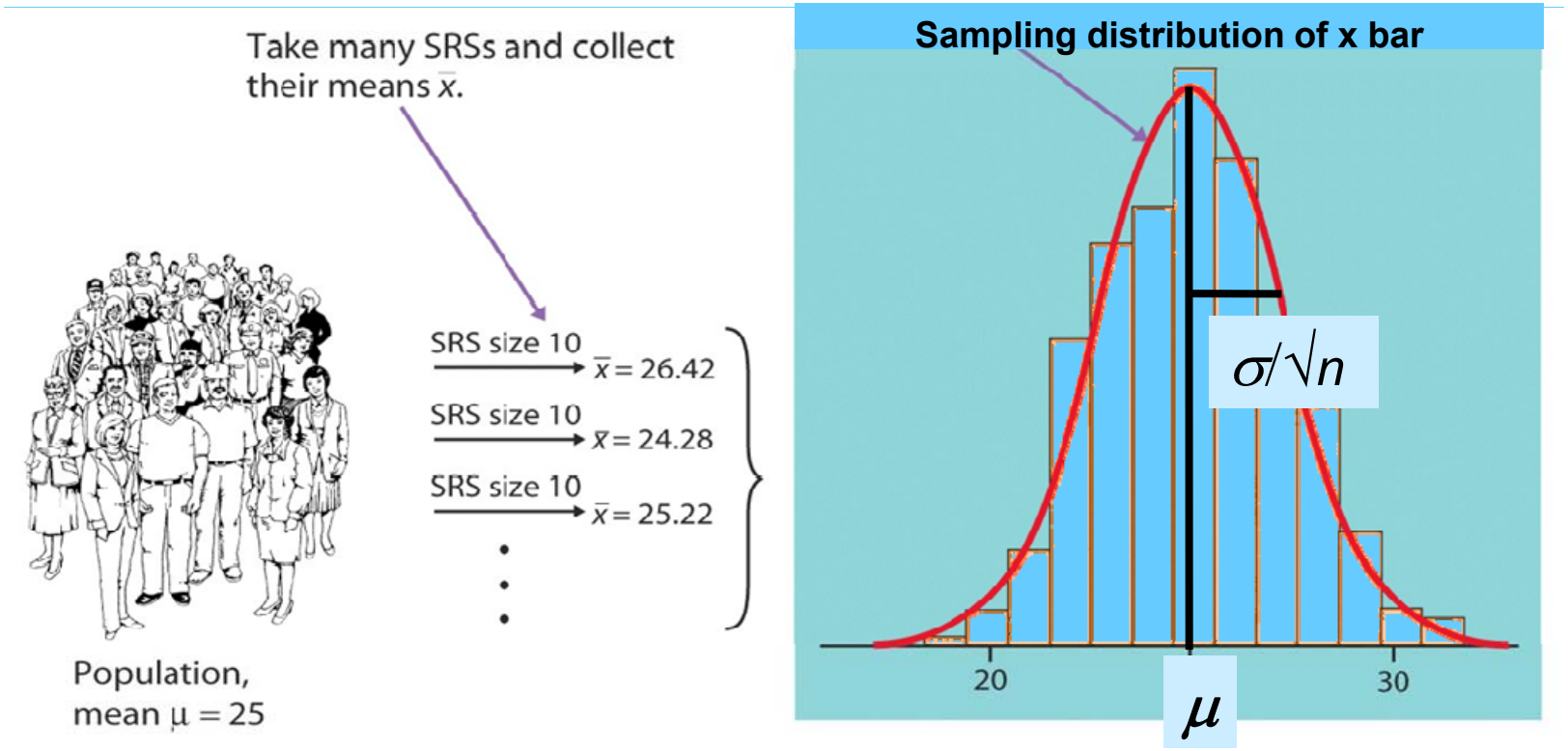
We take many random samples of a given size *n* from a population with mean μ and standard deviation σ.

Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



For any population with mean μ and standard deviation σ:

- The **mean**, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ : $\mu_{\bar{x}} = \mu$.
- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size : $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.



- Mean of a sampling distribution of $\bar{x}$

  There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an **unbiased estimate** of the population mean μ — it will be "correct on average" in many samples.

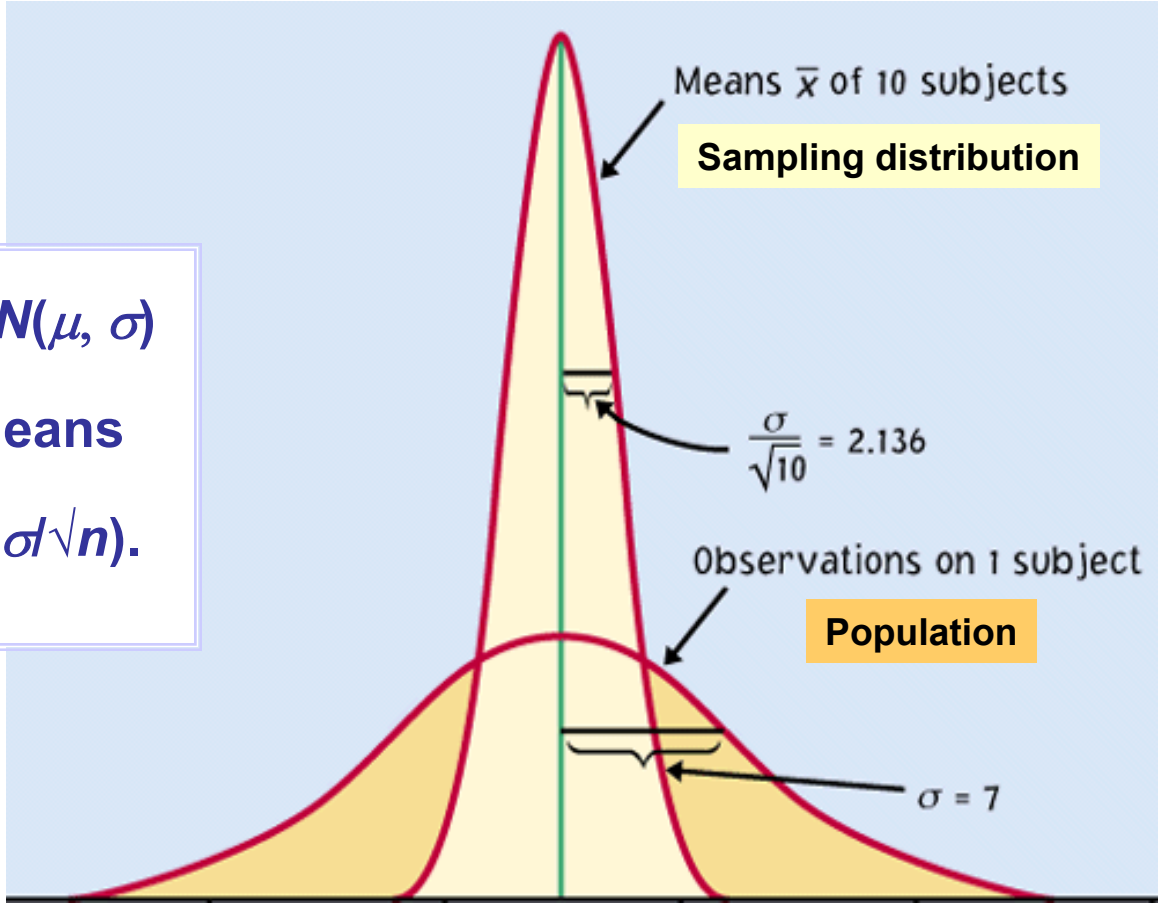
- Standard deviation of a sampling distribution of $\bar{x}$

  The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$. ➔ **Averages are less variable than individual observations**.

For normally distributed populations

When a variable in a population is normally distributed, the sampling distribution of $\bar{x}$ for all possible samples of size *n* is also normally distributed.

If the population is $N(\mu, \sigma)$ then the sample means distribution is $N(\mu, \sigma/\sqrt{n})$.



IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is:

A) Exactly normal, mean 112, standard deviation 20

B) Approximately normal, mean 112, standard deviation 20

C) Approximately normal, mean 112 , standard deviation 1.414

D) Approximately normal, mean 112, standard deviation 0.1

C) Approximately normal, mean 112 , standard deviation 1.414

Population distribution : $N(\mu = 112; \sigma = 20)$

Sampling distribution for $n = 200$ is $N(\mu = 112; \sigma/\sqrt{n} = 1.414)$

Application

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what is the probability that this patient will be misdiagnosed with Hypokalemia?

$$z = \frac{(x - \mu)}{\sigma} = \frac{3.5 - 3.8}{0.2}$$

$z = -1.5$, $P(z < -1.5) = 0.0668 \approx 7\%$

Instead, if measurements are taken on 4 separate days, what is the probability of a misdiagnosis?

$$z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} = \frac{3.5 - 3.8}{0.2/\sqrt{4}}$$

$z = -3$, $P(z < -1.5) = 0.0013 \approx 0.1\%$

Note: Make sure to standardize (z) using the standard deviation for the sampling distribution.

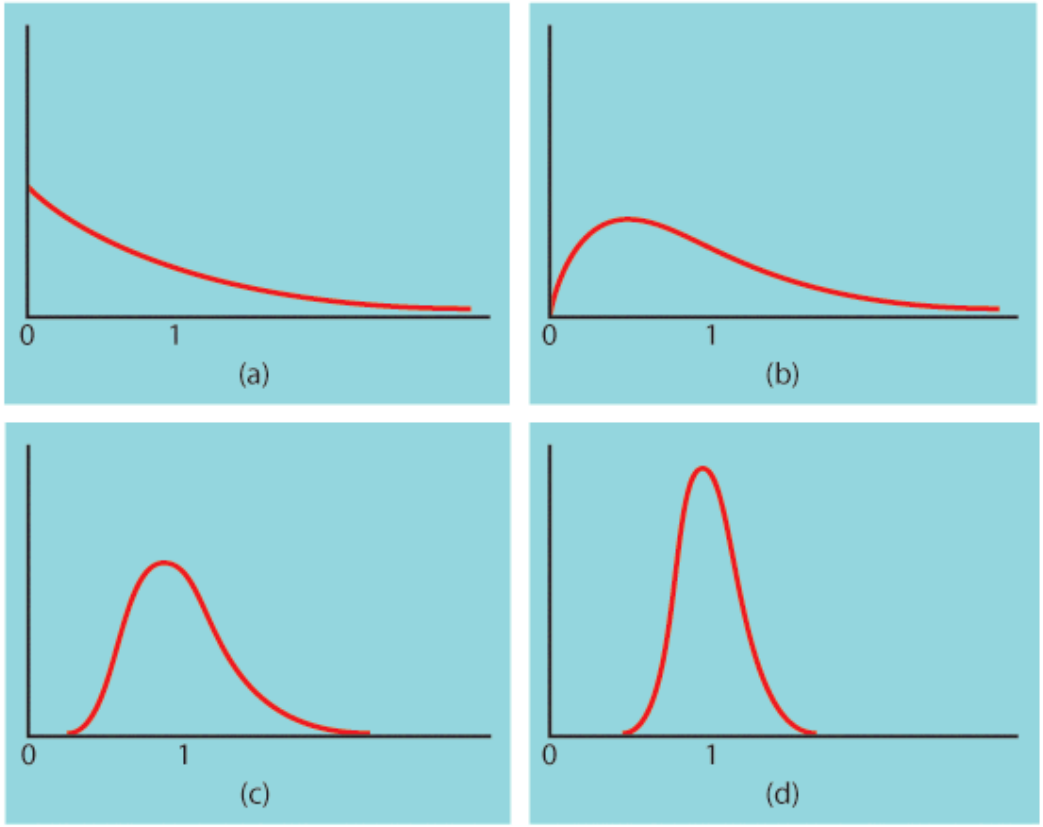
Practical note

- Large samples are not always attainable.
  - Sometimes the cost, difficulty, or preciousness of what is studied drastically limits any possible sample size.
  - Blood samples/biopsies: No more than a handful of repetitions are acceptable. Oftentimes, we even make do with just one.
  - Opinion polls have a limited sample size due to time and cost of operation. During election times, though, sample sizes are increased for better accuracy.
- Not all variables are normally distributed.
  - Income, for example, is typically strongly skewed.
  - Is $\bar{x}$ still a good estimator of μ then?

The central limit theorem

Central Limit Theorem: When randomly sampling from **any** population with mean μ and standard deviation σ, **when *n* is large enough,** the sampling distribution of $\bar{x}$ is approximately normal: **~ $N(\mu, \sigma/\sqrt{n})$**.

Population with strongly skewed distribution



Sampling distribution of $\bar{x}$ for n = 2 observations

Sampling distribution of $\bar{x}$ for n = 10 observations

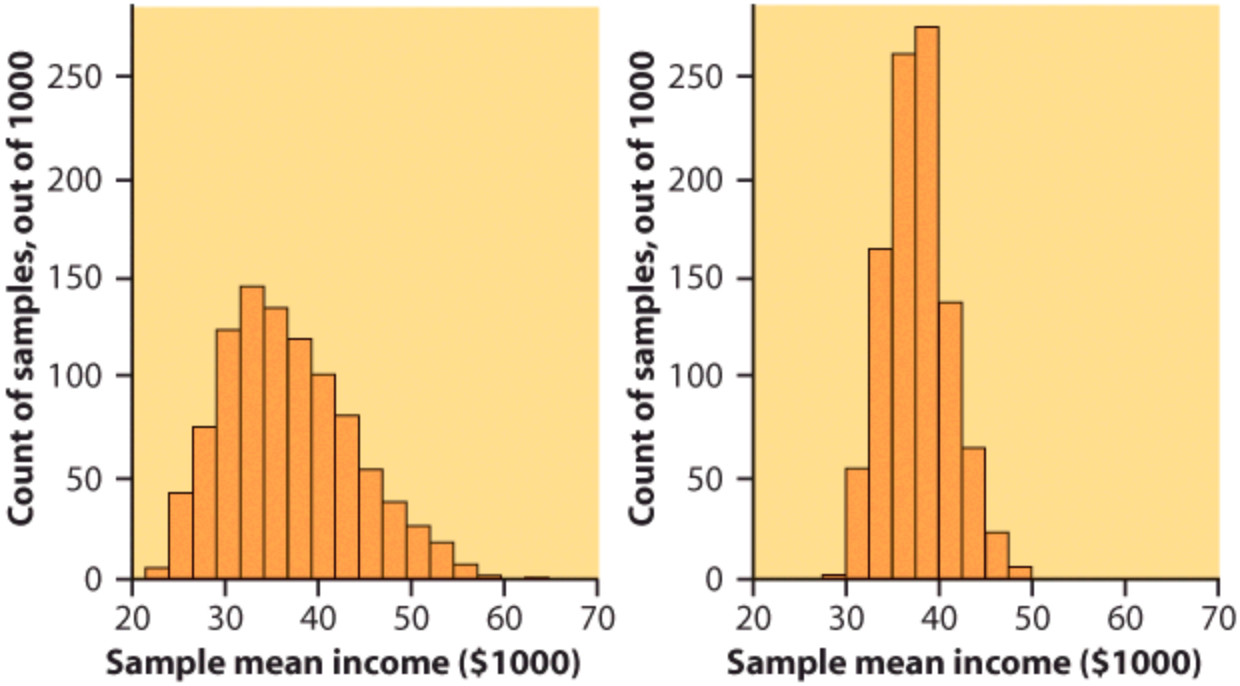
Sampling distribution of $\bar{x}$ for n = 25 observations

Income distribution

Let's consider the very large database of individual incomes from the Bureau of Labor Statistics as our population. It is strongly right skewed.

- We take 1000 SRSs of 100 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.
- We also take 1000 SRSs of 25 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.

Which histogram corresponds to samples of size 100? 25?



How large a sample size?

It depends on the population distribution. More observations are required if the population distribution is far from normal.

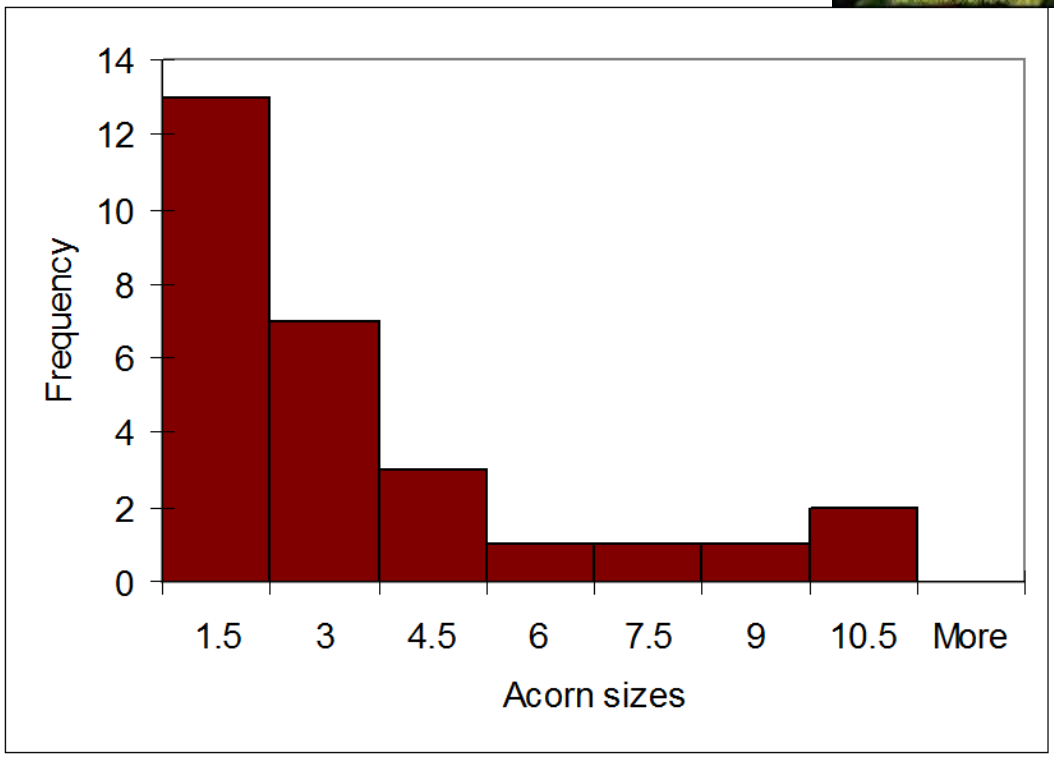
- A sample size of 25 is generally enough to obtain a normal sampling distribution from a strong skewness or even mild outliers.
- A sample size of 40 will typically be good enough to overcome extreme skewness and outliers.

In many cases, n = 25 isn't a huge sample. Thus, even for strange population distributions we can assume a normal sampling distribution of the mean and work with it to solve problems.

Sampling distributions

Atlantic acorn sizes (in cm^3)

— sample of 28 acorns:



- Describe the histogram. What do you assume for the population distribution?

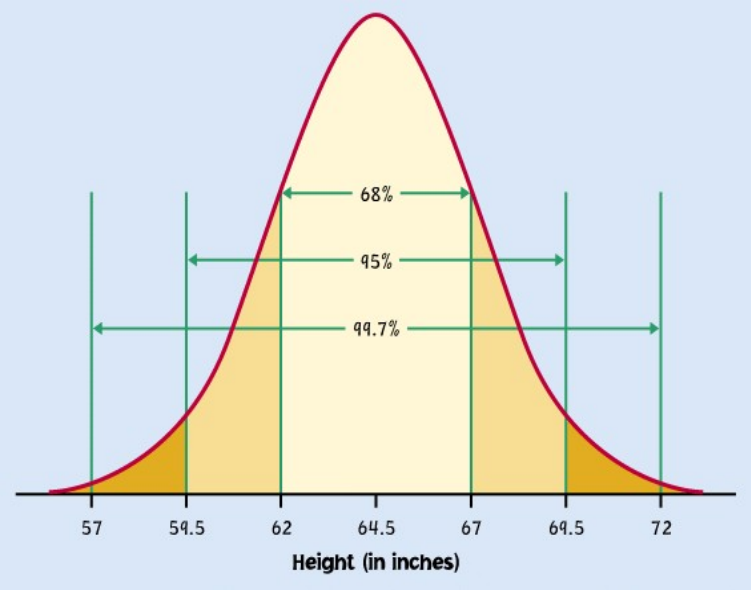
- What would be the shape of the sampling distribution of the mean:
  - For samples of size 5?
  - For samples of size 15?
  - For samples of size 50?

Further properties

Any linear combination of independent random variables is also normally distributed.

More generally, the central limit theorem is valid as long as we are sampling many small random events, even if the events have different distributions (as long as no one random event dominates the others).

Why is this cool? It explains why the normal distribution is so common.



Example: Height seems to be determined by a large number of genetic and environmental factors, like nutrition. The "individuals" are genes and environmental factors. Your height is a mean.

Weibull distributions

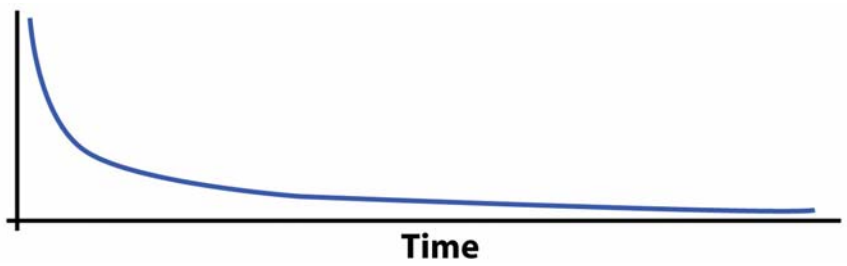
There are many probability distributions beyond the binomial and normal distributions used to model data in various circumstances.

Weibull distributions are used to model **time to failure/product lifetime** and are common in engineering to study product reliability.

Product lifetimes can be measured in units of time, distances, or number of cycles for example. Some applications include:

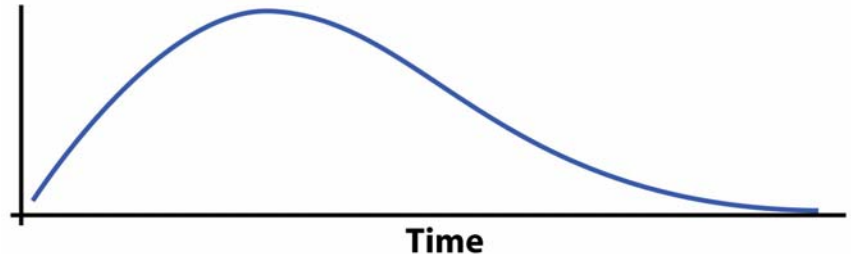
- Quality control (breaking strength of products and parts, food shelf life)
- Maintenance planning (scheduled car revision, airplane maintenance)
- Cost analysis and control (number of returns under warranty, delivery time)
- Research (materials properties, microbial resistance to treatment)

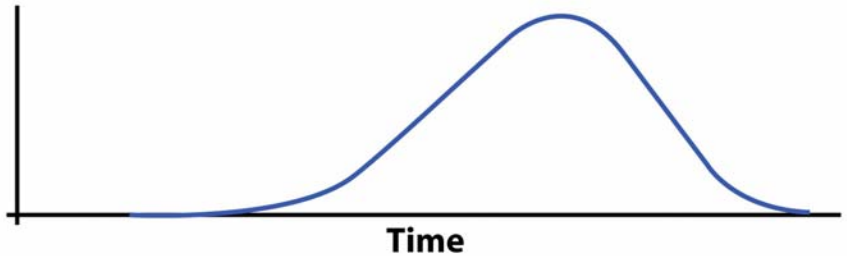
Density curves of three members of the Weibull family describing a different type of product time to failure in manufacturing:



Infant mortality: Many products fail immediately and the remainders last a long time. Manufacturers only ship the products after inspection.

Early failure: Products usually fail shortly after they are sold. The design or production must be fixed.





Old-age wear out: Most products wear out over time, and many fail at about the same age. This should be disclosed to customers.