What is Statistics?

Definition of Statistics

- **Statistics** is the science of collecting, organizing, analyzing, and interpreting data in order to make a decision.

- **Branches of Statistics**
  - The study of statistics has two major branches – descriptive(exploratory) statistics and inferential statistics.
    - **Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data.
    - **Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about population. A basic tool in the study of inferential statistics is probability.

Scatterplots and Correlation

- **Displaying relationships: Scatterplots**

- **Interpreting scatterplots**

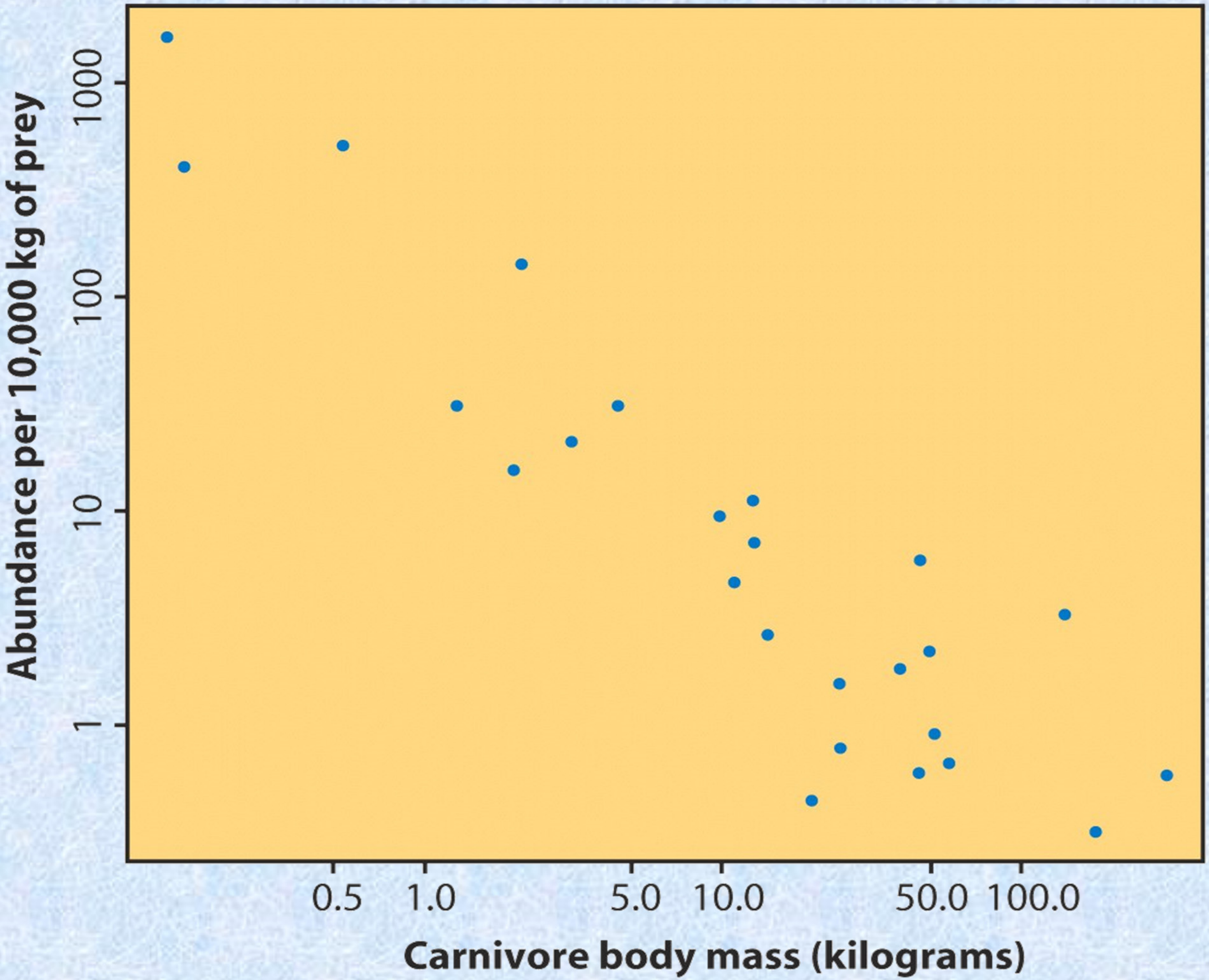
- **Adding categorical variables to scatterplots**

- **Measuring linear association: correlation *r***

- **Facts about correlation**

- Response variable measures **an outcome of a study**.
- An explanatory variable explains, **influences or cause changes in a response variable**.
- Independent variable and dependent variable.
- WARNING: The relationship between two variables can be strongly influenced by other variables that are lurking in the background.
- **Note:** There is not necessary to have a cause-and-effect relationship between explanatory and response variables.
- Example. Sales of personal computers and athletic shoes

Example - 1



1000
100
10
1
Abundance per 10,000 kg of prey
0.5
1.0
5.0
10.0
50.0
100.0
Carnivore body mass (kilograms)

Definitions

- **Sample space**: the set of all possible outcomes. We denote S
- **Event**: an outcome or a set of outcomes of a random phenomenon. An event is a subset of the sample space.
- **Probability** is the proportion of success of an event.
- **Probability model**: a mathematical description of a random phenomenon consisting of two parts: S and a way of assigning probabilities to events.

Probability distributions

- **Probability distribution of a random variable X**: it tells what values X can take and how to assign probabilities to those values.
  - Probability of discrete random variable: list of the possible value of X and their probabilities
  - Probability of continuous random variable: density curve.

Measuring linear association: correlation *r*

(The *Pearson Product-Moment Correlation Coefficient or Correlation Coefficient***)**

- The **correlation r** measures the strength and direction of the ***linear association*** between two quantitative variables, usually labeled X and Y.

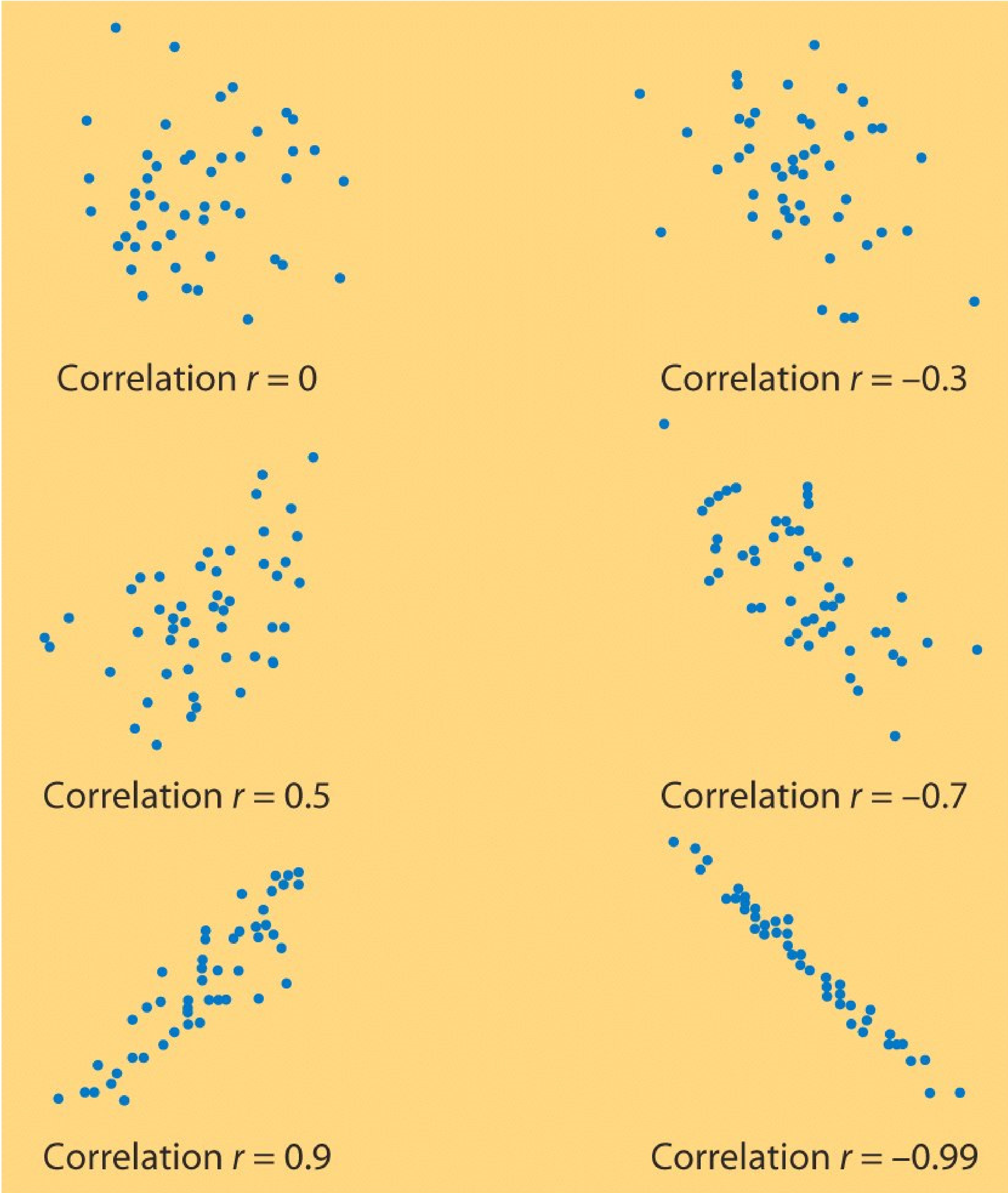
$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Facts about correlation

- What kind of variables do we use?
  - 1. No distinction between explanatory and response variables.
  - 2. Both variables should be quantitative
- Numerical properties
  - 1. $-1 \le r \le 1$
  - 2. r>0: positive association between variables
  - 3. r<0: negative association between variables
  - 4. If r =1or r = - 1, it indicates perfect linear relationship
  - 5. As |r| is getting close to 1, much stronger relationship

$$\begin{array}{ccc} \lhd - \textit{negative relationship} - \rhd & \lhd - \textit{positive relationship} - \rhd & \\ -1 & 0 & 1 \\ \lhd ---- \textit{stronger} & & \textit{stronger} ---- \rhd \end{array}$$

  - 6. Effected by a few outliers →not resistant.
  - 7. It doesn’t describe curved relationships
  - 8. Not easy to guess the value of r from the appearance of a scatter plot


Correlation $r = 0$
Correlation $r = -0.3$
Correlation $r = 0.5$
Correlation $r = -0.7$
Correlation $r = 0.9$
Correlation $r = -0.99$

Some necessary elements of

Probability theory and Statistics

The NORMAL DISTRIBUTION

The normal (or Gaussian) distribution, is a very commonly used (occurring) function in the fields of probability theory, and has wide applications in the fields of:

- Pattern Recognition;
- Machine Learning;
- Artificial Neural Networks and Soft computing;
- Digital Signal (image, sound , video etc.) processing
- Vibrations, Graphics etc.

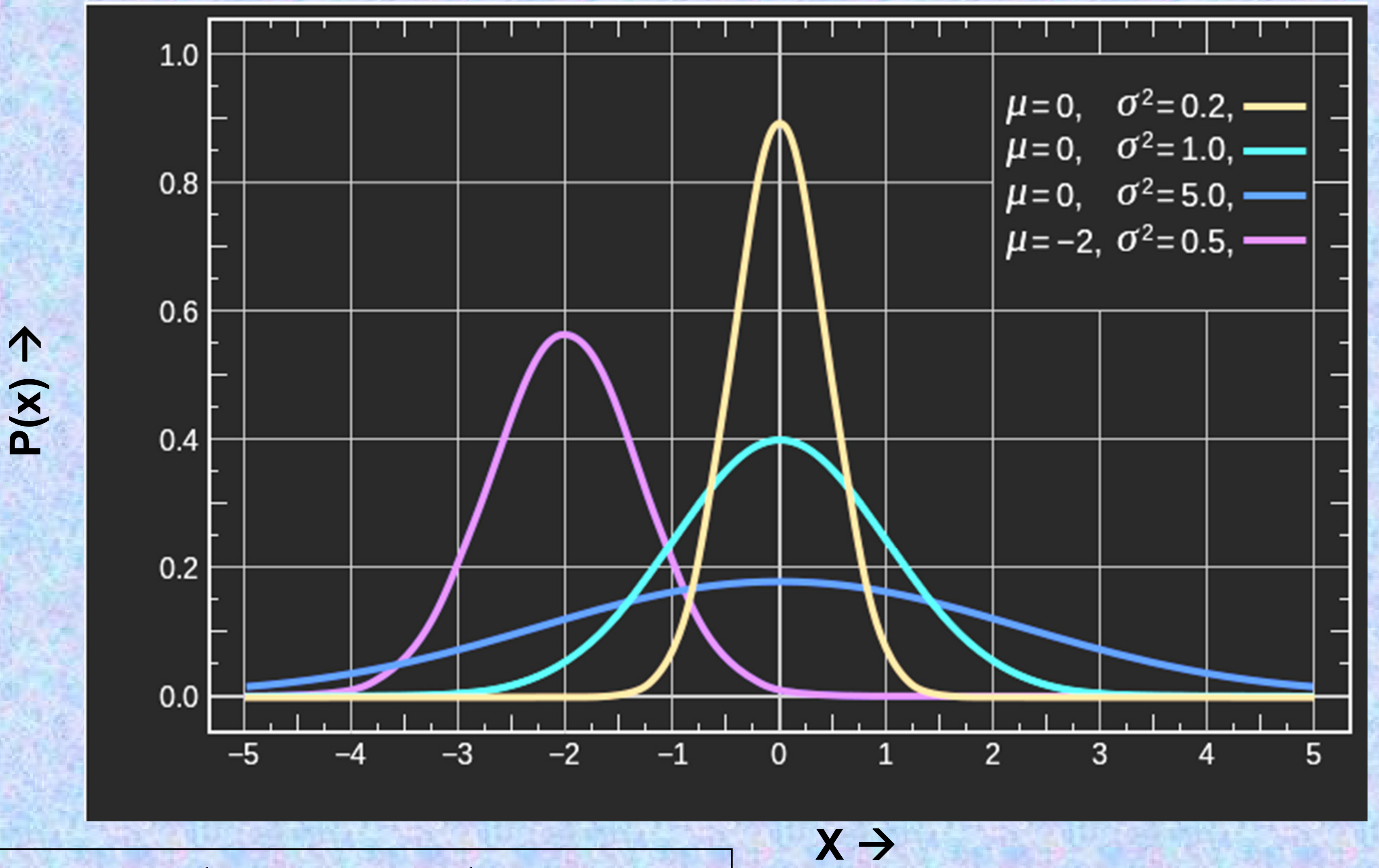
Its also called a BELL function/curve.

The formula for the normal distribution is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

The parameter μ is called the mean or expectation (or median or mode) of the distribution.

The parameter σ is the standard deviation; and variance is thus σ^2.



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

https://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg
(2013)

The normal distribution $p(x)$, with any mean μ and any positive deviation σ, has the following properties:

- **It is symmetric around the mean (μ) of the distribution.**
- **It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$.**
- **It has two inflection points (where the second derivative of f is zero and changes sign), located one standard deviation away from the mean, $x = \mu - \sigma$ and $x = \mu + \sigma$.**
- **It is log-concave.**
- **It is infinitely differentiable, indeed supersmooth of order 2.**

Also, the standard normal distribution p (with $\mu = 0$ and $\sigma = 1$) also has the following properties:

- **Its first derivative $p'(x)$ is: $-x.p(x)$.**
- **Its second derivative $p''(x)$ is: $(x^2 - 1).p(x)$**
- **More generally, its *n*-th derivative :**

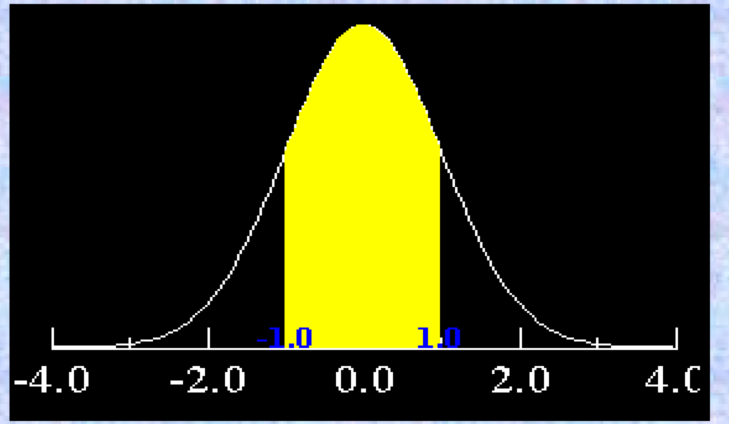
$p^{(n)}(x)$ is: $(-1)^n H_n(x) p(x)$,

where, H_n is the Hermite polynomial of order n.

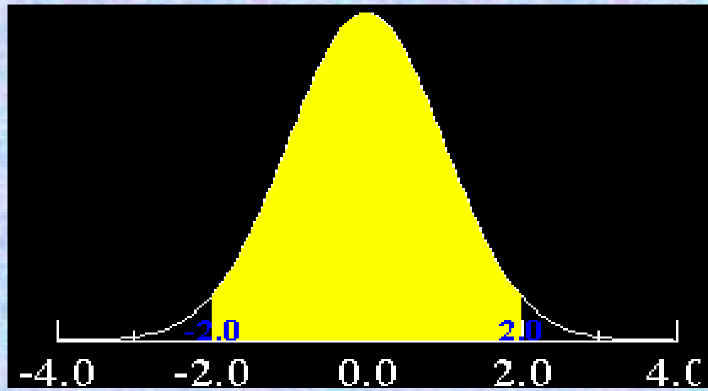
The 68 – 95 - 99.7% Rule:

All normal density curves satisfy the following property which is often referred to as the Empirical Rule:

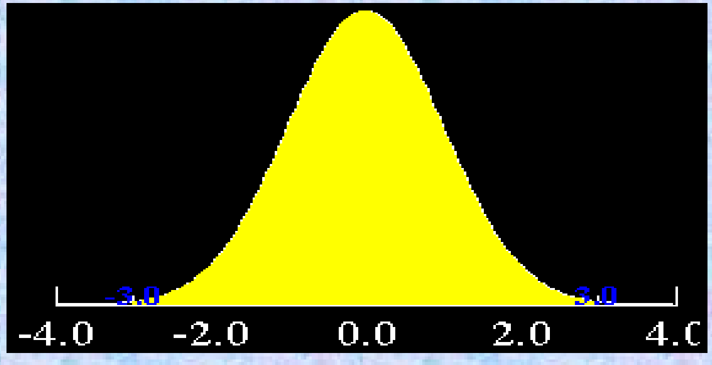
- 68% of the observations fall within 1 standard deviation of the mean, that is, between $(\mu - \sigma)$ and $(\mu + \sigma)$

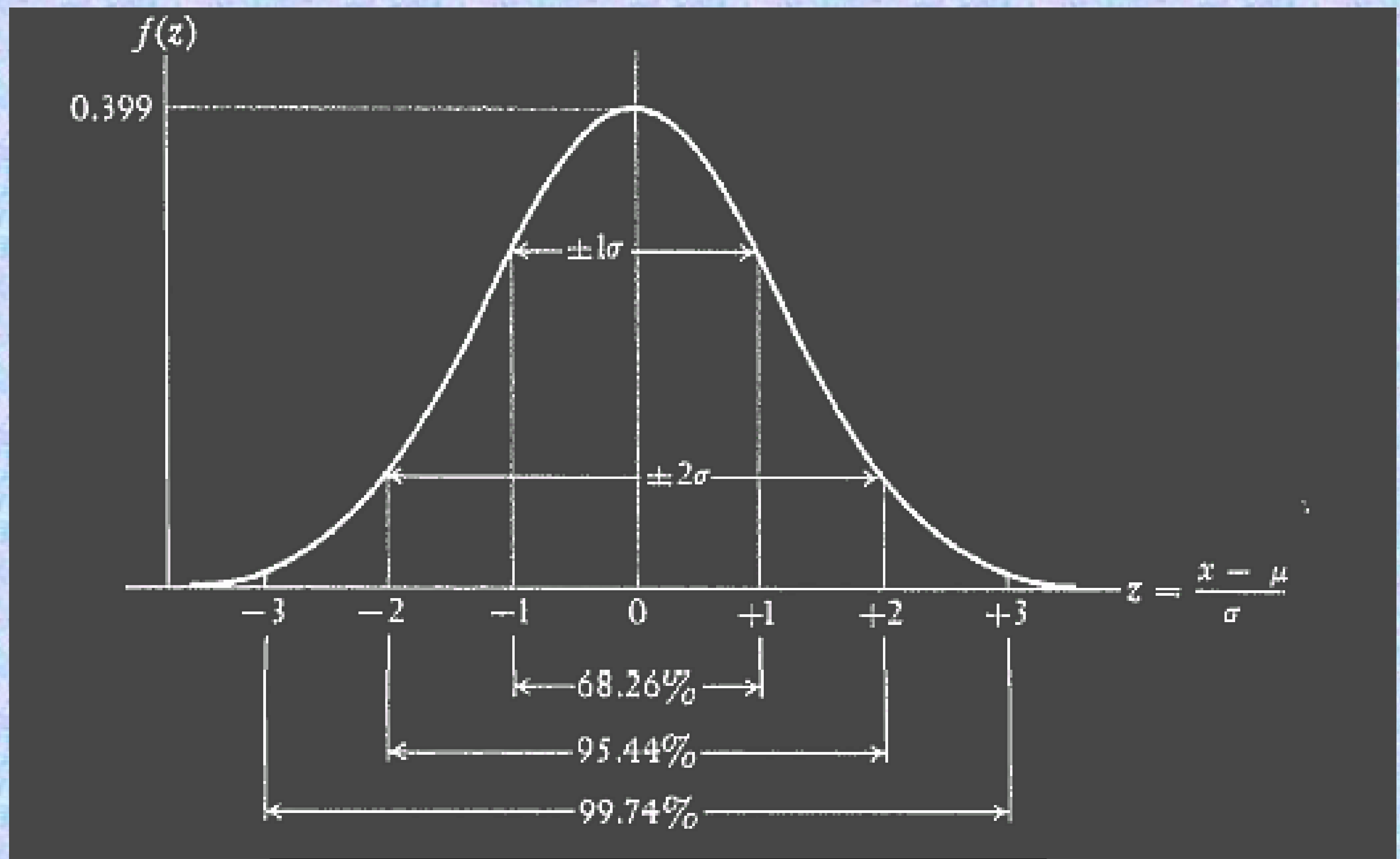


- 95% of the observations fall within 2 standard deviations of the mean, that is, between $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$



- 99.7% of the observations fall within 3 standard deviations of the mean, that is, between $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$



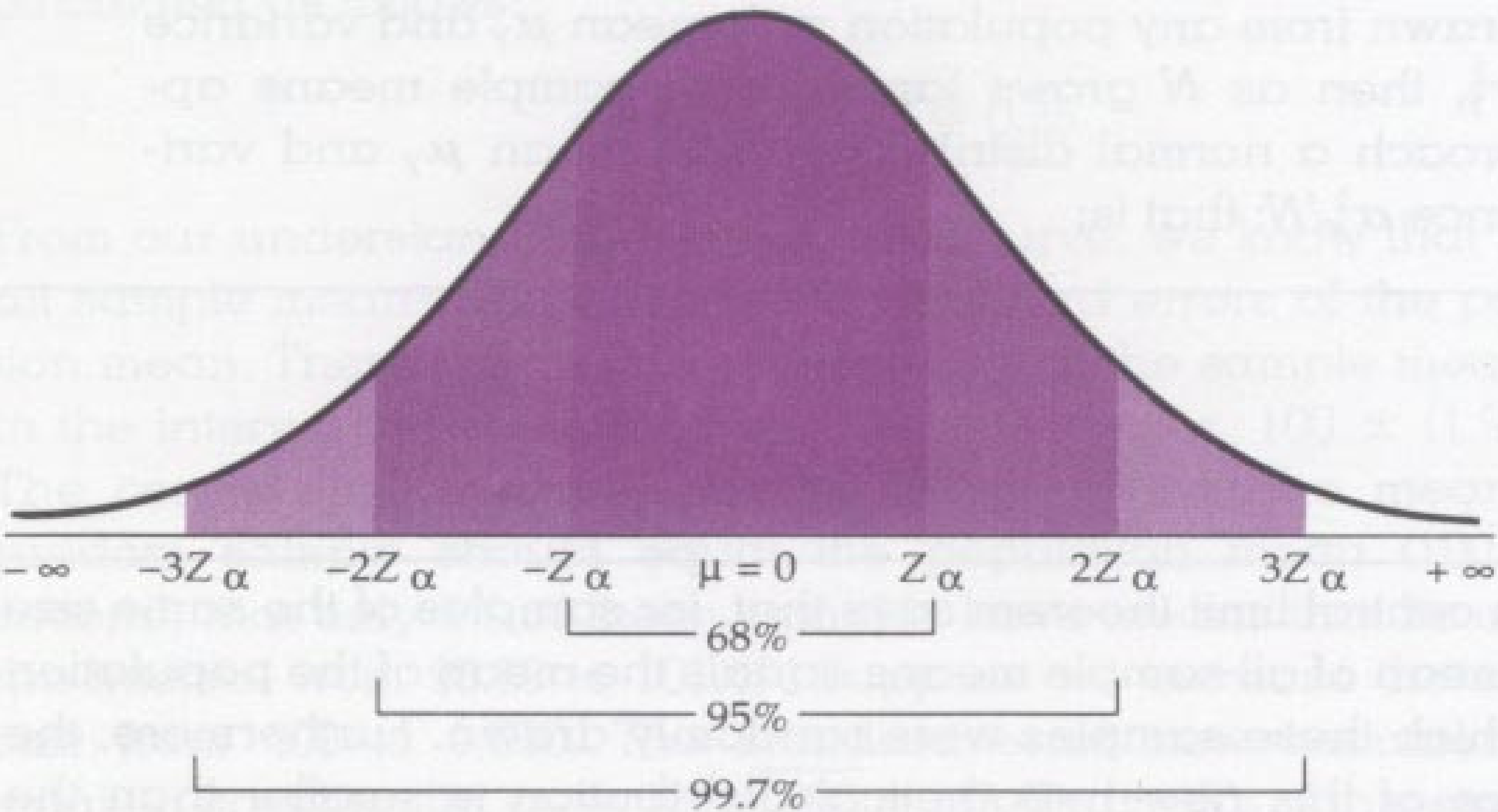


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

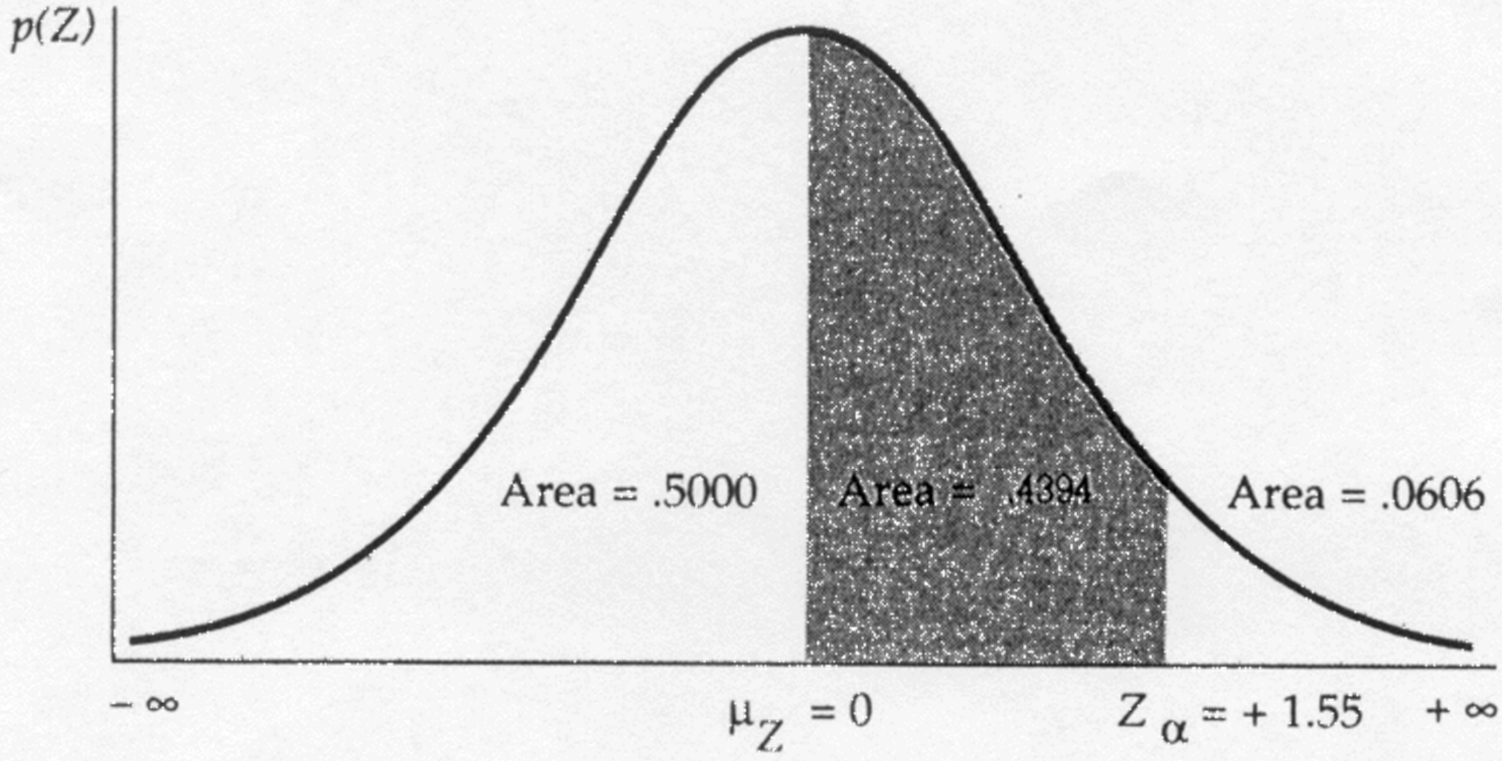
A normal distribution:

1. is **symmetrical** (both halves are *identical*);
2. is **asymptotic** (its *tails never touch* the underlying x-axis; the curve reaches to – ∞ and + ∞ and thus must be truncated);
3. has **fixed** and **known** *areas under the curve* (these fixed areas are marked off by units along the x-axis called **z-scores**; imposing truncation, the normal curve ends at + 3.00 z on the right and - 3.00 z on the left).

Areas Under the Normal Curve for Various Z Scores



Example of the Probability of Observing an Outcome in a Standar
Distribution



Conditional Distribution

The *conditional probability mass function* of Y given X is:

$$p(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p(x)}.$$

For continuous random variables, we can define the *conditional probability density function*:

Conditional probability: $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Multiplication rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.

$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

Rewriting the above equation yields:

$$f(x, y) = f(x) \cdot f(y|x).$$

The marginal density of Y can then be obtained from:

$$f(y) = \int_{-\infty}^{\infty} f(x) \cdot f(y|x) dx.$$

conditional probability which is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0.$$

Any other formula regarding conditional probability can be derived from the above formula. Specifically, if you have two random variables X and Y, you can write

$$P(X \in C|Y \in D) = \frac{P(X \in C, Y \in D)}{P(Y \in D)}, \text{ where } C, D \subset \mathbb{R}.$$

[redacted] the **conditional PMF**. Specifically, the conditional PMF of X given event A, is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i|A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}. \end{aligned}$$

Similarly, we define the **conditional CDF** of X given A as

$$F_{X|A}(x) = P(X \leq x|A).$$

Two discrete random variables X and Y are independent if

$$P_{XY}(x,y) = P_X(x)P_Y(y), \quad \text{for all } x, y.$$

Equivalently, X and Y are independent if

$$F_{XY}(x,y) = F_X(x)F_Y(y), \quad \text{for all } x, y.$$

For discrete random variables X and Y, the **conditional PMFs** of X given Y and vice versa are defined as

$$P_{X|Y}(x_i|y_j) = \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)},$$

$$P_{Y|X}(y_j|x_i) = \frac{P_{XY}(x_i, y_j)}{P_X(x_i)}$$

for any $x_i \in R_X$ and $y_j \in R_Y$.

So, if X and Y are independent, we have

$$\begin{aligned}
P_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\
&= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \\
&= \frac{P_X(x_i)P_Y(y_j)}{P_Y(y_j)} \\
&= P_X(x_i).
\end{aligned}$$

As we expect, for independent random variables, the conditional PMF is equal to the marginal PMF. In other words, knowing the value of Y does not provide any information about X.

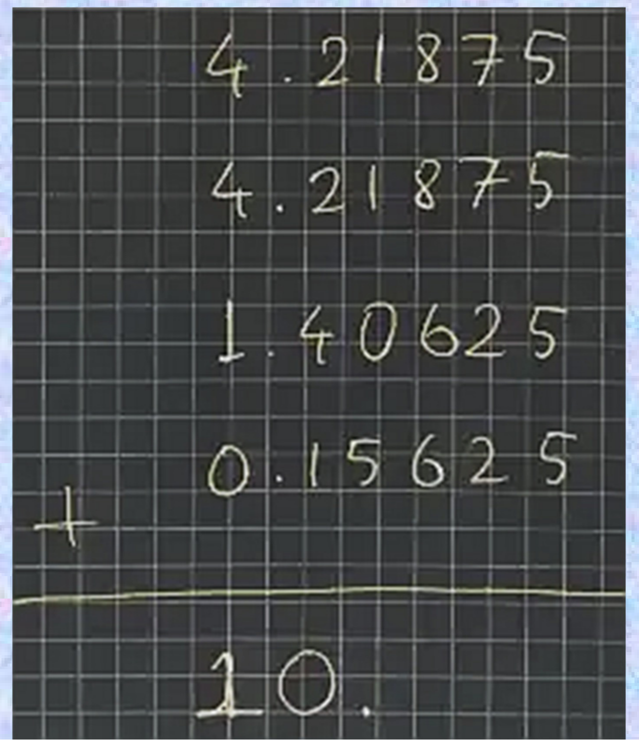
Expected Value of Random Variables

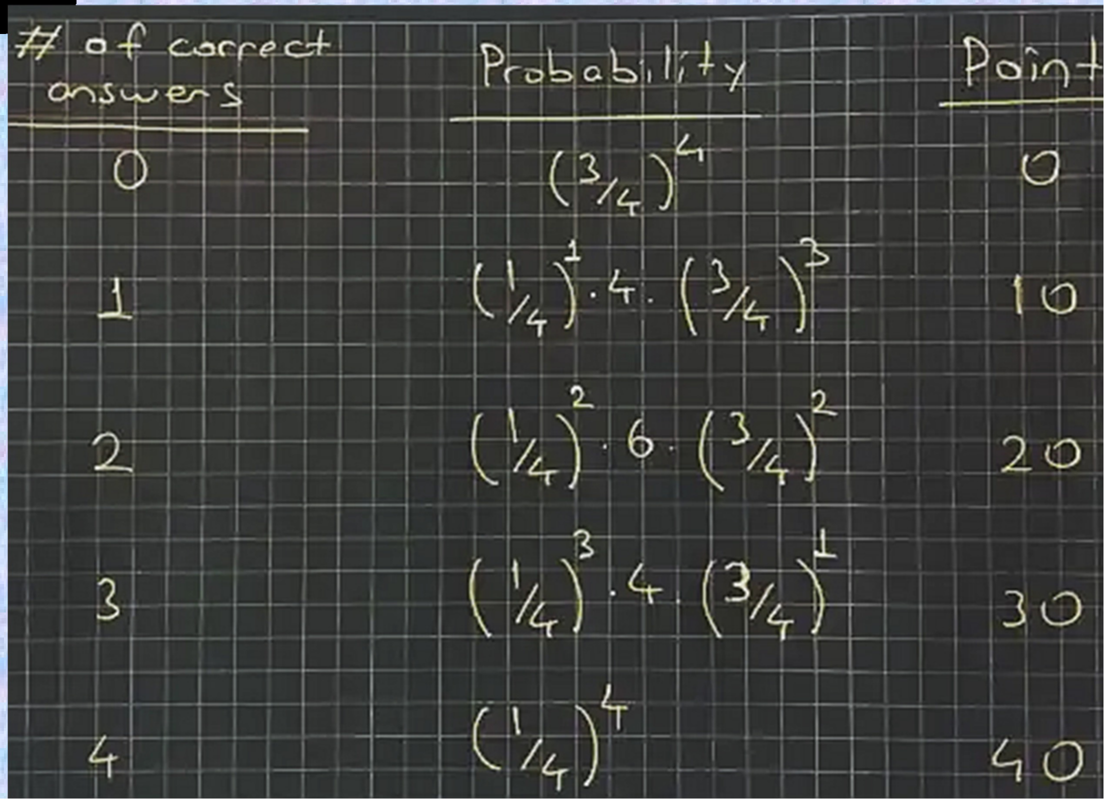
The expected value of a random variable is the weighted average of all possible values of the variable. The weight here means the probability of the random variable taking a specific value.

$$E[X] = \sum x_i p(x_i)$$

x_i = The values that X takes

$p(x_i)$ = The probability that X takes the value x_i



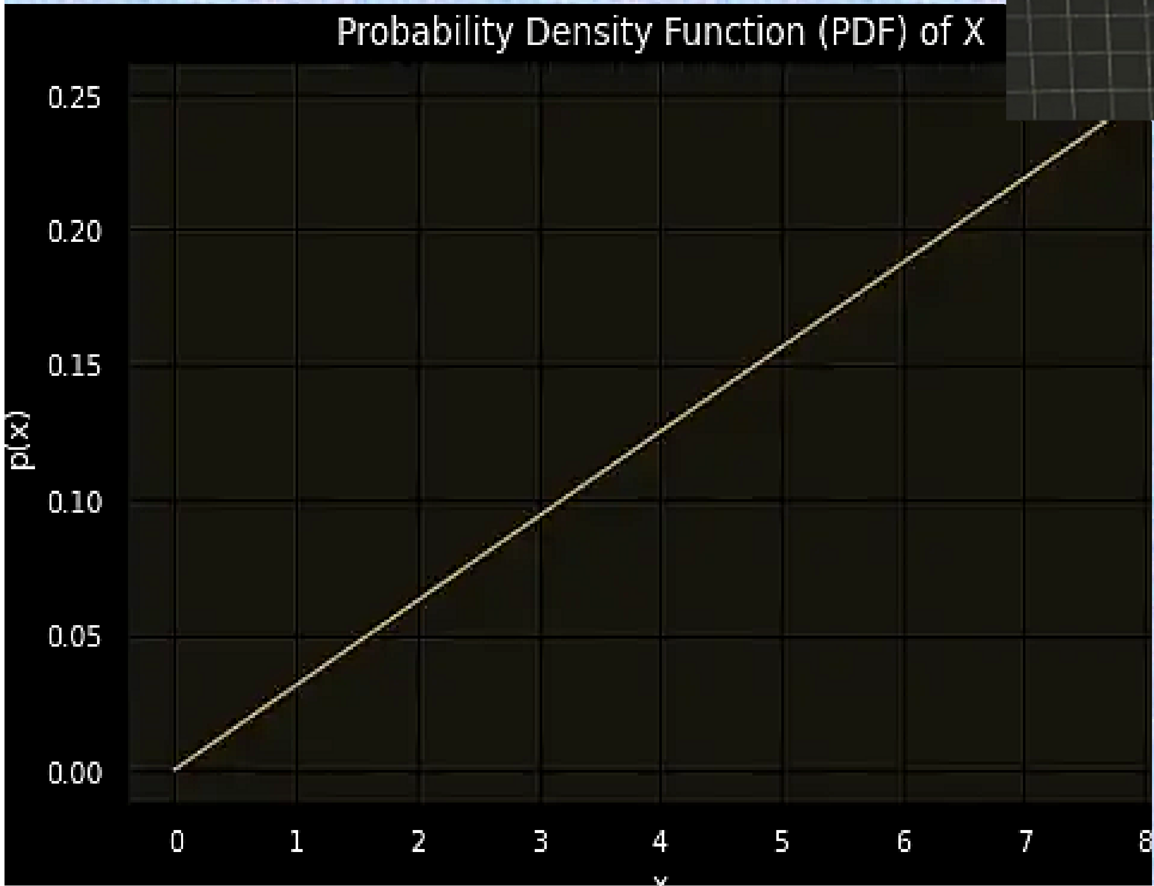


$$E[X] = \int_{xmin}^{xmax} x f(x) dx$$

f(x) is the PDF of X

$$E[x] = \int_5^{10} x\,0.2\,dx = 0.2\,\frac{x^2}{2}\Big|_5^{10}$$

$$= 0.1\,x^2\Big|_5^{10} = 0.1\,(100 - 25)$$



$$E[x] = \int_0^8 x\,0.03125\,x\,dx$$

$$= 0.03125\,\frac{x^3}{3}\Big|_0^8$$

$$= 0.03125\,\frac{(8)^3}{3}$$

$$= 5.33$$

Example Let X be a continuous random variable with support $R_X = [0, \infty)$ and probability density function

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \textit{if } x \in [0, \infty) \\ 0 & \textit{otherwise} \end{cases}$$

where $\lambda > 0$. Its expected value is

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{0}^{\infty} x \lambda \exp(-\lambda x) dx \end{aligned}$$

$$\begin{aligned} \mathrm{E}[Y] &= \sum_{x \in R_X} (a + bx) p_X(x) && \text{(by the transformation theorem)} \\ &= \sum_{x \in R_X} a p_X(x) + \sum_{x \in R_X} bx p_X(x) \\ &= a \sum_{x \in R_X} p_X(x) + b \sum_{x \in R_X} x p_X(x) \\ &= a + b \sum_{x \in R_X} x p_X(x) && \text{(because probabilities sum up to 1)} \\ &= a + b\mathrm{E}[X] && \text{(by the definition of } \mathrm{E}[X]\text{)} \end{aligned}$$

Expectation of $g(X)$

Let $g(X)$ be a function of X. We can imagine a long-term average of $g(X)$ just as we can imagine a long-term average of X. This average is written as $\mathbb{E}(g(X))$. Imagine observing X many times (N times) to give results $x_1, x_2, \ldots, x_N$. Apply the function g to each of these observations, to give $g(x_1), \ldots, g(x_N)$. The mean of $g(x_1), g(x_2), \ldots, g(x_N)$ approaches $\mathbb{E}(g(X))$ as the number of observations N tends to infinity.

Definition: Let X be a continuous random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

Definition: Let X be a discrete random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \sum_x g(x) f_X(x) = \sum_x g(x) \mathbb{P}(X = x).$$

Let X and Y be independent random variables, and g, h be functions. Then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$
$$\mathbb{E}\Big(g(X)h(Y)\Big) = \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big).$$

Probability as a conditional expectation

Define the indicator random variable: $I_A = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$

Then $\mathbb{E}(I_A) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

$$\mathbb{P}(A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big) = \mathbb{E}_Y\Big(\mathbb{P}(A \,|\, Y)\Big)$$

Law of Total Probability:

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A | Y = y_j) P_Y(y_j), \quad \text{for any set } A.$$

Law of Total Expectation:

1. If $B_1, B_2, B_3, ...$ is a partition of the sample space S,

$$EX = \sum_i E[X|B_i]P(B_i) \tag{5.3}$$

2. For a random variable X and a discrete random variable Y,

$$EX = \sum_{y_j \in R_Y} E[X|Y = y_j]P_Y(y_j) \tag{5.4}$$

Conditional Distribution and Conditional Expectation

The *conditional probability mass function* of Y given X is:

Conditional probability: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Multiplication rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.

$$p(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p(x)}.$$

For continuous random variables, we can define the *conditional probability density function*:

$$f(y|x) = \frac{f(x, y)}{f(x)}.$$

The *conditional expectation* of a random variable *Y* is the expected value of *Y* given [*X*=*x*], and is denoted: *E*[*Y*|*X*=*x*] or *E*[*Y*|*x*]. If the conditional probability density function is known, then the conditional expectation can be found using:

$$E[Y|X = x] = \begin{cases} \int_{-\infty}^{\infty} y \cdot f(y|x)dy & \text{if } Y \text{ is continuous} \\ \sum_y y \cdot p(y|x) & \text{if } Y \text{ is discrete} \end{cases} \quad (38)$$

To obtain the unconditional expectation of *Y*, we can take the expectation of *E*[*Y*|*X*]. The result is the *theorem of total expectation*:

$$E[Y] = \begin{cases} \int_{-\infty}^{\infty} E[Y|X = x] f(x)dx & \text{if } X \text{ is continuous} \\ \sum_x E[Y|X = x]p(x) & \text{if } X \text{ is discrete.} \end{cases} \quad (39)$$

Conditional Expectation of X:

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i),$$

$$E[X|Y = y_j] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y_j)$$

Iterated Expectations:

Let us look again at the law of total probability for expectation. Assuming $g(Y) = E[X|Y]$, we have

$$\begin{aligned} E[X] &= \sum_{y_j \in R_Y} E[X|Y = y_j] P_Y(y_j) \\ &= \sum_{y_j \in R_Y} g(y_j) P_Y(y_j) \\ &= E[g(Y)] \\ &= E[E[X|Y]]. \end{aligned}$$

[redacted]

Theorem 1 *Let X, Y, Z be random variables, $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist, we have*

(i) $E[a|Y] = a$

(ii) $E[aX + bZ|Y] = aE[X|Y] + bE[Z|Y]$

(iii) $E[X|Y] \geq 0$ *if* $X \geq 0$.

(iv) $E[X|Y] = E[X]$ *if* X *and* Y *are independent.*

(v) $E[E[X|Y]] = E[X]$

(vi) $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *In particular,* $|E[g(Y)|Y] = g(Y)$.

(vii) $E[X|Y, g(Y)] = E[X|Y]$

(viii) $E[E[X|Y, Z]|Y] = E[X|Y]$

Theorem 2 *For any function* $h : \mathbb{R} \to \mathbb{R}$,

$$E[(X - E[X|Y])^2] \leq E[(X - h(Y))^2]$$

and we have equality if and only if $h(Y) = E[X|Y]$.

This follows immediately from the law of total expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} = \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y).$$

Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables X and Y, we have:

i) $\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$ *Law of Total Expectation.*

Note that we can pick *any* *r.v.* Y, *to make the expectation as easy as we can.*

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \mid Y)\Big)$ *for any function* g.

we can give a proof of (1) in the special case where (X, Y, Z) are jointly continuous with a pdf $f(x, y, z)$:

$$E[X \mid Y = y, Z = z] = \frac{\int x \cdot f(x, y, z)\, dx}{\int f(x, y, z)\, dx},$$

$$\Downarrow$$

$$\begin{aligned} E[E[X \mid Y, Z = z] \mid Z = z] &= \iint \frac{\int x \cdot f(x, y, z)\, dx}{\int f(x, y, z)\, dx} \cdot f(x, y, z)\, dx\, dy \\ &= \int \left(\int x \cdot f(x, y, z)\, dx \right) \frac{\int f(x, y, z)\, dx}{\int f(x, y, z)\, dx}\, dy \\ &= \iint x \cdot f(x, y)\, dx\, dy \\ &= E[X \mid Z = z] \end{aligned}$$

You can give a similar proof in the case where X, Y, Z are jointly discrete, with a joint probability mass function $f(x, y, z) = P(X = x, Y = y, Z = z)$, for (x, y, z) ranging over some countable support set. Basically, you do this by replacing $\int$ with $\sum$ in the proof above.

One thing you can say is that

$$E[E[X \mid Y, Z] \mid Z] = E[X \mid Z] \tag{1}$$

$E[E[X|Y;Z]|Y=y]$. $E[X|Y;Z]$ is a random variable. Given that $Y=y$, its possible values are $E[X|Y=y;Z=z]$ where z varies over the range of Z. Given that $Y=y$, the probability that $E[X|Y;Z]=E[X|Y=y;Z=z]$ is just $P(Z=z|Y=y)$. Hence,

$$
\begin{aligned}
E[E[X|Y;Z]|Y=y] &= \sum_z E[X|Y=y,Z=z]P(Z=z|Y=y) \\
&= \sum_z \sum_x x\,P(X=x|Y=y,Z=z)P(Z=z|Y=y) \\
&= \sum_{z,x} x\,\frac{P(X=x,Y=y,Z=z)}{P(Y=y,Z=z)}\frac{P(Z=z,Y=y)}{P(Y=y)} \\
&= \sum_{z,x} x\,\frac{P(X=x,Y=y,Z=z)}{P(Y=y)} \\
&= \sum_x x\,\frac{P(X=x,Y=y)}{P(Y=y)} \\
&= \sum_x x\,P(X=x|Y=y) \\
&= E[X|Y=y]
\end{aligned}
$$

This follows immediately from the law of total expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} = \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y).$$

Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables X and Y, we have:

i) $\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$ *Law of Total Expectation.*

Note that we can pick *__any__* *r.v. Y, to make the expectation as easy as we can.*

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \mid Y)\Big)$ *for any function g.*

iii) $\boxed{Var(X) = \mathbb{E}_Y\Big(Var(X \mid Y)\Big) + Var_Y\Big(\mathbb{E}(X \mid Y)\Big)}$

Law of Total Variance.

(i) is a special case of (ii), so we just need to prove (ii). Begin at RHS:

$$\text{RHS} = \mathbb{E}_Y\Big[\mathbb{E}(g(X)\,|\,Y)\Big] = \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y)\right]$$

$$= \sum_y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y=y)\right]\mathbb{P}(Y=y)$$

(iii) Wish to prove $\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X\,|\,Y)] + \text{Var}_Y[\mathbb{E}(X\,|\,Y)]$. Begin at RHS:

$$\mathbb{E}_Y[\text{Var}(X\,|\,Y)] + \text{Var}_Y[\mathbb{E}(X\,|\,Y)]$$

$$= \mathbb{E}_Y\left\{\mathbb{E}(X^2\,|\,Y) - (\mathbb{E}(X\,|\,Y))^2\right\} + \left\{\mathbb{E}_Y\left\{[\mathbb{E}(X\,|\,Y)]^2\right\} - \Big[\underbrace{\mathbb{E}_Y(\mathbb{E}(X\,|\,Y))}_{\mathbb{E}(X)\text{ by part (i)}}\Big]^2\right\}$$

$$= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2\,|\,Y)\}}_{\mathbb{E}(X^2)\text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X\,|\,Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X\,|\,Y)]^2\} - (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$= \text{Var}(X) = \text{LHS}. \qquad \square$$

Theorem 2.4: The Partition Theorem (Law of Total Probability)

Let $B_1, \ldots, B_m$ form a partition of Ω. Then for any event A,

$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$

Proof of partition formula

$$\begin{aligned}
\sum_i \mathbf{E}(X \mid A_i)\,\mathbf{P}(A_i) &= \sum_i \int_\Omega X(\omega)\,\mathbf{P}(d\omega \mid A_i)\cdot\mathbf{P}(A_i) \\
&= \sum_i \int_\Omega X(\omega)\,\mathbf{P}(d\omega \cap A_i) \\
&= \sum_i \int_\Omega X(\omega) I_{A_i}(\omega)\,\mathbf{P}(d\omega) \\
&= \sum_i \mathbf{E}(X I_{A_i}),
\end{aligned}$$

where I_{A_i} is the indicator function of the set A_i.

If the partition $\{A_i\}_{i=0}^{n}$ is finite, then, by linearity, the previous expression becomes

$$\mathbf{E}\left(\sum_{i=0}^{n} X I_{A_i}\right) = \mathbf{E}(X),$$

5.2. Expectation and Variance of Standard Normal Distribution. Assume $X \backsim \mathcal{N}(0,1)$. Then

$$\mathbf{E}X = \int_{-\infty}^{+\infty} xe^{-x^2/2}\,dx = 0,$$

because the function inside the integral is odd. We can also say that X is symmetric with respect to zero, so $\mathbf{E}X = 0$. Now,

$$\mathbf{E}X^2 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^2e^{-x^2/2}\,dx = 1.$$

Why is this? We know that

$$\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}.$$

Let $u = e^{-x^2/2}$, $v = x$. Integrate by parts: note that $uv = xe^{-x^2/2} = 0$ for $x = \pm\infty$. So

$$\int_{-\infty}^{+\infty} e^{-x^2/2}\,dx = \int_{-\infty}^{+\infty} u\,dv = uv\Big|_{x=-\infty}^{x=+\infty} - \int_{-\infty}^{+\infty} v\,du$$

$$= -\int_{-\infty}^{+\infty} x\,de^{-x^2/2} = -\int_{-\infty}^{+\infty} x\,(-x)e^{-x^2/2}\,dx = \int_{-\infty}^{+\infty} x^2e^{-x^2/2}\,dx.$$

This is equal to $\sqrt{2\pi}$, which proves $\mathbf{E}X^2 = 1$. So $\operatorname{Var} X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = 1$. This proves that

$$\boxed{X \sim \mathcal{N}(0,1) \;\Rightarrow\; \mathbf{E}X = 0,\ \operatorname{Var} X = 1}$$

Normal Density: $$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

Bivariate Normal Density:

$$p(x,y) = \frac{e^{-\frac{1}{2(1-\rho_{xy}^2)}[(\frac{x-\mu_x}{\sigma_x})^2 - \frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + (\frac{y-\mu_y}{\sigma_y})^2]}}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho_{xy}^2)}}$$

μ - Mean; σ - S.D.; ρ_{xy} - Correlation Coefficient

Visualize ρ as equivalent to the orientation of tilted asymmetric Gaussian filter.

For x as a discrete random variable, the expected value of x:

$$E(x) = \sum_{i=1}^{n} x_i P(x_i) = \mu_x$$

E(x) is also called the first moment of the distribution.
The kth moment is defined as:

$$E(x^k) = \sum_{i=1}^{n} x_i^k P(x_i)$$

$P(x_i)$ is the probability of $x = x_i$.

Covariance of x and y, is defined as: $$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$$

Covariance indicates how much x and y vary together. The value depends on how much each variable tends to deviate from its mean, and also depends on the degree of association between x and y.

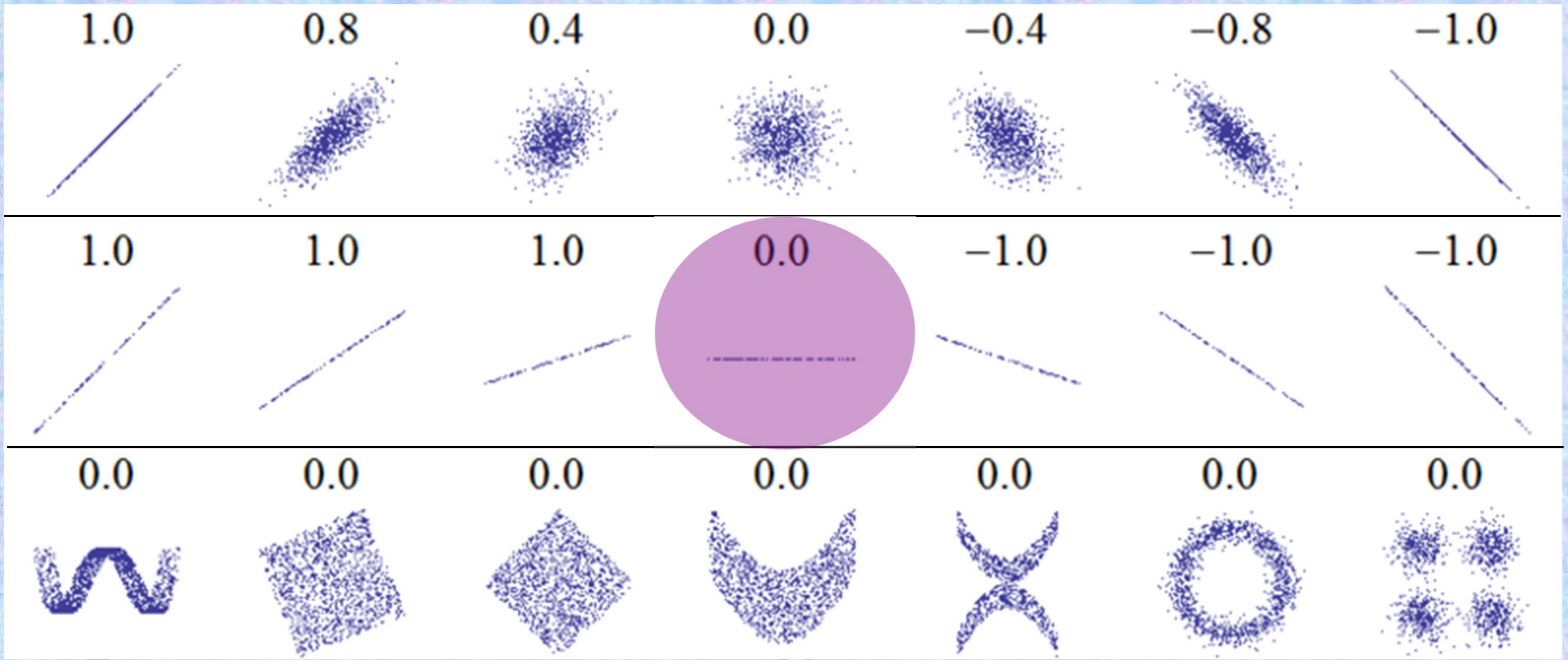
Correlation between x and y: $$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})]$$

Property of correlation coefficient: $$-1 \le \rho_{xy} \le 1$$

For Z = ax + by ;

$$E[(z-\mu_z)^2] = a^2\sigma_x^2 + 2ab\sigma_{xy} + b^2\sigma_y^2;$$

$$If \quad \sigma_{xy} = 0, \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2$$



$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})]$$

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\ \sqrt{E(Y^2) - E^2(Y)}}$$

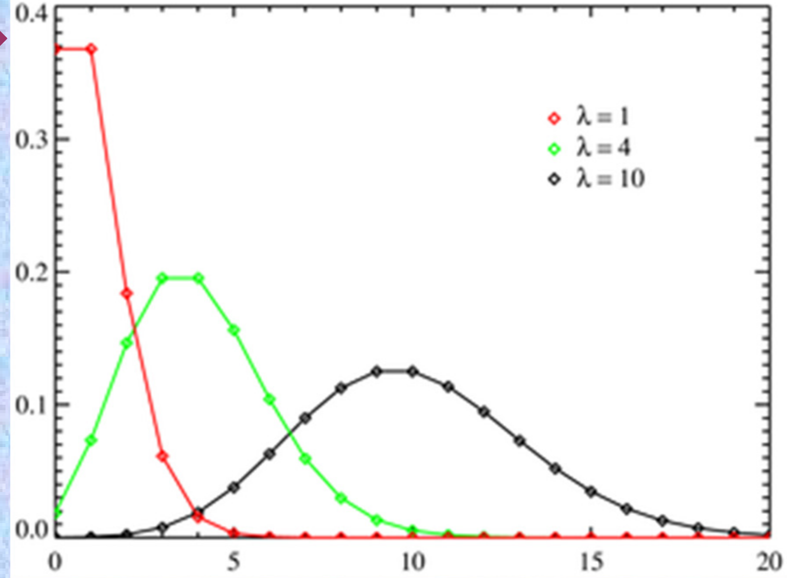
The correlation coefficient can also be viewed as the cosine of the angle between the two vectors ($\mathfrak{R}^D$) of samples drawn from the two random variables.

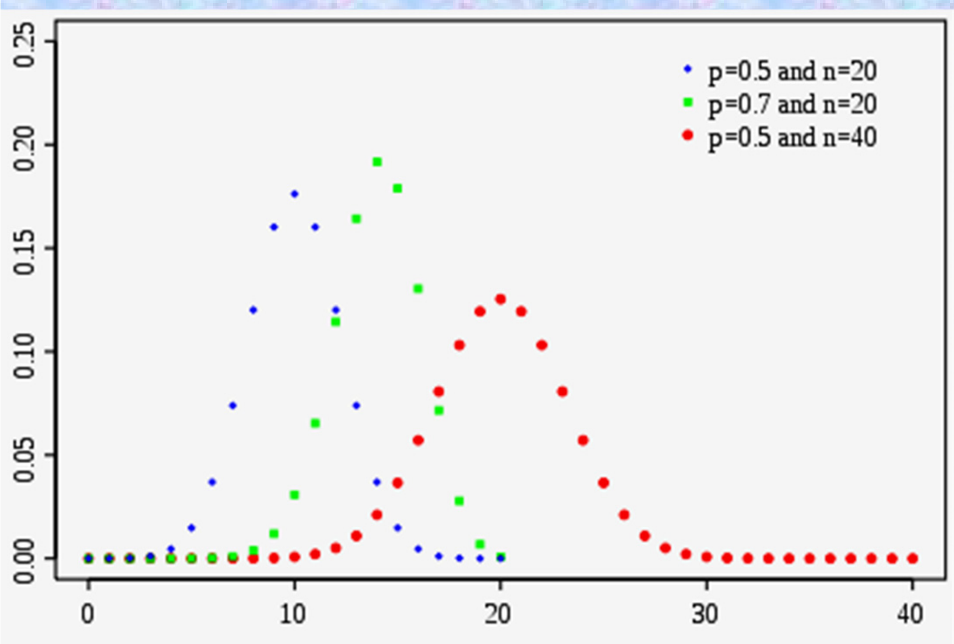
This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero.

Other PDFs:

Poisson →

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}; \quad \lambda > 0$$

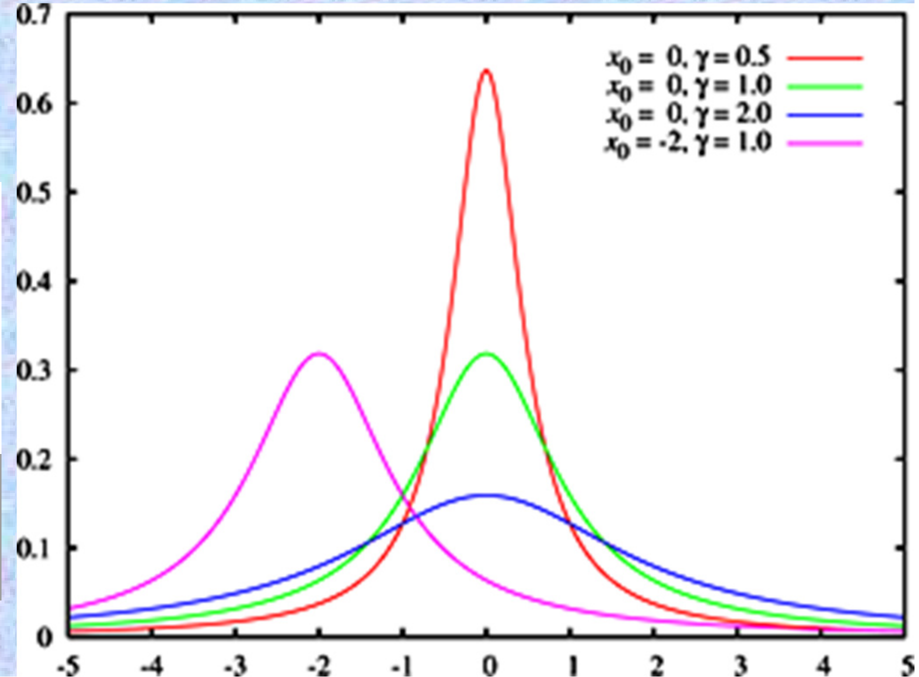




$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

← Binomial

Cauchy



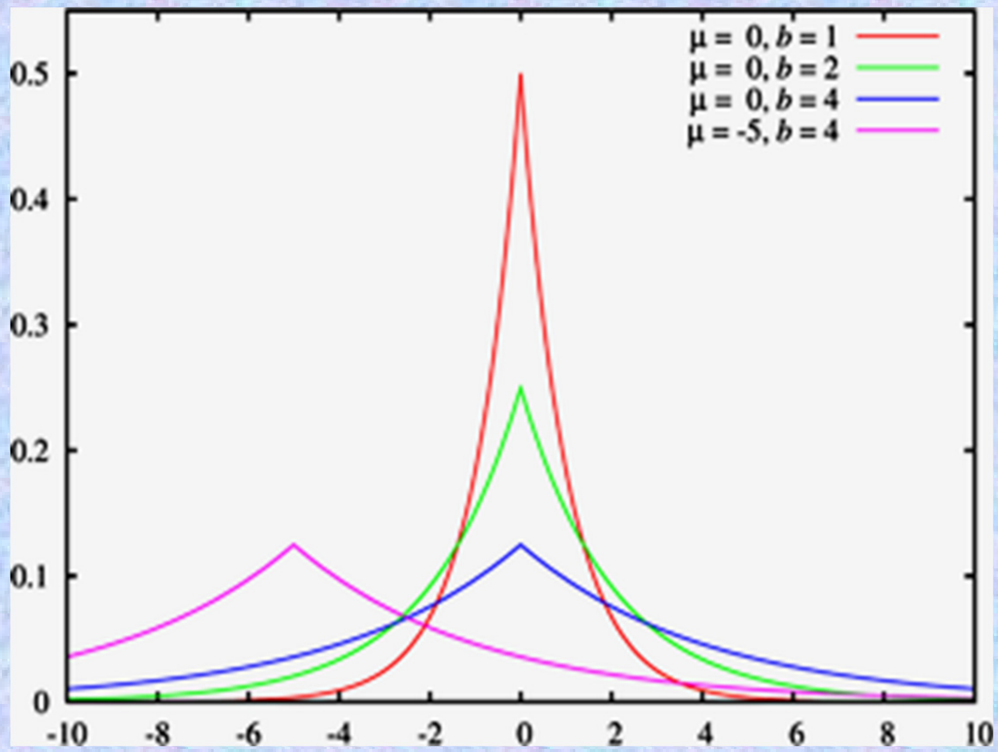
$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

$$= \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$

LAPLACE:

$$f(x|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

$$= \frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu\end{cases}$$



Read about:

- **Central Limit Theorem**
- **Uniform Distribution**
- **Geometric Distribution**
- **Quantile-Quantile (QQ) Plot**
- **Probability-Probability (P-P) Plot**

Double Exponential Density:

$$P(x) = \frac{1}{2b} e^{-\left| {x-a}/{b} \right|};$$

Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
Normal distribution	$f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform distribution (continuous)	$f(x\mid a,b)=\begin{cases}\frac{1}{b-a} & \text{for } a\le x\le b,\\ 0 & \text{for } x<a \text{ or } x>b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	$f(x\mid\lambda)=\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Poisson distribution	$f(x\mid\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ

The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = \operatorname{E}[X]$:

$$\operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)^2\right].$$

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X-\operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

In other words, the variance of X is equal to the mean of the square of X minus the square of the mean of X.

A formula for calculating the variance of an entire population of size N is:

$$\sigma^2 = \overline{(x^2)} - \bar{x}^2 = \frac{\sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2/N}{N}.$$

Using Bessel's correction to calculate an unbiased estimate of the population variance from a finite sample of n observations

$$s^2 = \left(\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2\right)\cdot\frac{n}{n-1}.$$

Discrete random variable [edit]

If the generator of random variable X is discrete with probability mass function $x_1 \mapsto p_1, x_2 \mapsto p_2, \ldots, x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

or equivalently,

$$\operatorname{Var}(X) = \left(\sum_{i=1}^{n} p_i x_i^2\right) - \mu^2,$$

where μ is the expected value. That is,

$$\mu = \sum_{i=1}^{n} p_i x_i.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

The variance of a collection of n equally likely values can be written as

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \mu^2,$$

where μ is the average value. That is,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If the random variable X has a probability density function $f(x)$, and $F(x)$ is the corresponding cumulative distribution function, then

$$\begin{aligned}\operatorname{Var}(X)=\sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx \\ &= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu \int_{\mathbb{R}} x f(x)\,dx + \mu^2 \int_{\mathbb{R}} f(x)\,dx \\ &= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \int_{\mathbb{R}} x\,dF(x) + \mu^2 \int_{\mathbb{R}} dF(x) \\ &= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu\cdot\mu + \mu^2\cdot 1 \\ &= \int_{\mathbb{R}} x^2\,dF(x) - \mu^2,\end{aligned}$$

or equivalently,

$$\operatorname{Var}(X)=\int_{\mathbb{R}} x^2 f(x)\,dx - \mu^2,$$

where μ is the expected value of X given by

$$\mu = \int_{\mathbb{R}} x f(x)\,dx = \int_{\mathbb{R}} x\,dF(x).$$

Variance (σ^2) of a discrete random variable X is defined as

$$\sigma^2 = V(X) = E\left(X - E(X)\right)^2$$

$$\begin{aligned}
\operatorname{Var}(X) = \sigma^2 &= \int_{\mathbb{R}} (x-\mu)^2 f(x)\,dx \\
&= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\mu \int_{\mathbb{R}} x f(x)\,dx + \mu^2 \int_{\mathbb{R}} f(x)\,dx \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \int_{\mathbb{R}} x\,dF(x) + \mu^2 \int_{\mathbb{R}} dF(x) \\
&= \int_{\mathbb{R}} x^2\,dF(x) - 2\mu \cdot \mu + \mu^2 \cdot 1 \\
&= \int_{\mathbb{R}} x^2\,dF(x) - \mu^2,
\end{aligned}$$

$$\begin{aligned}
E(X-\mu)^2 &= E(X^2 - 2X\mu + \mu^2) \\
&= E(X^2) - 2E(X)\mu + E(\mu^2) \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2 \\
&= E(X^2) - E(X)^2
\end{aligned}$$

$$\begin{aligned}
V[X] &= E[(X - E[X])^2] \\
&= E[X^2 - 2XE[X] + E[X]^2] \\
&= E[X^2] - 2E[XE[X]] + E[X]^2 \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - E[X]^2
\end{aligned}$$

$$\sum_x x^2 p(x) - \mu^2,$$

$$2E(X)\cdot X + E^2(X)\}$$

$$= E\left(X^2\right) - 2\cdot E(X)E(X) + E^2(X) = E\left(X^2\right) - E^2(X).$$

Definition [edit]

Throughout this article, boldfaced unsubscripted $\mathbf{X}$ and $\mathbf{Y}$ are used to refer to random vectors, and unboldfaced subscripted X_i and Y_i are used to refer to scalar random variables.

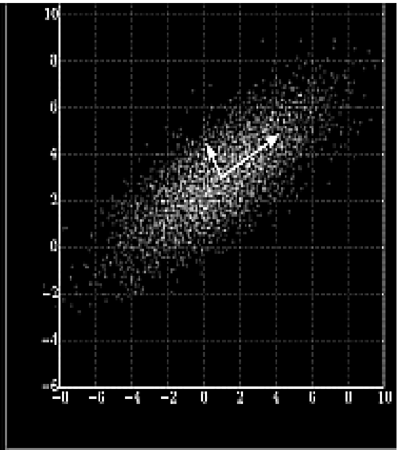
If the entries in the column vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$$

are random variables, each with finite variance and expected value, then the covariance matrix $\operatorname{K}_{\mathbf{X}\mathbf{X}}$ is the matrix whose (i, j) entry is the covariance[1]: p. 177

$$\operatorname{K}_{X_i X_j} = \operatorname{cov}[X_i, X_j] = \operatorname{E}[(X_i - \operatorname{E}[X_i])(X_j - \operatorname{E}[X_j])]$$

where the operator E denotes the expected value (mean) of its argument.



Sample points from a bivariate Gaussian distribution with a standard deviation of 3 in roughly the lower left-upper right direction and of 1 in the orthogonal direction. Because the *x* and *y* components co-vary, the variances of x and y do not fully describe the distribution. A 2×2 covariance matrix is needed; the directions of the arrows correspond to the eigenvectors of this covariance matrix and their lengths to the square roots of the eigenvalues.

Conflicting nomenclatures and notations [edit]

Nomenclatures differ. Some statisticians, following the probabilist William Feller in his two-volume book *An Introduction to Probability Theory and Its Applications*,[2] call the matrix $\operatorname{K}_{\mathbf{X}\mathbf{X}}$ the **variance** of the random vector $\mathbf{X}$, because it is the natural generalization to higher dimensions of the 1-dimensional variance. Others call it the **covariance matrix**, because it is the matrix of covariances between the scalar components of the vector $\mathbf{X}$.

$$\operatorname{var}(\mathbf{X}) = \operatorname{cov}(\mathbf{X}, \mathbf{X}) = \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{X} - \operatorname{E}[\mathbf{X}])^{\mathrm{T}}\right].$$

Both forms are quite standard, and there is no ambiguity between them. The matrix $\operatorname{K}_{\mathbf{X}\mathbf{X}}$ is also often called the *variance-covariance matrix*, since the diagonal terms are in fact variances.

By comparison, the notation for the cross-covariance matrix *between* two vectors is

$$\operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \operatorname{K}_{\mathbf{X}\mathbf{Y}} = \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{Y} - \operatorname{E}[\mathbf{Y}])^{\mathrm{T}}\right].$$

Basic properties

For $\mathrm{K}_{\mathbf{XX}} = \operatorname{var}(\mathbf{X}) = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}}\right]$ and $\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$, where $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ is a n-dimensional random variable, the following basic properties apply:[4]

1. $\mathrm{K}_{\mathbf{XX}} = \mathrm{E}(\mathbf{XX}^{\mathrm{T}}) - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^{\mathrm{T}}$
2. $\mathrm{K}_{\mathbf{XX}}$ is positive-semidefinite, i.e. $\mathbf{a}^T \mathrm{K}_{\mathbf{XX}}\, \mathbf{a} \geq 0 \quad \text{for all } \mathbf{a} \in \mathbb{R}^n$
3. $\mathrm{K}_{\mathbf{XX}}$ is symmetric, i.e. $\mathrm{K}_{\mathbf{XX}}^{\mathrm{T}} = \mathrm{K}_{\mathbf{XX}}$
4. For any constant (i.e. non-random) $m \times n$ matrix $\mathbf{A}$ and constant $m \times 1$ vector $\mathbf{a}$, one has $\operatorname{var}(\mathbf{AX} + \mathbf{a}) = \mathbf{A}\operatorname{var}(\mathbf{X})\mathbf{A}^{\mathrm{T}}$
5. If $\mathbf{Y}$ is another random vector with the same dimension as $\mathbf{X}$, then $\operatorname{var}(\mathbf{X} + \mathbf{Y}) = \operatorname{var}(\mathbf{X}) + \operatorname{cov}(\mathbf{X}, \mathbf{Y}) + \operatorname{cov}(\mathbf{Y}, \mathbf{X}) + \operatorname{var}(\mathbf{Y})$ where $\operatorname{cov}(\mathbf{X}, \mathbf{Y})$ is the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$.

For random vectors $\mathbf{X}$ and $\mathbf{Y}$, each containing random elements whose expected value and variance exist, the **cross-covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is defined by[1]: p.336

$$\mathrm{K}_{\mathbf{XY}} = \operatorname{cov}(\mathbf{X}, \mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^{\mathrm{T}}] \quad \textbf{(Eq.1)}$$

where $\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$ and $\mu_{\mathbf{Y}} = \mathrm{E}[\mathbf{Y}]$ are vectors containing the expected values of $\mathbf{X}$ and $\mathbf{Y}$. The vectors $\mathbf{X}$ and $\mathbf{Y}$ need not have the same dimension, and either might be a scalar value.

The cross-covariance matrix is the matrix whose (i, j) entry is the covariance

$$\mathrm{K}_{X_i Y_j} = \operatorname{cov}[X_i, Y_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(Y_j - \mathrm{E}[Y_j])]$$

For the cross-covariance matrix, the following basic properties apply:[2]

1. $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = \operatorname{E}[\mathbf{X}\mathbf{Y}^{\mathrm T}] - \mu_{\mathbf{X}}\mu_{\mathbf{Y}}{}^{\mathrm T}$
2. $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = \operatorname{cov}(\mathbf{Y},\mathbf{X})^{\mathrm T}$
3. $\operatorname{cov}(\mathbf{X}_1 + \mathbf{X}_2,\mathbf{Y}) = \operatorname{cov}(\mathbf{X}_1,\mathbf{Y}) + \operatorname{cov}(\mathbf{X}_2,\mathbf{Y})$
4. $\operatorname{cov}(A\mathbf{X} + \mathbf{a}, B^{\mathrm T}\mathbf{Y} + \mathbf{b}) = A\,\operatorname{cov}(\mathbf{X},\mathbf{Y})\,B$
5. If $\mathbf{X}$ and $\mathbf{Y}$ are independent (or somewhat less restrictedly, if every random variable in $\mathbf{X}$ is uncorrelated with every random variable in $\mathbf{Y}$), then $\operatorname{cov}(\mathbf{X},\mathbf{Y}) = 0_{p\times q}$

where $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are random $p \times 1$ vectors, $\mathbf{Y}$ is a random $q \times 1$ vector, $\mathbf{a}$ is a $q \times 1$ vector, $\mathbf{b}$ is a $p \times 1$ vector, A and B are $q \times p$ matrices of constants, and $0_{p\times q}$ is a $p \times q$ matrix of zeroes.

Given a sample consisting of *n* independent observations $x_1,\ldots, x_n$ of a *p*-dimensional random vector $X \in \mathbf{R}^{p\times 1}$ (a *p*×1 column-vector), an unbiased estimator of the (*p*×*p*) covariance matrix

$$\Sigma = \operatorname{E}\left[\left(X - \operatorname{E}[X]\right)\left(X - \operatorname{E}[X]\right)^{\mathrm T}\right]$$

is the sample covariance matrix

$$\mathbf{Q} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^{\mathrm T},$$

where x_i is the *i*-th observation of the *p*-dimensional random vector, and the vector

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the sample mean. This is true regardless of the distribution of the random variable *X*, provided of course that the theoretical means and covariances exist. The reason

Which matrices are covariance matrices?

let $\mathbf{b}$ be a $(p \times 1)$ real-valued vector, then

$$\operatorname{var}(\mathbf{b}^{\mathrm{T}}\mathbf{X}) = \mathbf{b}^{\mathrm{T}} \operatorname{var}(\mathbf{X})\mathbf{b},$$

which must always be nonnegative, since it is the variance of a real-valued random variable, so a covariance matrix is always a positive-semidefinite matrix.

The above argument can be expanded as follows:

$$w^{\mathrm{T}} \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{X} - \operatorname{E}[\mathbf{X}])^{\mathrm{T}}\right] w = \operatorname{E}\left[w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{X} - \operatorname{E}[\mathbf{X}])^{\mathrm{T}} w\right]$$
$$= \operatorname{E}\left[\left(w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])\right)^2\right] \geq 0,$$

where the last inequality follows from the observation that $w^{\mathrm{T}}(\mathbf{X} - \operatorname{E}[\mathbf{X}])$ is a scalar.

Conversely, every symmetric positive semi-definite matrix is a covariance matrix. To see this, suppose M is a $p \times p$ symmetric positive-semidefinite matrix. From the finite-dimensional case of the spectral theorem, it follows that M has a nonnegative symmetric square root, which can be denoted by $\mathrm{M}^{1/2}$. Let $\mathbf{X}$ be any $p \times 1$ column vector-valued random variable whose covariance matrix is the $p \times p$ identity matrix. Then

$$\operatorname{var}(\mathbf{M}^{1/2}\mathbf{X}) = \mathbf{M}^{1/2} \operatorname{var}(\mathbf{X}) \mathbf{M}^{1/2} = \mathbf{M}.$$

$$= \operatorname{E}[b(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top} b^{\top}]$$

$$= b\operatorname{E}[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}]b^{\top}$$

$$= b\operatorname{Var}[X]b^{\top}$$

PROB. & STAT. Contd.

Sample mean is defined as: $\tilde{x} = \sum_{i=1}^{n} x_i P(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$ where, $P(x_i) = 1/n$.

Sample Variance is: $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})^2$

Higher order moments may also be computed: $E(x_i - \tilde{x})^3;\ E(x_i - \tilde{x})^4$

Covariance of a bivariate distribution:

$$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] = \frac{1}{n}\sum_{i=1}^{n}(x-\tilde{x})(y-\tilde{y})$$

Second, third,… moments of the distribution p(x) are the expected values of: x^2, x^3,…

The k^{th} central moment is defined as:

$$E[(x-\mu_x)^k] = \sum_{i=1}^{n} (x-\mu_x)^k P(x_i)$$

Thus, the second central moment (also called Variance) of a random variable x is defined as:

$$\sigma_x^2 = E[\{x - E(x)\}^2] = E[(x-\mu_x)^2]$$

S.D. of x is σ_x.

$$\sigma_x^2 = E[\{x - E(x)\}^2] = E[(x-\mu_x)^2]$$

$$= E(x^2) - 2\mu_x^2 + \mu_x^2 = E(x^2) - \mu_x^2$$

Thus

$$E(x^2) = \sigma^2 + \mu^2$$

If z is a new variable: z= ax + by; Then E(z) = E(ax + by)=aE(x) + bE(y).

$$M_X(t) = \mathrm{E}[e^{tX}] = \int_{x=-\infty}^{\infty} e^{tx} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx =$$

$$\int_{x=-\infty}^{\infty} \frac{e^{-(x^2-2tx+t^2)/2} e^{t^2/2}}{\sqrt{2\pi}}\, dx = e^{t^2/2} \int_{x=-\infty}^{\infty} \frac{e^{-(x-t)^2/2}}{\sqrt{2\pi}}\, dx.$$

But this last integrand is a normal density with mean t and variance 1, thus integrates to 1. Hence

$$M_X(t) = e^{t^2/2}.$$

Now we recall that

$$\mathrm{E}[X^k] = \left[\frac{d^k M_X(t)}{dt^k}\right]_{t=0},$$

so let's calculate successive derivatives:

$$\begin{aligned}
M_X'(t) &= te^{t^2/2} \\
M_X''(t) &= e^{t^2/2} + t^2 e^{t^2/2} = (1+t^2)e^{t^2/2} \\
M_X'''(t) &= 2te^{t^2/2} + (1+t^2)te^{t^2/2} = (3t+t^3)e^{t^2/2} \\
M_X^{(4)}(t) &= (3+3t^2)e^{t^2/2} + (3t^2+t^4)e^{t^2/2} = (3+6t^2+t^4)e^{t^2/2},
\end{aligned}$$

and it is fairly easy to continue this. Now simply evaluate all of these at $t = 0$ to get

$$\begin{aligned}
\mathrm{E}[X] &= 0 \\
\mathrm{E}[X^2] &= 1 \\
\mathrm{E}[X^3] &= 0 \\
\mathrm{E}[X^4] &= 3.
\end{aligned}$$

MAXIMUM LIKELIHOOD ESTIMATE (MLE)

The ML estimate (MLE) of a parameter is that value which, when substituted into the probability distribution (or density), produces that distribution for which the probability of obtaining the entire observed set of samples is maximized.

Problem: **Find the maximum likelihood estimate for μ in a normal distribution.**

Normal Density: $$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

Assuming all random samples to be independent:

$$p(x_1,,,,x_n) = p(x_1).....p(x_n) = \prod_{i=1}^{n} p(x_i)$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\frac{x-\mu}{\sigma})^2]$$

Taking derivative (w.r.t. μ) of the LOG of the above:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu).2 = \frac{1}{\sigma^2}[\sum_{i=1}^{n} x_i - n\mu]$$

Setting this term = 0, we get:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \tilde{x}$$

Also read about MAP estimate – Baye's is an example.

Sampling Distributions

http://grid.cs.gsu.edu/~skarmakar/math1070_slides.html

Sampling Distribution Introduction

- In real life calculating parameters of populations is prohibitive because populations are very large.
- Rather than investigating the whole population, we take a sample, calculate a **statistic** related to the **parameter** of interest, and make an inference.
- The **sampling distribution** of the **statistic** is the tool that tells us how close is the statistic to the parameter.

What are the main types of <u>sampling</u> and how is each done?

Simple Random Sampling: A simple random sample (**SRS**) of size *n* is produced by a scheme which ensures that each subgroup of the population of size *n* has an equal probability of being chosen as the sample.

Stratified Random Sampling: Divide the population into "strata". There can be any number of these. Then choose a simple random sample from each stratum. Combine those into the overall sample. That is a stratified random sample. (Example: Church A has 600 women and 400 women as members. One way to get a stratified random sample of size 30 is to take a SRS of 18 women from the 600 women and another SRS of 12 men from the 400 men.)

Multi-Stage Sampling: Sometimes the population is too large and scattered for it to be practical to make a list of the entire population from which to draw a SRS. For instance, when the a polling organization samples US voters, they do not do a SRS. Since voter lists are compiled by counties, they might first do a sample of the counties and then sample within the selected counties. This illustrates two stages.

< SRC: WIKI *>*

In statistics, a **simple random sample** is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals. This process and technique is known as simple random sampling, and should not be confused with systematic random sampling. A simple random sample is an unbiased surveying technique.

Systematic sampling (Sys-S) is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equi-probability method. In this approach, progression through the list is treated circularly, with a return to the top once the end of the list is passed. The sampling starts by selecting an element from the list at random and then every k-th element in the frame is selected, where k, the sampling interval (sometimes known as the *skip*): this is calculated as: $k = N/n$
where n is the sample size, and N is the population size.

Systematic sampling (Sys-S) Example: Suppose a supermarket wants to study buying habits of their customers, then using systematic sampling they can choose every 10th or 15th customer entering the supermarket and conduct the study on this sample.

This is random sampling with a system. From the sampling frame, a starting point is chosen at random, and choices thereafter are at regular intervals. For example, suppose you want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

Sampling With Replacement and Sampling Without Replacement

Consider a population of potato sacks, each of which has either 12, 13, 14, 15, 16, 17, or 18 potatoes, and all the values are equally likely. Suppose that, in this population, there is exactly one sack with each number. So the whole population has seven sacks.

Sampling with replacement:

If I sample two with replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I replace it. Then I pick another. Every one of them still has 1/7 probability of being chosen. And there are exactly 49 different possibilities here.

Sampling without replacement:

If I sample two without replacement, then I first pick one (say 14). I had a 1/7 probability of choosing that one. Then I pick another. At this point, there are only six possibilities: 12, 13, 15, 16, 17, and 18. So there are only 42 different possibilities here (again assuming that we distinguish between the first and the second.)

Sampling distribution

- The sampling distribution of a statistic (not parameter) is the **distribution of values taken by the statistic** (not parameter) **in <u>all</u> possible samples** of the **<u>same size</u>** from the **same population**.

Sample Statistics as Estimators of Population Parameters

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.

A **population parameter** is a numerical measure of a summary characteristic of a population.

- An **estimator** of a population parameter is a sample statistic used to estimate or predict the population parameter.
- An **estimate** of a parameter is a *particular* numerical value of a sample statistic obtained through sampling.
- A **point estimate** is a single value used as an estimate of a population parameter.

Estimators

- The sample mean, $\overline{X}$, is the most common estimator of the population mean, μ.
- The sample variance, s^2, is the most common estimator of the population variance, σ^2.
- The sample standard deviation, s, is the most common estimator of the population standard deviation, σ.
- The sample proportion, $\hat{p}$, is the most common estimator of the population proportion, p.

Sampling Distribution of $\overline{X}$

- The **sampling distribution of $\overline{X}$** is the probability distribution of all possible values the random variable $\overline{X}$ may assume when a sample of size *n* is taken from a specified population.

Sampling Distribution of the Mean

- An example
  - A die is thrown infinitely many times. Let X represent the number of spots showing on any throw.
  - The probability distribution of X is

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

E(X) = 1(1/6) + 2(1/6) + 3(1/6)+ ……………………= 3.5

V(X) = $(1-3.5)^2(1/6)$ + $(2-3.5)^2(1/6)$ + ………….. …= 2.92

Throwing a dice twice – sampling distribution of sample mean

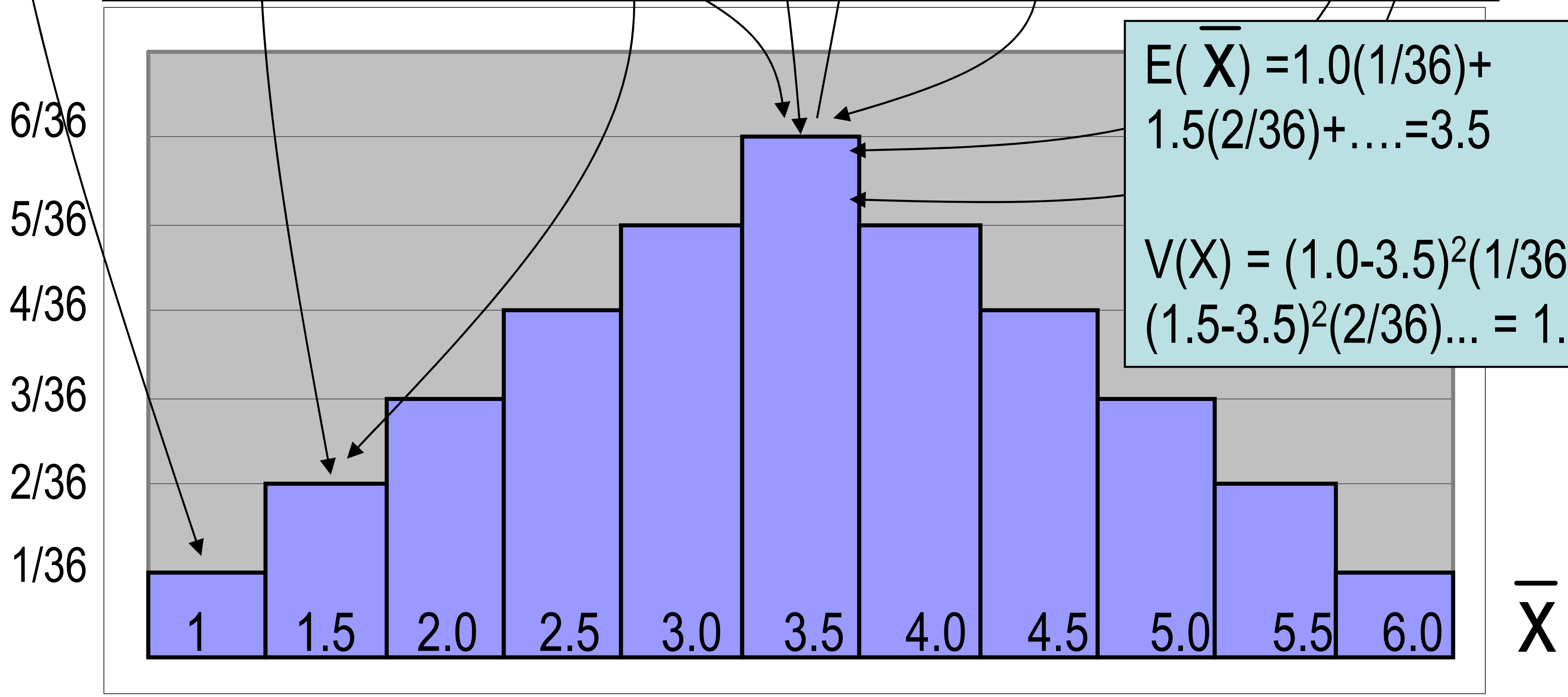
- Suppose we want to estimate μ from the mean $\bar{x}$ of a sample of size n = 2.
- What is the distribution of $\bar{x}$?

Throwing a die twice – sample mean

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

$$Note: \ \mu_{\bar{x}} = \mu_x \quad and \quad \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{2}$$



$E(\bar{X}) = 1.0(1/36) + 1.5(2/36) + \ldots = 3.5$

$V(X) = (1.0-3.5)^2(1/36) + (1.5-3.5)^2(2/36) \ldots = 1.46$

Sampling Distribution of the Mean

$n = 5$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .5833 \; (= \frac{\sigma^2_x}{5})$

6

$n = 10$

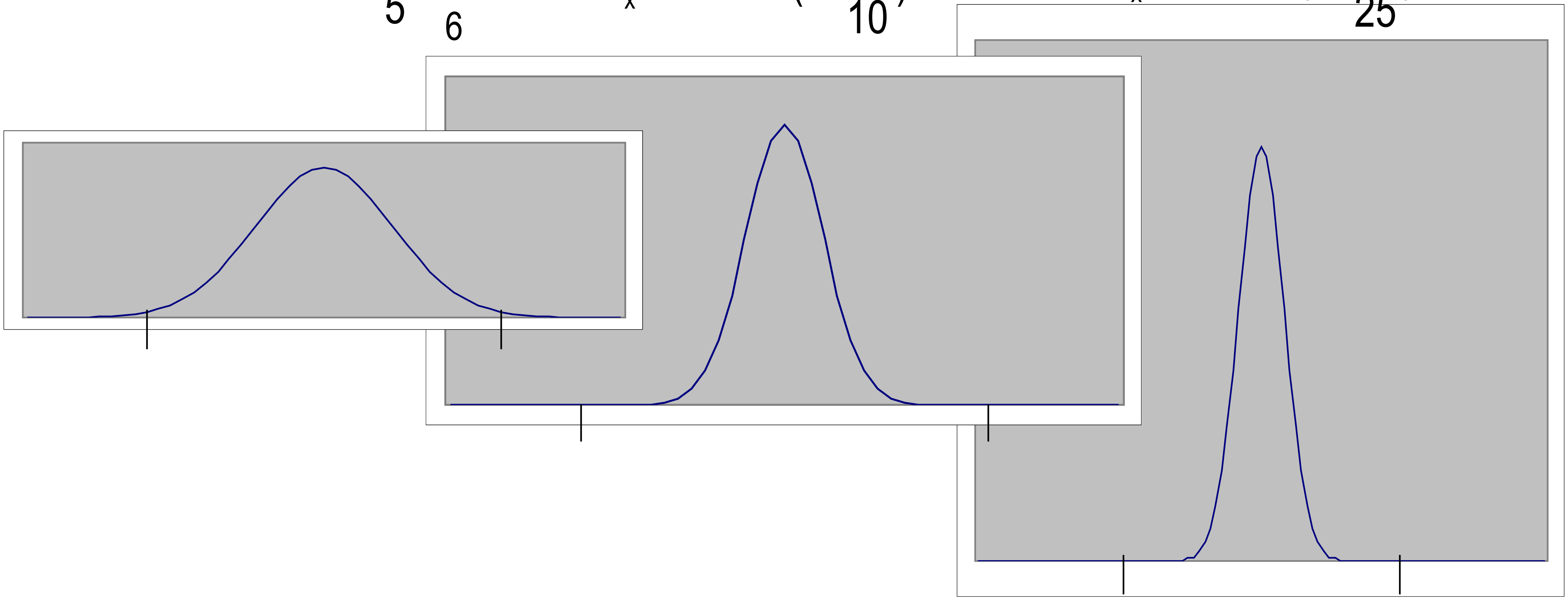
$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .2917 \; (= \frac{\sigma^2_x}{10})$

$n = 25$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .1167 \; (= \frac{\sigma^2_x}{25})$



Sampling Distribution of the Mean

$n = 5$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .5833 \; (= \frac{\sigma^2_x}{5})$

$n = 10$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .2917 \; (= \frac{\sigma^2_x}{10})$

$n = 25$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .1167 \; (= \frac{\sigma^2_x}{25})$

Notice that $\sigma^2_{\bar{x}}$ is smaller than σ^2_x. The larger the sample size the smaller $\sigma^2_{\bar{x}}$. Therefore, $\bar{X}$ tends to fall closer to μ, as the sample size increases.

Relationships between Population Parameters and the Sampling Distribution of the Sample Mean

The **expected value of the sample mean** is equal to the population mean:

$$E(\overline{X}) = \mu_{\overline{X}} = \mu_X$$

The **variance of the sample mean** is equal to the population variance divided by the sample size:

$$V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$$

The **standard deviation of the sample mean, known as the standard error of the mean,** is equal to the population standard deviation divided by the square root of the sample size:

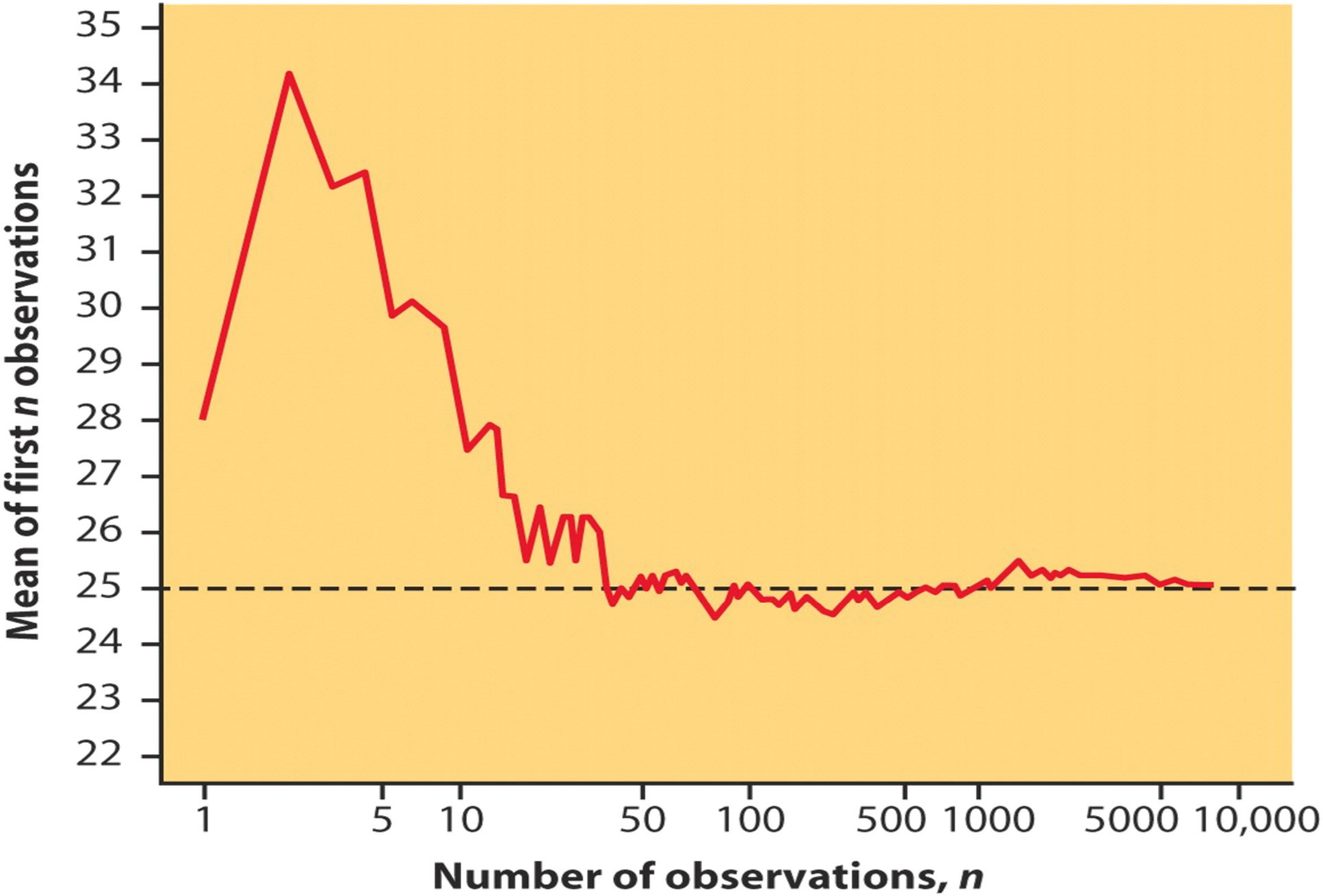
$$\text{s.e.} = SD(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Law of Large Number

LAW OF LARGE NUMBERS

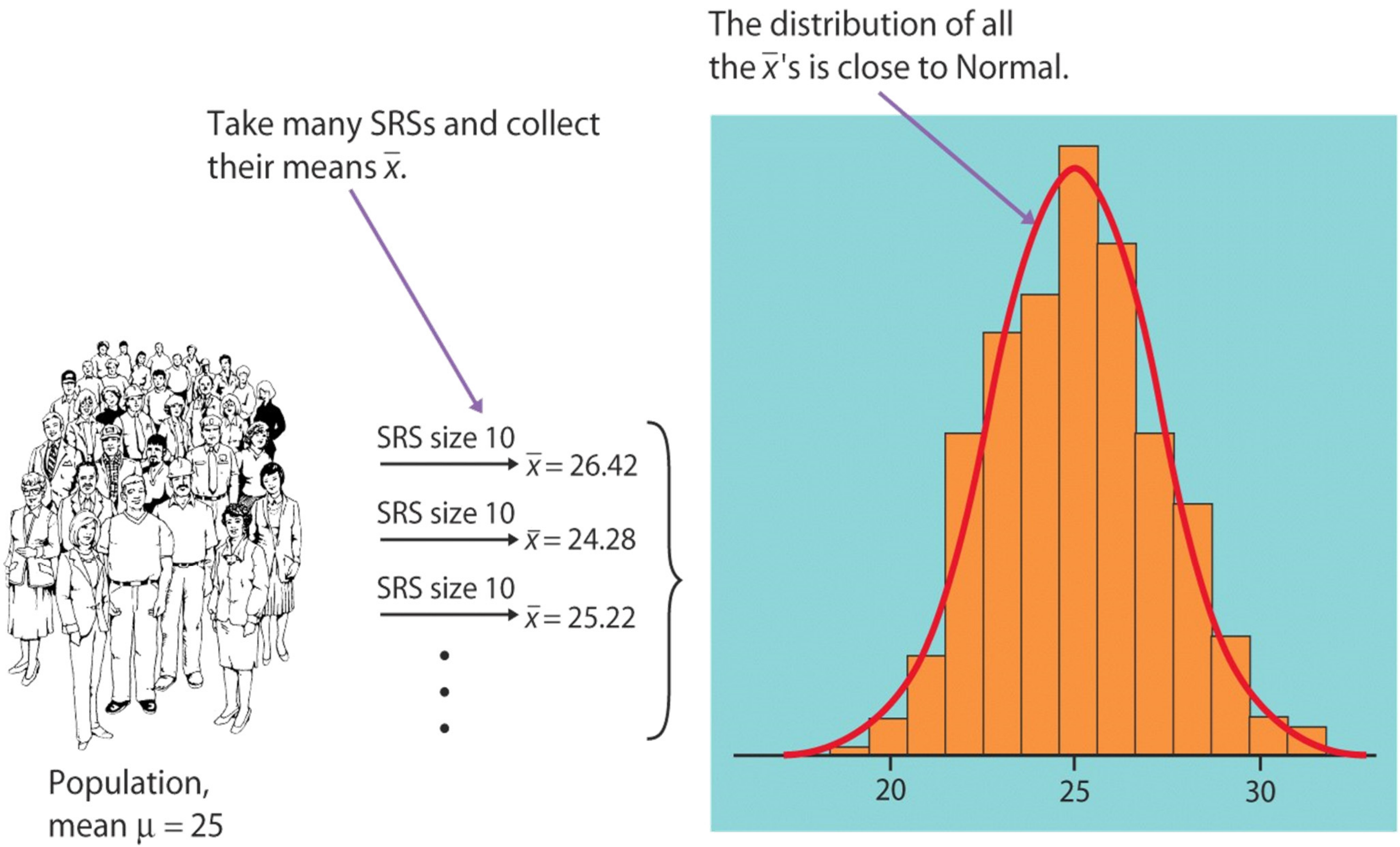
Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\overline{x}$ of the observed values gets closer and closer to the mean μ of the population.

How sample means approach the population mean (μ=25).



Example

- what would happen in many samples?



Recall Some Features of the Sampling Distribution

- It will approximate a normal curve even if the population you started with does NOT look normal

- Sampling distribution serves as a bridge between the sample and the population

Mean of a sample mean $\overline{x}$

First Property: The Mean

- The mean of the sampling distribution of the mean equals the mean of the population

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

Standard Deviation of a sample mean $\bar{x}$

Second Property: The Standard Error

- The standard error of the mean is an approximate measure of the amount by which sample means deviate from the population mean

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

Third Property: Sample Size and the Standard Deviation

- The larger the sample size, the smaller the standard deviation of the mean $\bar{x}$

Or

- As n increases, the standard deviation of the mean decreases

Example

- Population standard deviation = 100

$$For\ n = 10,\ \sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{10}} = 31.62$$

$$For\ n = 100,\ \sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10.00$$

$$For\ n = 1000,\ \sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{1000}} = 3.16$$

Sampling distribution of a sample mean $\bar{x}$

- Definition: For a random variable x and a given sample size n, the distribution of the variable $\bar{x}$, that is the distribution of all possible sample means, is called the sampling distribution of the sample mean.

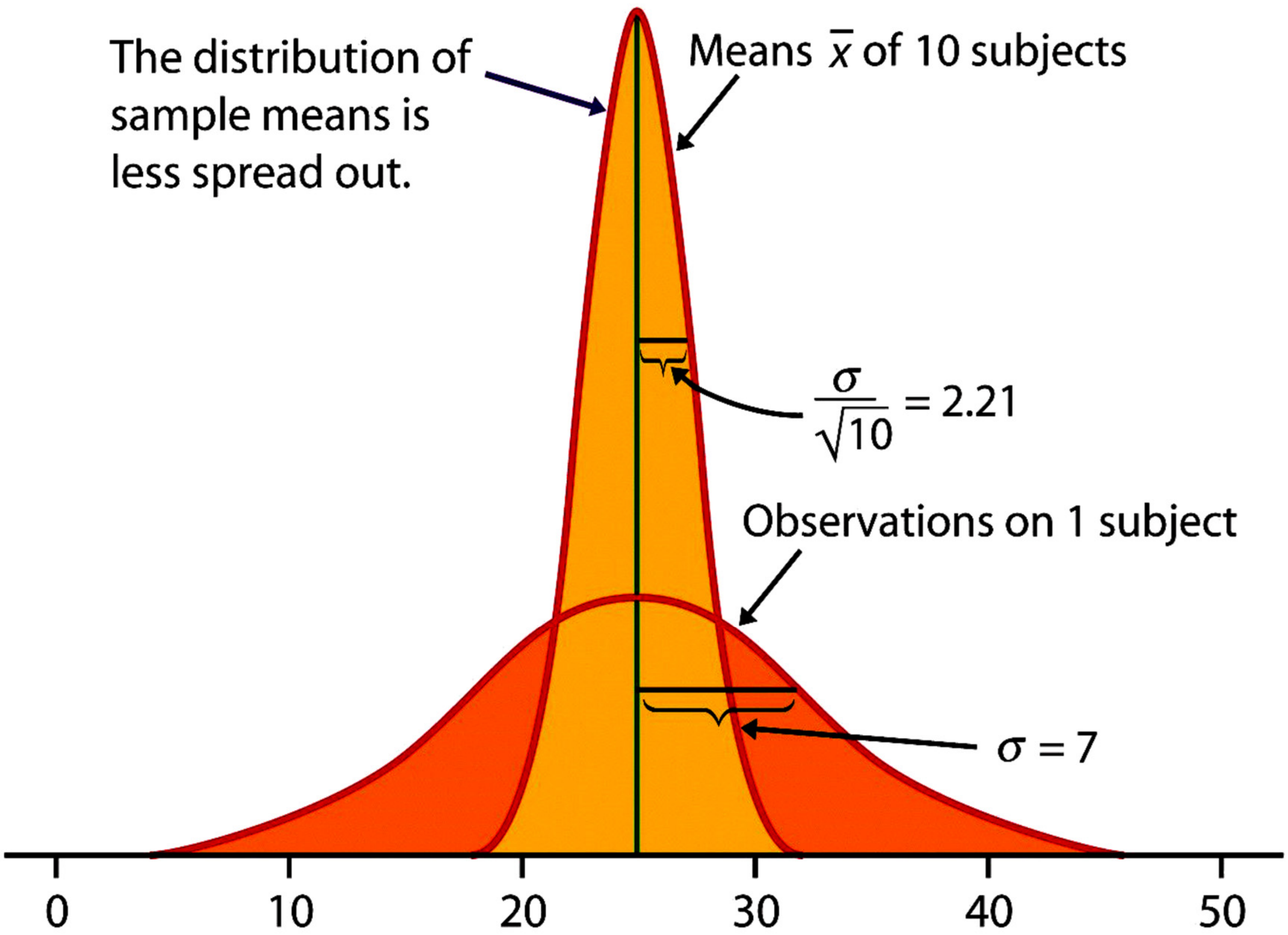
<u>Sampling distribution of the sample mean</u>

- Case 1. Population follows Normal distribution
  - Draw an SRS of size **n** from any population.
  - Repeat sampling.
  - Population follows a Normal distribution with mean **μ** and standard deviation **σ**.
  - Sampling distribution of $\overline{x}$ follows normal distribution as follows: **N(μ, σ/√n).**

$\sigma / \sqrt{n}$

Example

(The population distribution follow a Normal distribution, then so does the sample mean)



The central limit theorem

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

This theorem tells us:

1. Small samples: Shape of sampling distribution is less normal
2. Large sample: Shape of sampling distribution is more normal.

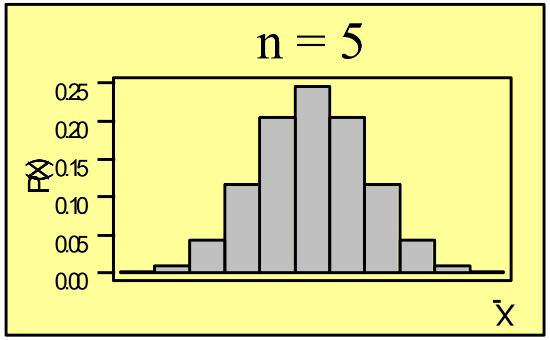
Sampling distribution of the sample mean

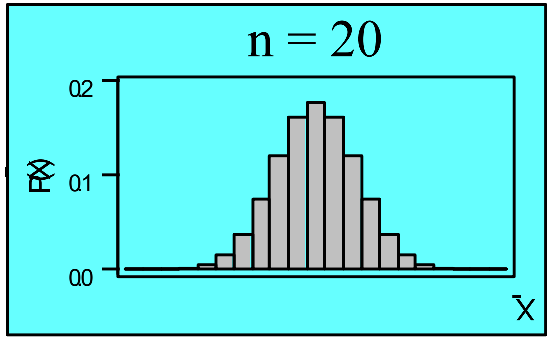
- Case 2. Population follows any distribution (CLT: Central limit theorem)
  - Draw an SRS of size **n** from any population.
  - Repeat sampling.
  - Population follows *a distribution* with mean **μ** and standard deviation **σ**.
  - When **n is large** (n>=30), sampling dist of $\overline{x}$ follows approximately Normal distribution as follows **N(μ, σ/√n).**

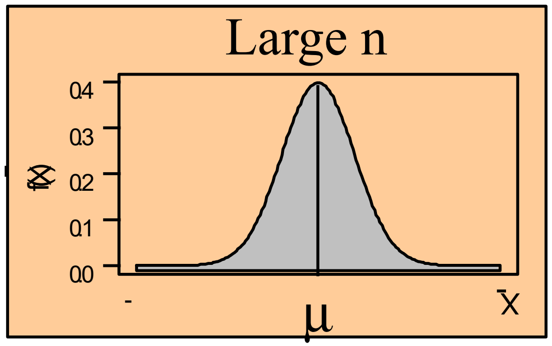
The Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean will tend to be a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large (n >30).

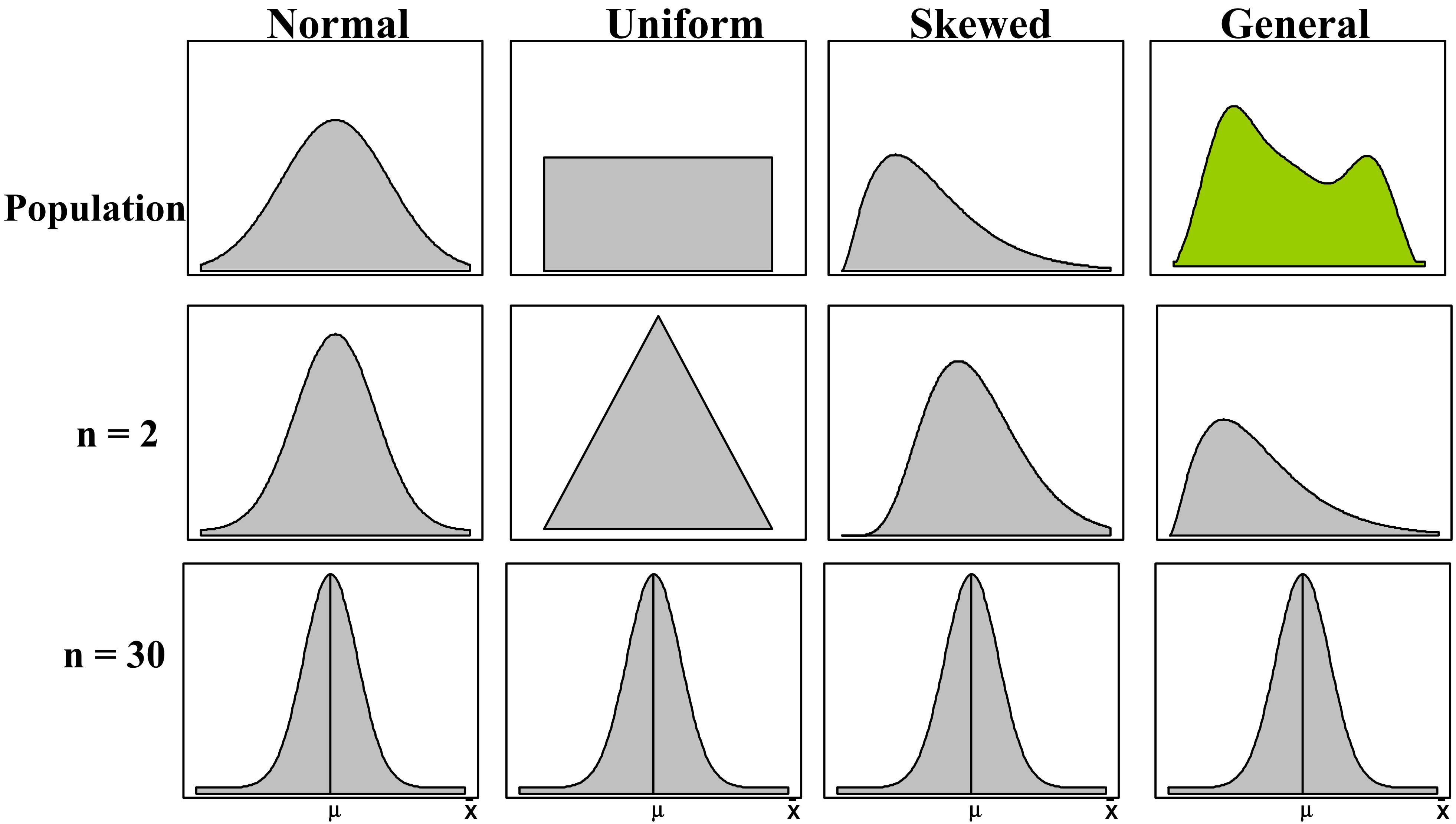
For "large enough" n: $\bar{X} \sim N(\mu, \sigma^2/n)$







The Central Limit Theorem Applies to Sampling Distributions from **Any** Population



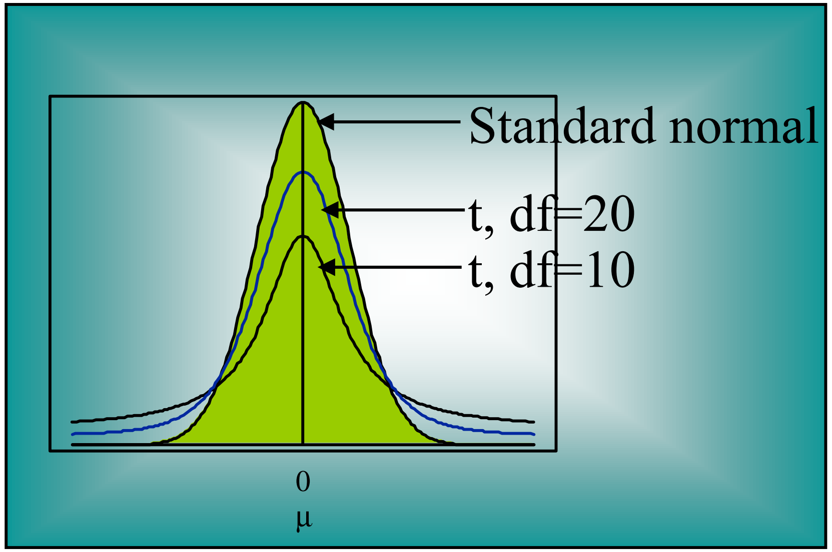
Student's *t* Distribution

If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, *s*. If the population is normal, the resulting statistic:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

has a ***t distribution* with *(n - 1) degrees of freedom***.

- The *t* is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of *t* is 0.
- The variance of *t* is greater than 1, but approaches 1 as the number of degrees of freedom increases.
- The *t* distribution approaches a standard normal as the number of degrees of freedom increases.
- When the sample size is small (<30) we use t distribution.



Sampling Distributions

Finite Population Correction Factor

If the sample size is more than 5% of the population size and the sampling is done without replacement, then a correction needs to be made to the standard error of the means.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \bullet \sqrt{\frac{N-n}{N-1}}$$

Sampling Distribution of $\bar{x}$

Standard Deviation of $\bar{x}$

Finite Population

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

Infinite Population

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

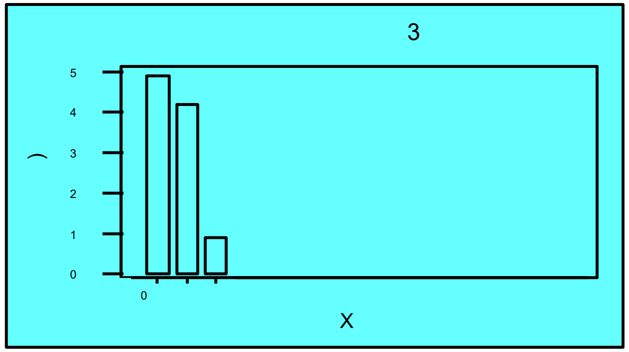
- A finite population is treated as being infinite if $n/N \leq .05$.
- $\sqrt{(N-n)/(N-1)}$ is the finite correction factor.
- $\sigma_{\bar{x}}$ is referred to as the standard error of the mean.

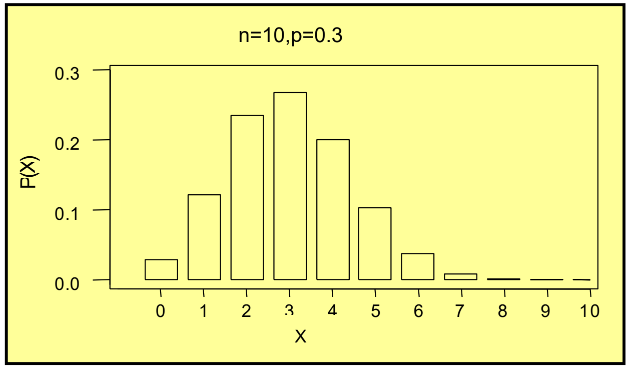
The Sampling Distribution of the Sample Proportion, $\hat{p}$

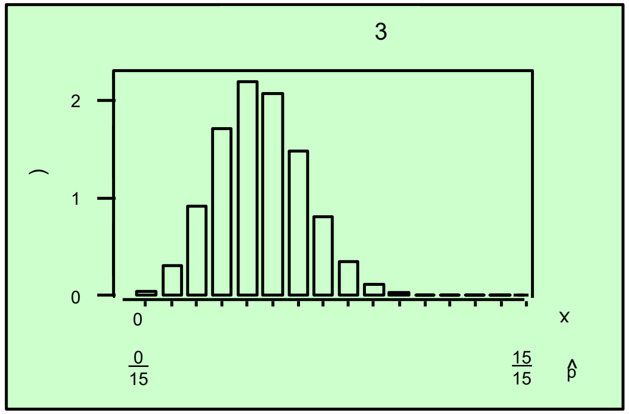
The **sample proportion** is the percentage of successes in *n* binomial trials. It is the number of successes, *X*, divided by the number of trials, *n*.

Sample proportion: $$\hat{p} = \frac{X}{n}$$

As the sample size, *n,* increases, the sampling distribution of $\hat{p}$ approaches a **normal distribution** with mean *p* and standard deviation $\sqrt{\frac{p(1-p)}{n}}$







Statistical inference: CLT, confidence intervals, p-values

The process of making guesses about the truth from a sample

Sample statistics

$$\hat{\mu} = \overline{X}_n = \frac{\sum_{i=1}^{n} x}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X}_n)^2}{n-1}$$

*hat notation ^ is often used to indicate "estitmate"

Truth (not observable)

Population parameters

$$\mu = \frac{\sum_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample (observation)

Make guesses about the whole population

Statistics vs. Parameters

- **<u>Sample Statistic</u>** – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient
  - E.g., the mean Vit-D level in a sample of 100 men is 63 nmol/L
  - E.g., the correlation coefficient between vit-D and cognitive function in the sample of 100 men is 0.15

- **<u>Population parameter</u>** – the true value/true effect in the entire population of interest
  - E.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L
  - E.g., the true correlation between vitamin D and cognitive function in all middle-aged and older European men is 0.15

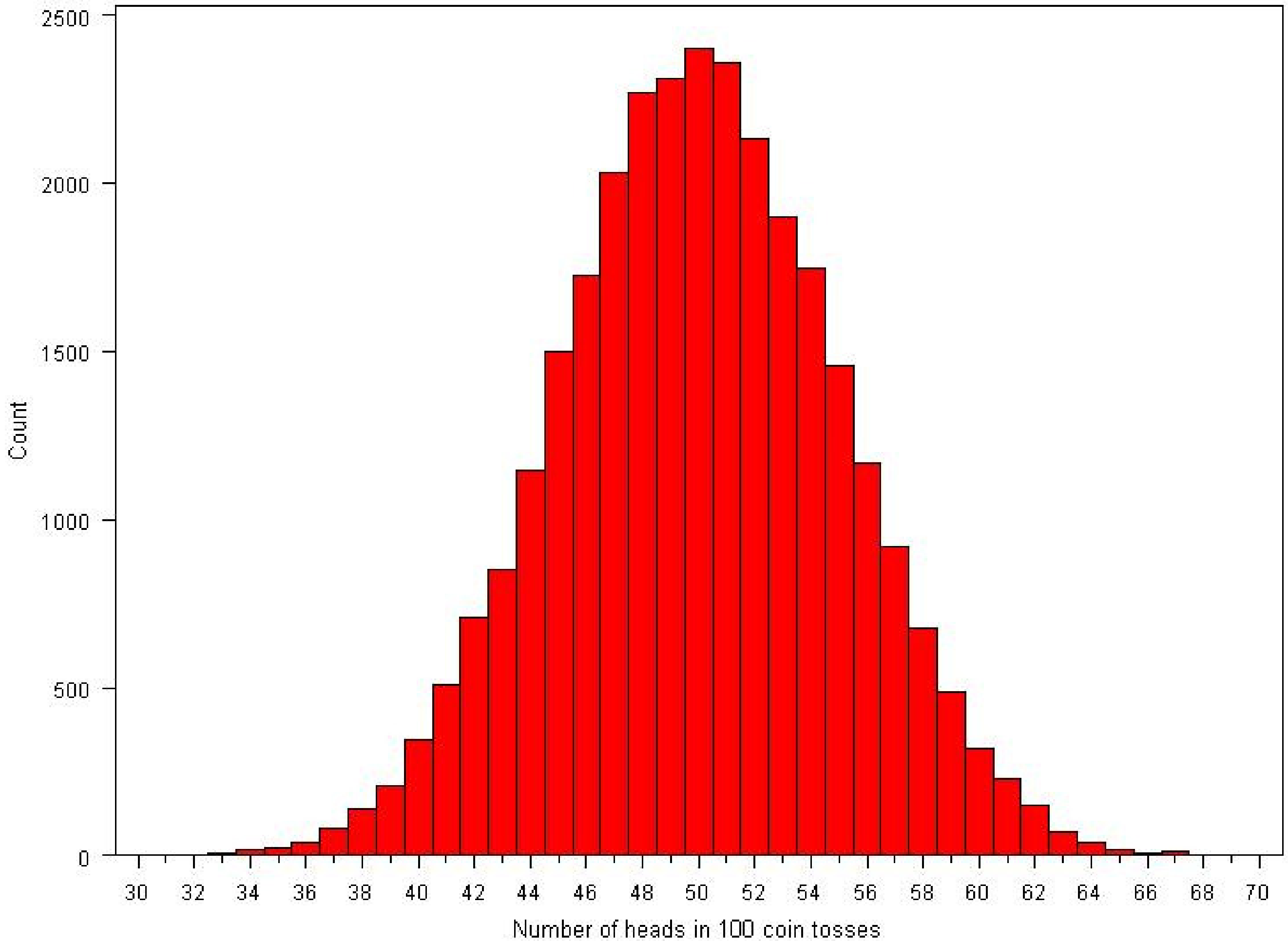
Distribution of a statistic…

- Statistics follow distributions too…
- *But the distribution of a statistic is a theoretical construct.*
- Statisticians ask a thought experiment: how much would the value of the statistic fluctuate if one could repeat a particular study over and over again with different samples of the same size?
- By answering this question, statisticians are able to pinpoint exactly how much uncertainty is associated with a given statistic.

Distribution of a statistic

- Two approaches to determine the distribution of a statistic:
  - 1. Computer simulation
    - Repeat the experiment over and over again virtually!
    - More intuitive; can directly observe the behavior of statistics.
  - 2. Mathematical theory
    - Proofs and formulas!
    - More practical; use formulas to solve problems.

Coin tosses…



Conclusions:

We usually get between 40 and 60 heads when we flip a coin 100 times.

It's extremely unlikely that we will get 30 heads or 70 heads (didn't happen in 30,000 experiments!).

Distribution of the sample mean, computer simulation…

- 1. Specify the underlying distribution of vitamin D in all European men aged 40 to 79.
  - Right-skewed
  - Standard deviation = 33 nmol/L
  - True mean = 62 nmol/L (this is arbitrary; does not affect the distribution)
- 2. Select a random sample of 100 virtual men from the population.
- 3. Calculate the mean vitamin D for the sample.
- 4. Repeat steps (2) and (3) a large number of times (say 1000 times).
- 5. Explore the distribution of the 1000 means.

Mathematical Theory…
The **Central Limit Theorem**!

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $$\mu_{\bar{x}} = \mu$$

2. have standard deviation: $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*). It all comes back to Z!

Symbol Check

$\mu_{\bar{x}}$ The mean of the sample means.

$\sigma_{\bar{x}}$ The standard deviation of the sample means. *Also called "the standard error of the mean."*

Mathematical Proof (optional!)

If X is a random variable from any distribution with known mean, E(x), and variance, Var(x), then the expected value and variance of the average of n observations of X is:

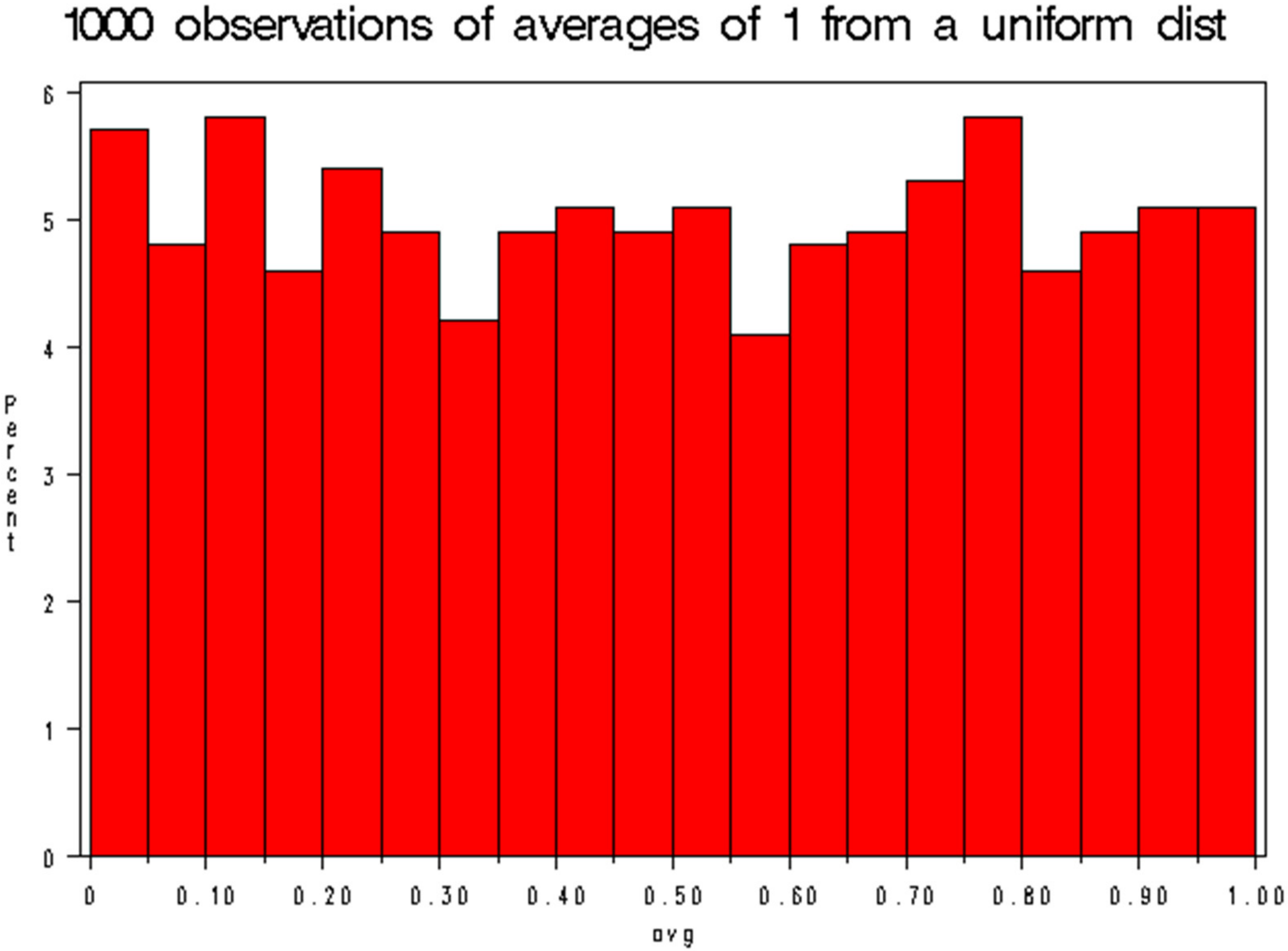
$$E(\bar{X}_n) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} E(x)}{n} = \frac{nE(x)}{n} = E(x)$$

$$Var(\bar{X}_n) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} Var(x)}{n^2} = \frac{nVar(x)}{n^2} = \frac{Var(x)}{n}$$

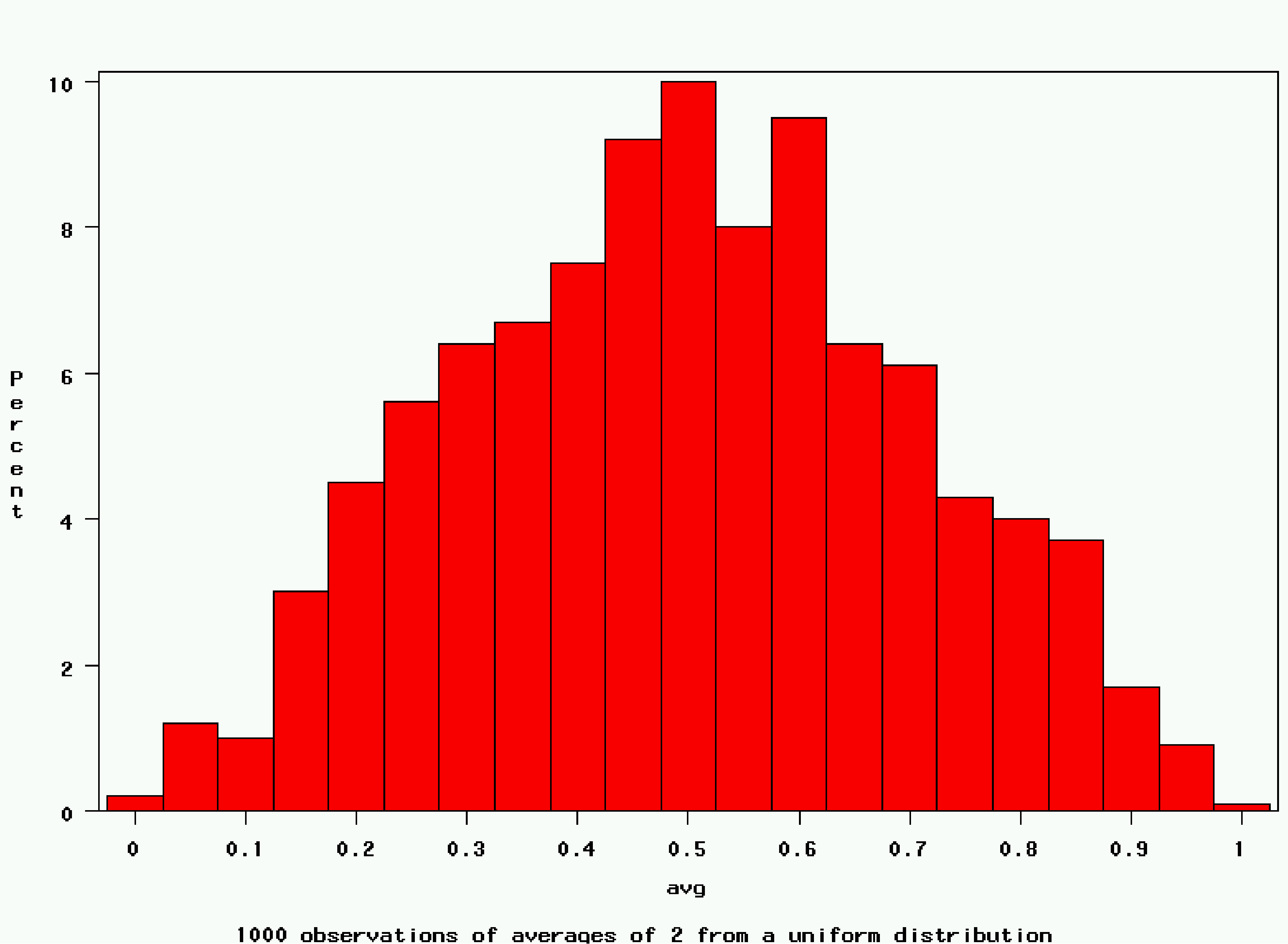
Computer simulation of the CLT:

1. Pick any probability distribution and specify a mean and standard deviation.
2. Tell the computer to randomly generate 1000 observations from that probability distributions
   - E.g., the computer is more likely to spit out values with high probabilities
3. Plot the "observed" values in a histogram.
4. Next, tell the computer to randomly generate 1000 averages-of-2 (randomly pick 2 and take their average) from that probability distribution. Plot "observed" averages in histograms.
5. Repeat for averages-of-10, and averages-of-100.

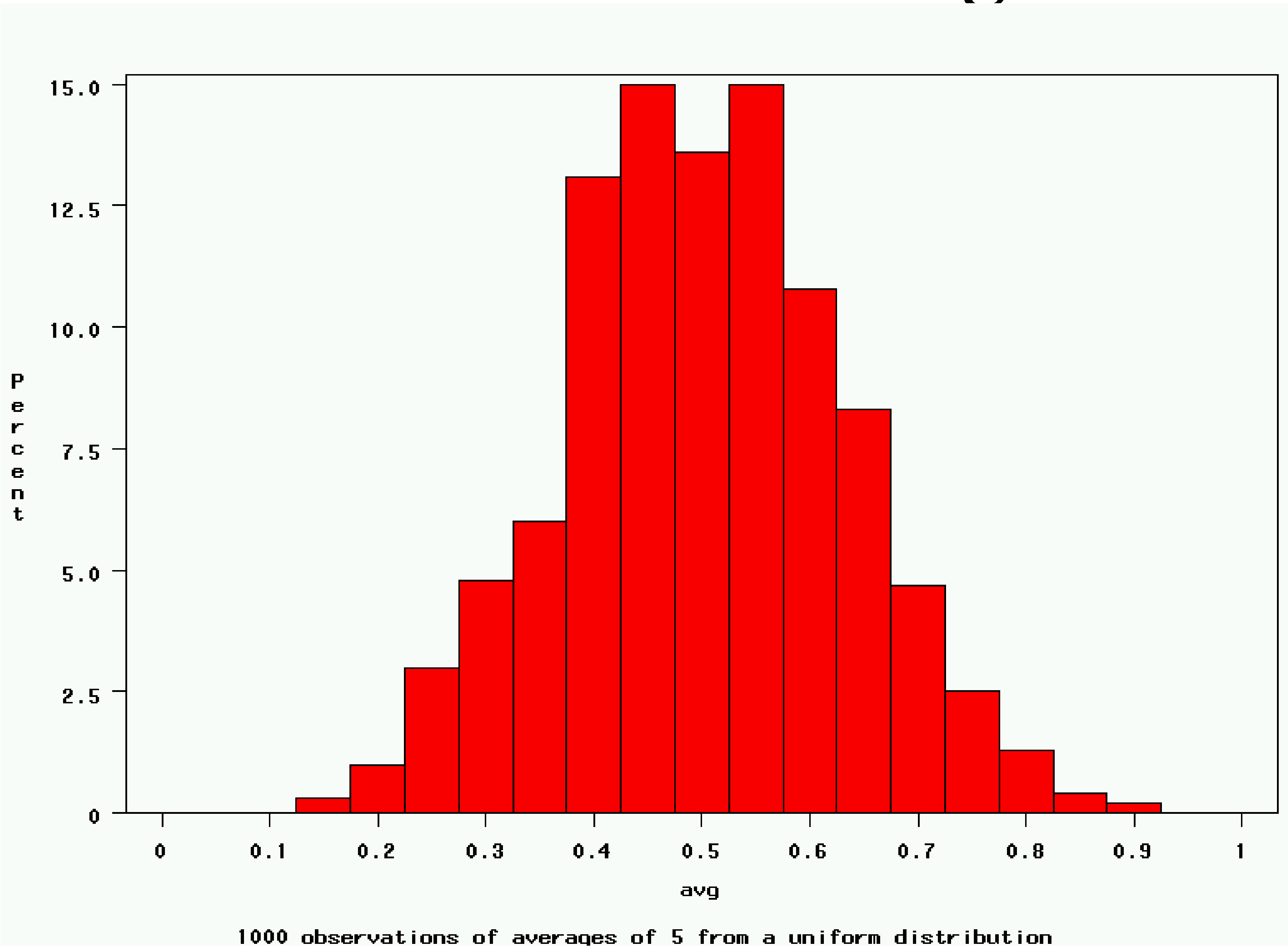
Uniform on [0,1]: average of 1 (original distribution)



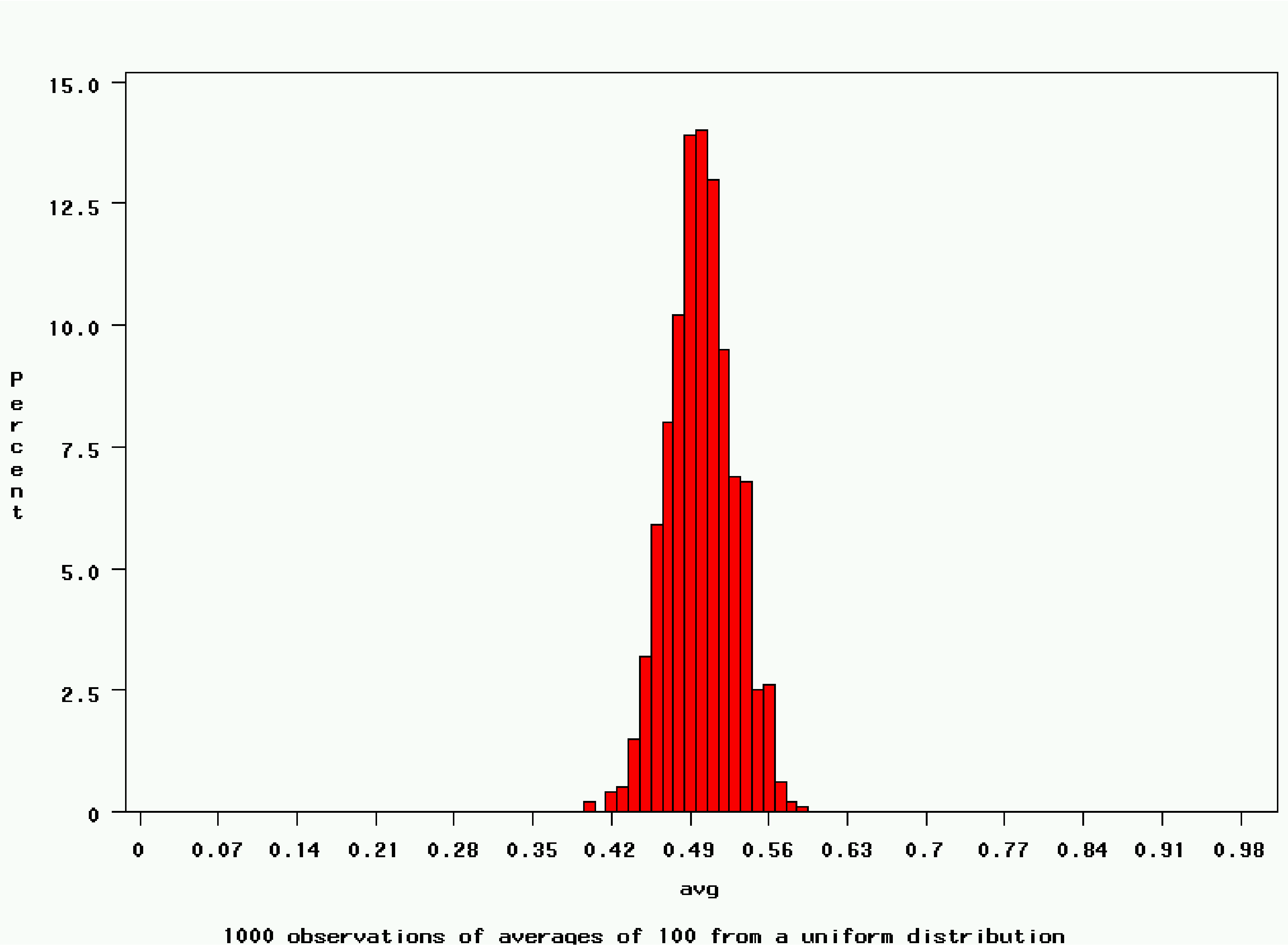
Uniform: 1000 averages of 2



Uniform: 1000 averages of 5



Uniform: 1000 averages of 100



The Central Limit Theorem: (revisited)

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*)

Distribution of the sample mean

Statistical inference about the population mean is of prime practical importance. Inferences about this parameter are based on the sample mean and its sampling distribution.

Mean and Standard Deviation of $\overline{X}$

The distribution of the sample mean, based on a random sample of size n, has

$$E(\overline{X}) = \mu \qquad (= \text{Population mean})$$

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n} \qquad \left(= \frac{\text{Population variance}}{\text{Sample size}} \right)$$

$$\text{sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}} \qquad \left(= \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}} \right)$$

$\overline{X}$ Is Normal When Sampling from a Normal Population

In random sampling from a **normal** population with mean μ and standard deviation σ, the sample mean $\overline{X}$ has the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Central Limit Theorem

Whatever the population, the distribution of $\overline{X}$ is approximately normal when n is large.

In random sampling from an arbitrary population with mean μ and standard deviation σ, when n is large, the distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Consequently,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{is approximately } N(0,\ 1)$$

Standardizing in mathematical statistics [edit]

Further information: Normalization (statistics)

In mathematical statistics, a random variable *X* is **standardized** by subtracting its expected value $\operatorname{E}[X]$ and dividing the difference by its standard deviation $\sigma(X) = \sqrt{\operatorname{Var}(X)}$:

$$Z = \frac{X - \operatorname{E}[X]}{\sigma(X)}$$

If the random variable under consideration is the sample mean of a random sample $X_1, \dots, X_n$ of *X*:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then the standardized version is

$$Z = \frac{\bar{X} - \operatorname{E}[X]}{\sigma(X)/\sqrt{n}}.$$

T-score [edit]

"T-score" redirects here. It is not to be confused with t-statistic.

A **T-score** is a standard score Z shifted and scaled to have a mean of 50 and a standard deviation of 10.[3][4][5]

Calculation from raw score [edit]

The standard score of a raw score x[1] is

$$z = \frac{x - \mu}{\sigma}$$

where:

It measures the sigma distance of actual data from the average.

The Z value provides an assessment of how off-target a process is operating.

Applications [edit]

Main article: Z-test

The z-score is often used in the z-test in standardized testing – the analog of the Student's t-test for a population whose parameters are known, rather than estimated. As it is very unusual to know the entire population, the t-test is much more widely used.

Also, standard score can be used in the calculation of prediction intervals. A prediction interval [L,U], consisting of a lower endpoint designated L and an upper endpoint designated U, is an interval such that a future observation X will lie in the interval with high probability γ, i.e.

$$P(L < X < U) = \gamma,$$

For the standard score Z of X it gives:[2]

$$P\left(\frac{L - \mu}{\sigma} < Z < \frac{U - \mu}{\sigma}\right) = \gamma.$$

By determining the quantile z such that

$$P\left(-z < Z < z\right) = \gamma$$

it follows:

Example on probability calculations for the sample mean

Consider a population with mean 82 and standard deviation 12.

If a random sample of size 64 is selected, what is the probability that the sample mean will lie between 80.8 and 83.2?

Solution: We have μ = 82 and σ = 12. Since n = 64 is large, the central limit theorem tells us that the distribution of the sample mean is approximately normal with

$$E(\overline{X}) = \mu = 82, \; sd(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{64}} = 1.5$$

Converting to the standard normal variable:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 82}{1.5}$$

Thus,

$$P[80.8 < \overline{X} < 83.2]$$

$$= P[(80.8 - 82)/1.5 < Z < (83.2 - 82)/1.5]$$

$$= P[-.8 < Z < .8] = .7881 - .2119 = .5762$$

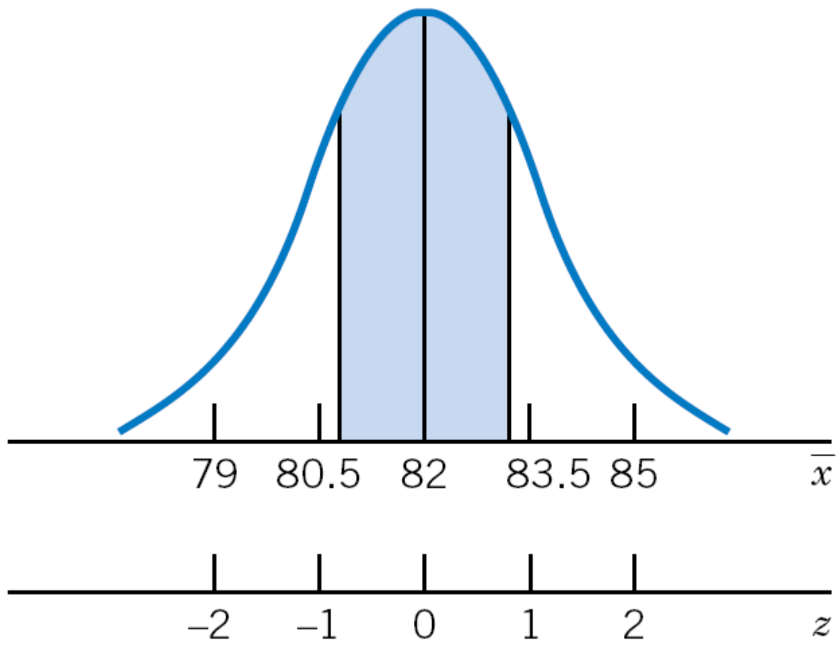
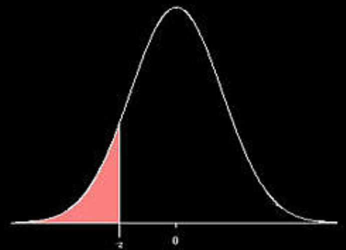
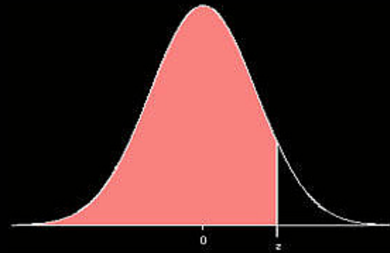


Table of Standard Normal Probabilities for Negative Z-scores



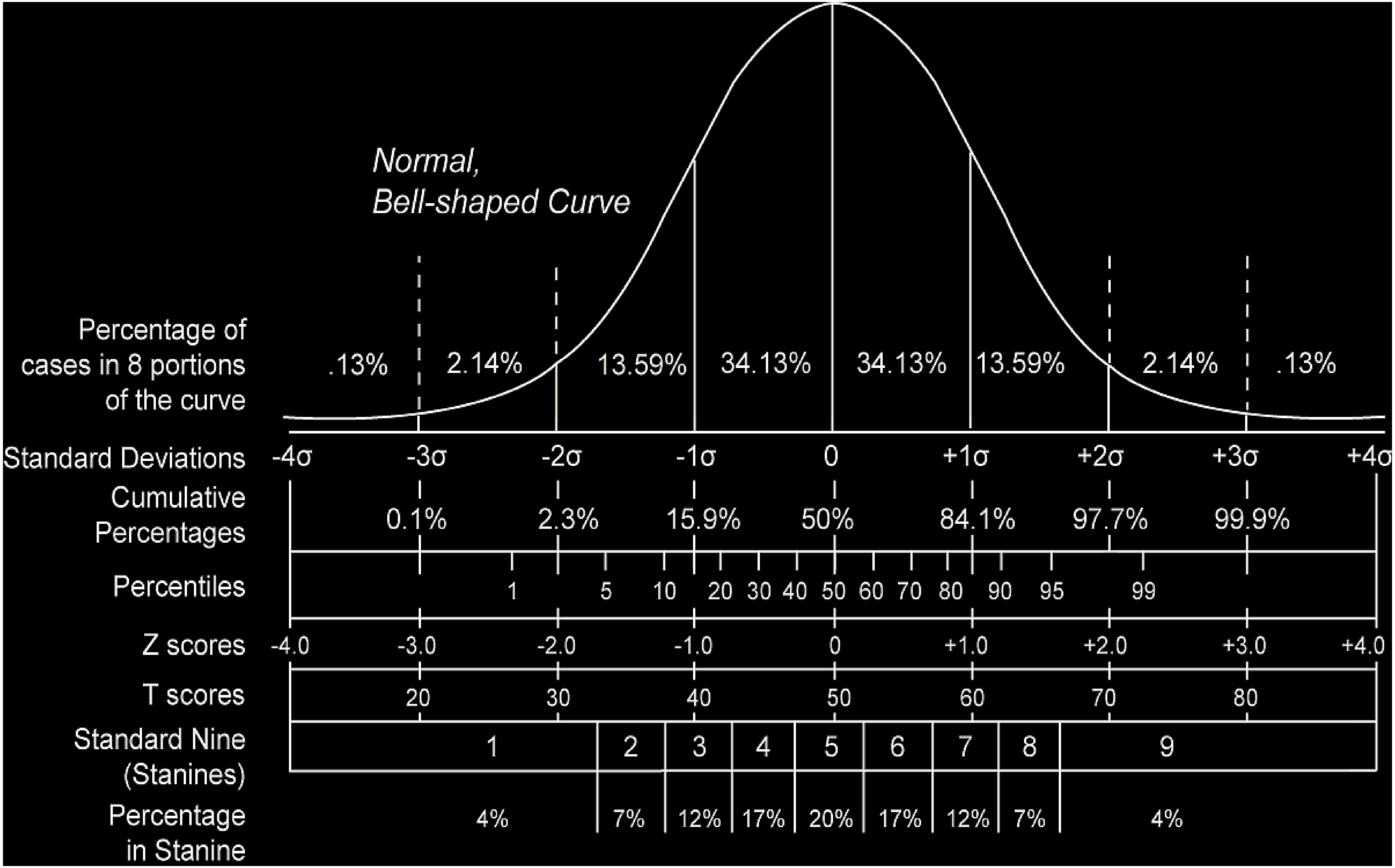
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score



Normal, Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ -3σ -2σ -1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percentages
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
Percentiles
1 5 10 20 30 40 50 60 70 80 90 95 99
Z scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0
T scores
20 30 40 50 60 70 80
Standard Nine (Stanines)
1 2 3 4 5 6 7 8 9
Percentage in Stanine
4% 7% 12% 17% 20% 17% 12% 7% 4%

The Central Limit Theorem more formally

The Central Limit Theorem

If repeated random samples of size N are drawn from a population that is normally distributed along some variable Y, having a mean μ and a standard deviation σ, then the sampling distribution of all theoretically possible sample means will be a normal distribution having a mean μ and a standard deviation $\hat{\sigma}$ given by $\frac{s_Y}{\sqrt{n}}$

[Sirkin (1999), p. 239]

	Mean	*Standard Deviation*	*Variance*
Universe	μ_Y	σ_Y	σ_Y^2
Sampling Distribution	μ_Y	$\hat{\sigma}_Y$	$\hat{\sigma}_Y^2$
Sample	$\bar{Y}$	s_Y	s_Y^2

The Standard Error

$$\hat{\sigma} = \frac{s_Y}{\sqrt{N}}$$

where s_Y = sample standard deviation
and N = sample size

Let's assume that we have a **random sample** of **200** USC undergraduates. Note that this is both a large and a random sample, hence the *Central Limit Theorem* applies to *any* statistic that we calculate from it. Let's pretend that we asked these 200 randomly-selected USC students to tell us their **grade point average** (GPA). (Note that our statistical calculations assume that all 200 [a] knew their current GPA and [b] were telling the truth about it.) We calculated the **mean** GPA for the sample and found it to be **2.58**. Next, we calculated the **standard deviation** for these self-reported GPA values and found it to be **0.44**.

The **standard error** is nothing more than the ***standard deviation*** of the *sampling distribution*. The Central Limit Theorem tells us how to estimate it:

$$\hat{\sigma} = \frac{s_Y}{\sqrt{N}}$$

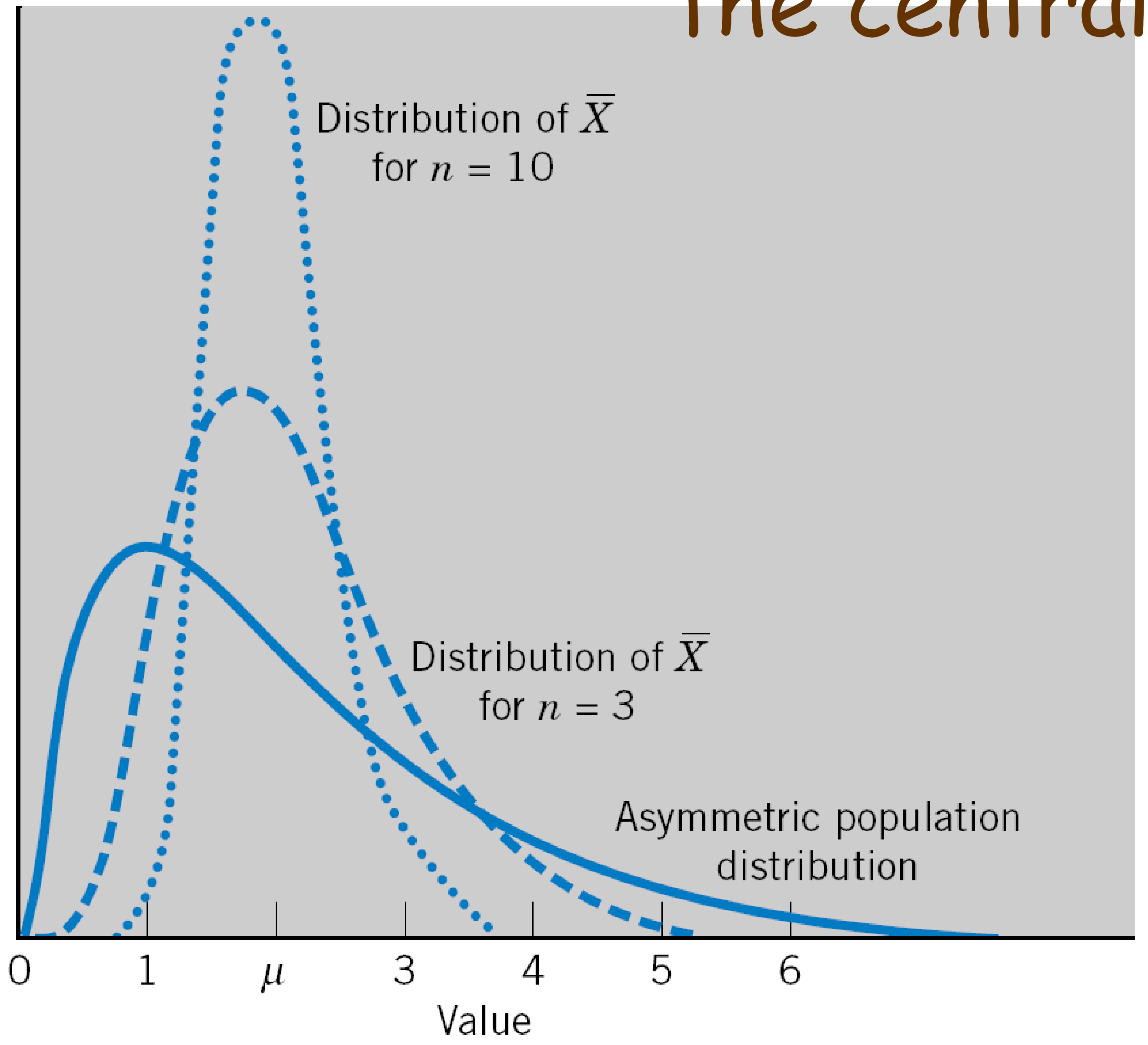
The **standard error** is estimated by ***dividing*** the **standard deviation** of the sample by the ***square root*** of the **size of the sample**. In our example,

$$\hat{\sigma} = \frac{0.44}{\sqrt{200}}$$

$$\hat{\sigma} = \frac{0.44}{14.142}$$

$$\hat{\sigma} = 0.031$$

An example illustrating the central limit theorem



Distributions of $\overline{X}$ for n = 3 and n = 10 in sampling from an asymmetric population.

Recapitulation

1. The Central Limit Theorem holds only for ***large***, ***random*** samples.
2. When the Central Limit Theorem holds, the ***mean*** of the ***sampling distribution*** μ is equal to the ***mean*** in the ***universe*** (also μ).
3. When the Central Limit Theorem holds, the ***standard deviation*** of the *sampling distribution* (called the ***standard error***, $\hat{\sigma}_Y$) is estimated by

$$\hat{\sigma} = \frac{s_Y}{\sqrt{N}}$$

Recapitulation (continued)

4. When the Central Limit Theorem holds, the sampling distribution is ***normally shaped***.
5. All normal distributions are ***symmetrical***, ***asymptotic***, and have areas that are ***fixed*** and ***known***.

In statistics, a **confidence interval** (**CI**) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that potentially includes the unobservable true parameter of interest.

How frequently the observed interval contains the true parameter if the experiment is repeated is called the **confidence level**. In other words, if confidence intervals are constructed in separate experiments on the same population following the same process, the proportion of such intervals that contain the true value of the parameter will match the given confidence level. <WIKI>

Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. However, the interval computed from a particular sample does not necessarily include the true value of the parameter.

Confidence intervals are commonly reported in tables or graphs, to show the reliability of the estimates. For example, a confidence interval can be used to describe how reliable survey results are.

In applied practice, confidence intervals are typically stated at the 95% confidence level. However, when presented graphically, confidence intervals can be shown at several confidence levels, for example 90%, 95% and 99%.

Certain factors may affect the confidence interval size including size of sample, level of confidence, and population variability. A larger sample size normally will lead to a better estimate of the population parameter.

In statistical inference, the concept of a **confidence distribution (CD)** has often been loosely referred to as a distribution function on the parameter space that can represent confidence intervals of all levels for a parameter of interest.

In statistics, a **confidence region** is a multi-dimensional generalization of a confidence interval. It is a set of points in an n-dimensional space, often represented as an ellipsoid around a point which is an estimated solution to a problem, although other shapes can occur.

A **confidence band** is used in statistical analysis to represent the uncertainty in an estimate of a curve or function based on limited or noisy data.

The explanation of a **confidence interval** can amount to something like: "The confidence interval represents values for the population parameter for which the difference between the parameter and the observed estimate is not **statistically significant** at the 10% level“ (assuming 90% confidence interval as an example). In fact, this relates to one particular way in which a confidence interval may be constructed.

The following applies: If the true value of the parameter lies outside the 90% confidence interval once it has been calculated, then a sampling event has occurred which had a probability of 10% (or less) of happening by chance.

In statistical hypothesis testing**,** **statistical significance** (or a statistically significant result) is attained whenever the observed **p-value** of a test statistic is less than the **significance level** defined for the study.

The **p-value** is the probability of obtaining results at least as extreme as those observed, given that the null hypothesis is true. The **significance level, α**, is the probability of rejecting the null hypothesis, given that it is true.

In statistics, the p-value is the probability that, using a given statistical model, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or more extreme than the actual observed results.

The *p*-value is defined as the probability, under the assumption of hypothesis H, of obtaining a result equal to or more extreme than what was actually observed.

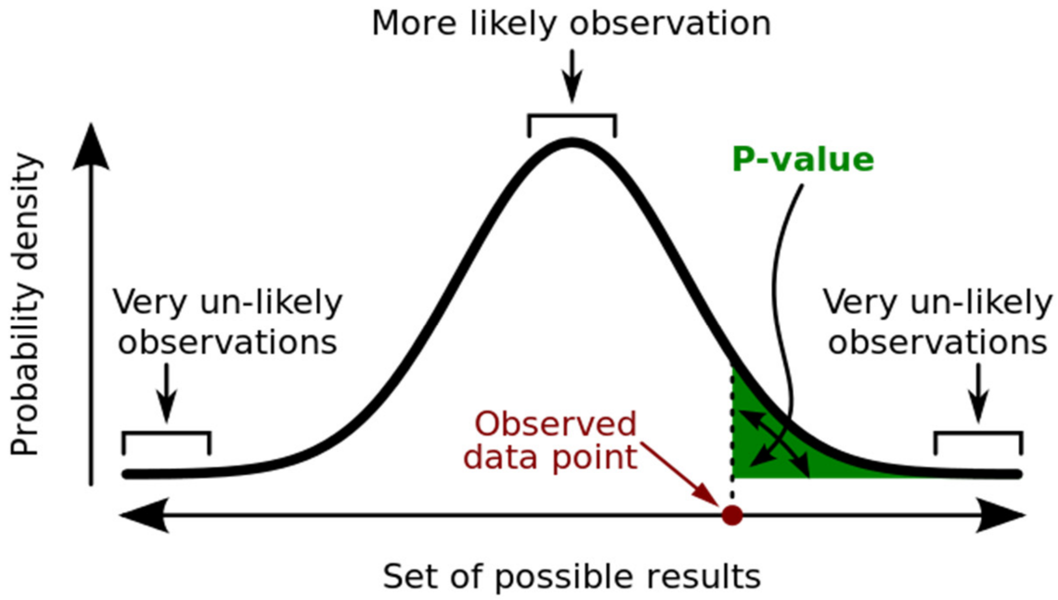
The p-value is used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence.

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

The smaller the *p*-value, the larger the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation. The hypothesis H is rejected if any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value α, which is referred to as the level of significance. Unlike the *p*-value, the α level is not derived from any observational data and does not depend on the underlying hypothesis; the value of α is instead determined by the consensus of the research community that the investigator is working in.

Statistical significance plays a pivotal role in **statistical hypothesis testing**. It is used to determine whether the **null hypothesis** should be rejected or retained. The null hypothesis is the default assumption that nothing happened or changed. For the null hypothesis to be rejected, an observed result has to be statistically significant, i.e. the observed ***p*-value** is less than the pre-specified significance level.

To determine whether a result is statistically significant, a researcher calculates a *p*-value, which is the probability of observing an effect given that the null hypothesis is true. The null hypothesis is rejected if the *p*-value is less than a predetermined level, α. **α is called the significance level**, and is the probability of rejecting the null hypothesis given that it is true (a type I error – false hit). It is usually set at or below 5%.

For example, when α is set to 5%, the conditional probability of a type I error, *given that the null hypothesis is true*, is 5%, and a statistically significant result is one where the observed *p*-value is less than 5%. When drawing data from a sample, this means that the rejection region comprises 5% of the sampling distribution. These 5% can be allocated to one side of the sampling distribution, as in a one-tailed test, or partitioned to both sides of the distribution as in a two-tailed test, with each tail (or rejection region) containing 2.5% of the distribution.

A 95**% confidence interval does not mean** that 95% of the sample data lie within the interval.

A 95% confidence interval does not mean that for a given realised interval calculated from sample data there is a 95% probability the population parameter lies within the interval. Once an experiment is done and an interval calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval.

A **confidence interval** is not a range of plausible values for the sample parameter, though it may be understood as an estimate of plausible values for the population parameter.

A particular confidence interval of 95% calculated from an experiment does not mean that there is a 95% probability of a sample parameter from a repeat of the experiment falling within this interval.

What is the difference between an alpha level and a p-value?

Science sets a conservative standard to meet for a researcher to claim that s/he has made a discovery of a real phenomenon. The standard is the alpha level, usually set of .05.

Assuming that the null hypothesis is true, this means we may reject the null only if the observed data are so unusual that they would have occurred by chance at most 5 % of the time. The smaller the alpha, the more stringent the test (the more unlikely it is to find a statistically significant result).

Once the alpha level has been set, a statistic (like r) is computed. Each statistic has an associated probability value called a p-value, or the likelihood of an observed statistic occurring due to chance, given the sampling distribution.

Alpha sets the standard for how extreme the data must be before we can reject the null hypothesis. The p-value indicates how extreme the data are. We compare the p-value with the alpha to determine whether the observed data are statistically significantly different from the null hypothesis:

If the p-value is less than or equal to the alpha ($p \leq .05$), then we reject the null hypothesis, and we say the result is statistically significant.

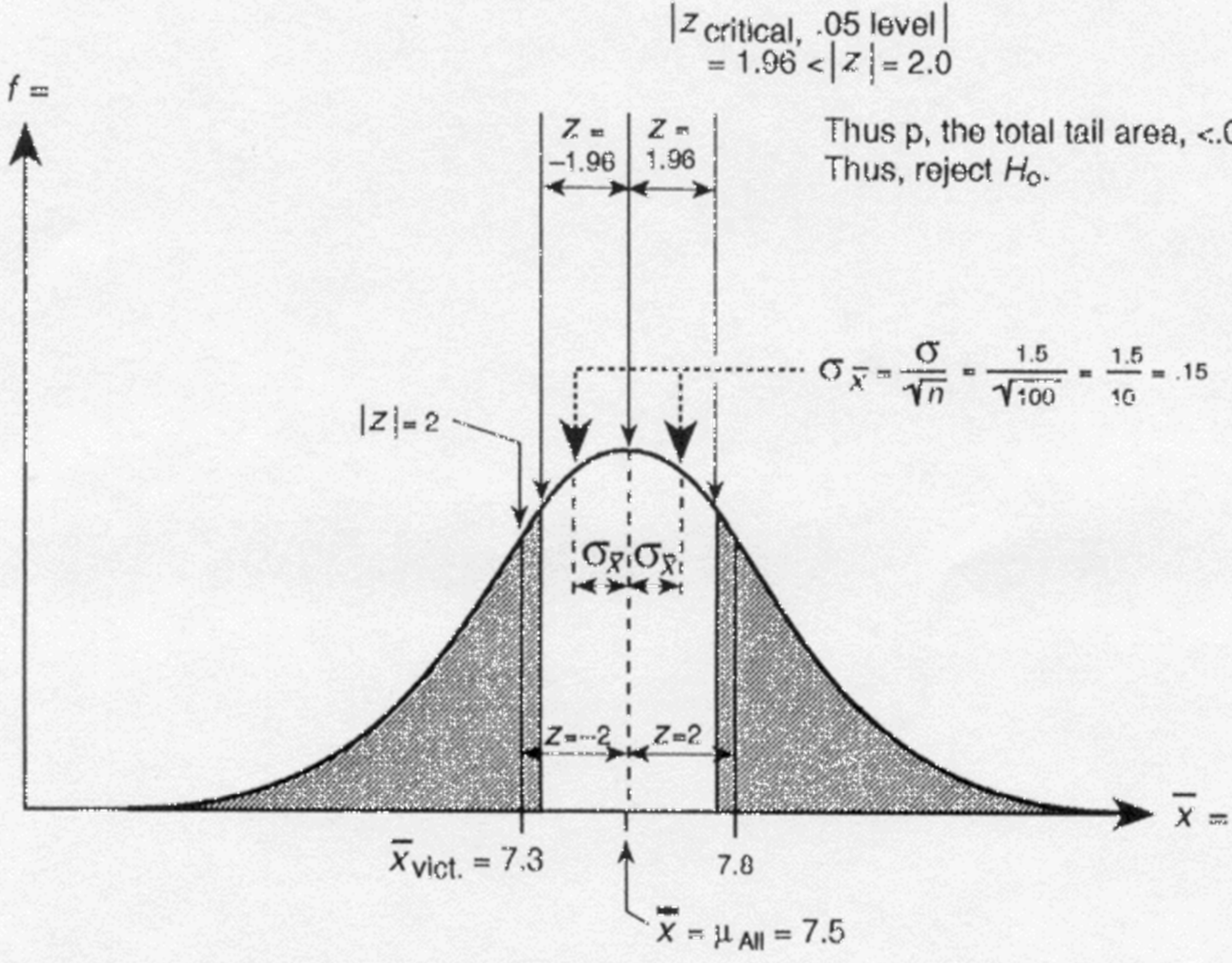
If the p-value is greater than alpha ($p > .05$), then we fail to reject the null hypothesis, and we say that the result is statistically nonsignificant (n.s.).

The basic breakdown of how to calculate a confidence interval for a population mean is as follows:

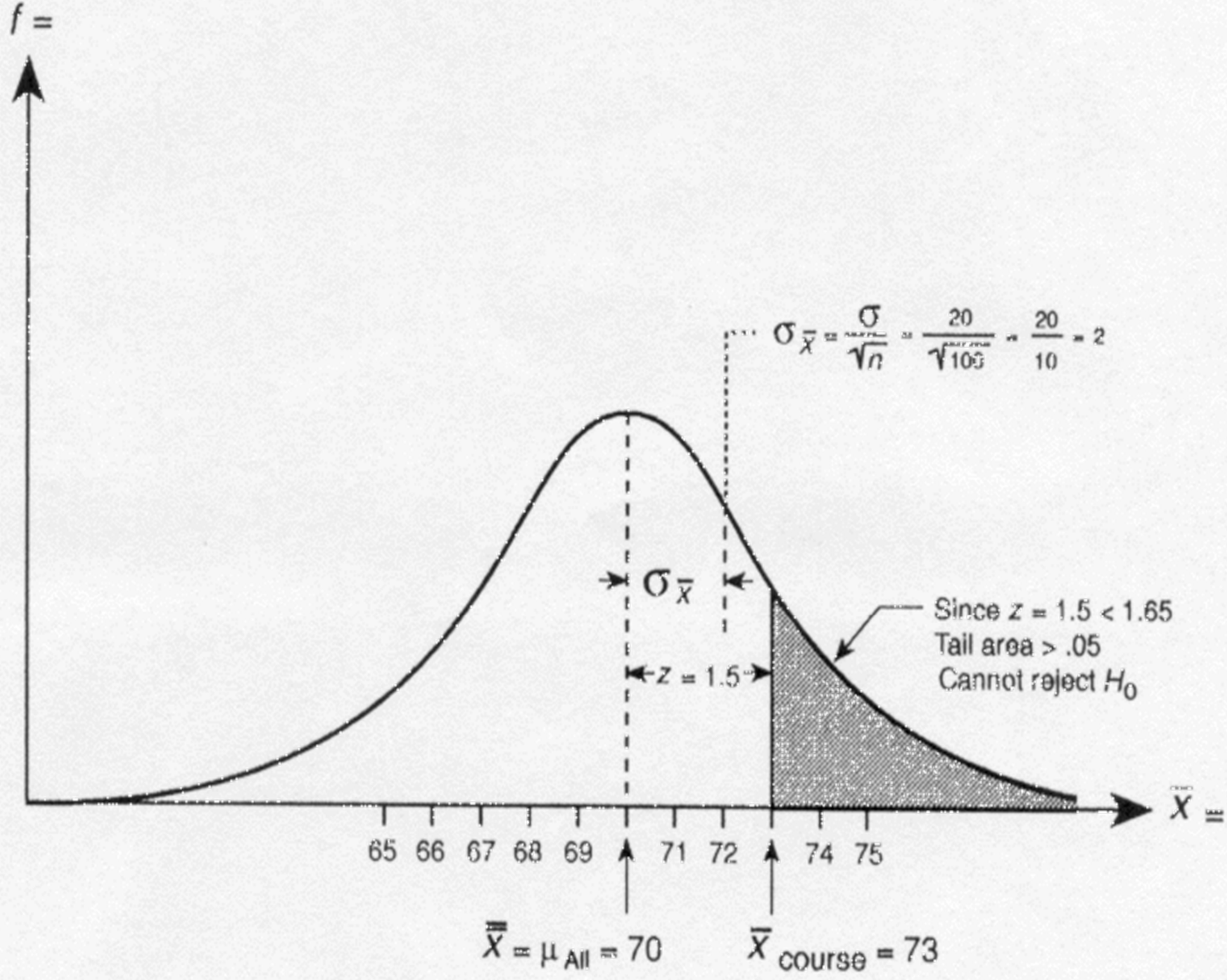
1. Identify the sample mean, $\bar{x}$. While $\bar{x}$ differs from μ, population mean, they are still calculated the same way: $\sum \frac{x_i}{n}$.
2. Identify whether the standard deviation is known, σ, or unknown, s.
   - If standard deviation is known then z* is used as the critical value. This value is only dependent on the confidence level for the test. Typical two sided confidence levels are:[23]

99%	2.576
98%	2.326
95%	1.96
90%	1.645

   - If the standard deviation is unknown then t* is used as the critical value. This value is dependent on the confidence level (C) for the test and degrees of freedom. The degrees of freedom is found by subtracting one from the number of observations, $n - 1$. The critical value is found from the t-distribution table. In this table the critical value is written as $t_\alpha(r)$, where r is the degrees of freedom and $\alpha = \frac{1-C}{2}$.
3. Plug the found values into the appropriate equations:
   - For a known standard deviation: $\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$
   - For an unknown standard deviation: $\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$



|z critical, .05 level|
= 1.96 < |z| = 2.0
f =
Thus p, the total tail area, <.05.
Thus, reject H₀.
z = −1.96
z = 1.96
σx̄ = σ/√n = 1.5/√100 = 1.5/10 = .15
|z| = 2
σx̄ σx̄
z = −2
z = 2
x̄ =
x̄vict. = 7.3
7.8
x̿ = μAll = 7.5



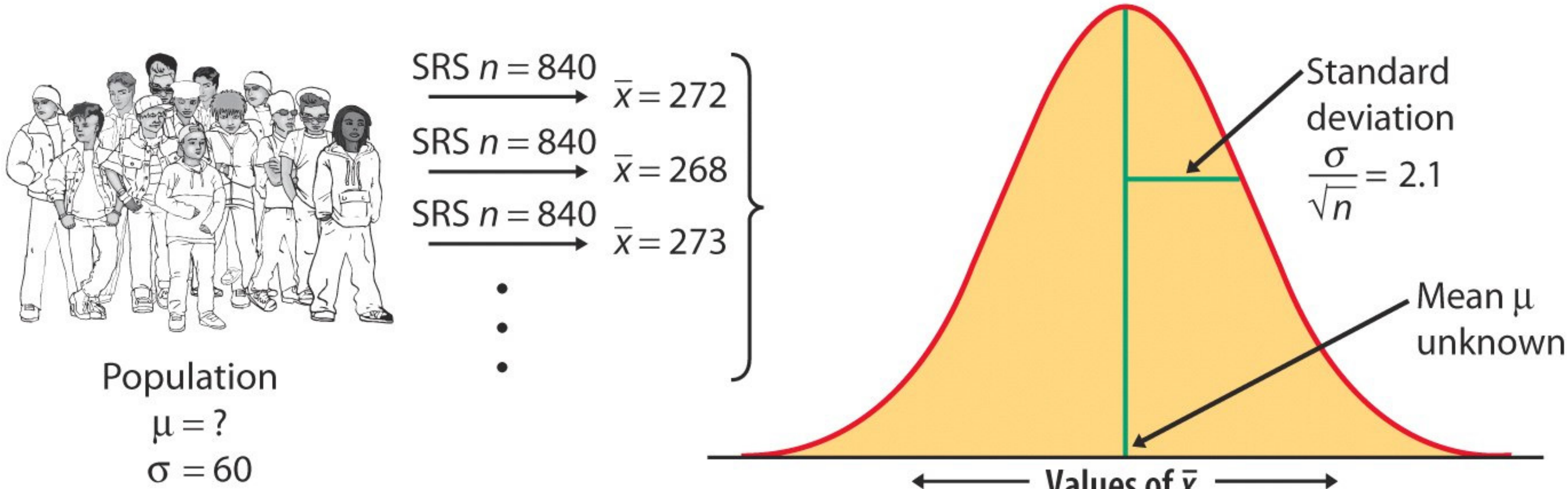
f =
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$
$\sigma_{\bar{X}}$
Since $z = 1.5 < 1.65$
Tail area > .05
Cannot reject H_0
$z = 1.5$
$\bar{X}$ =
65 66 67 68 69 71 72 74 75
$\bar{\bar{X}} = \mu_{All} = 70$
$\bar{X}_{course} = 73$

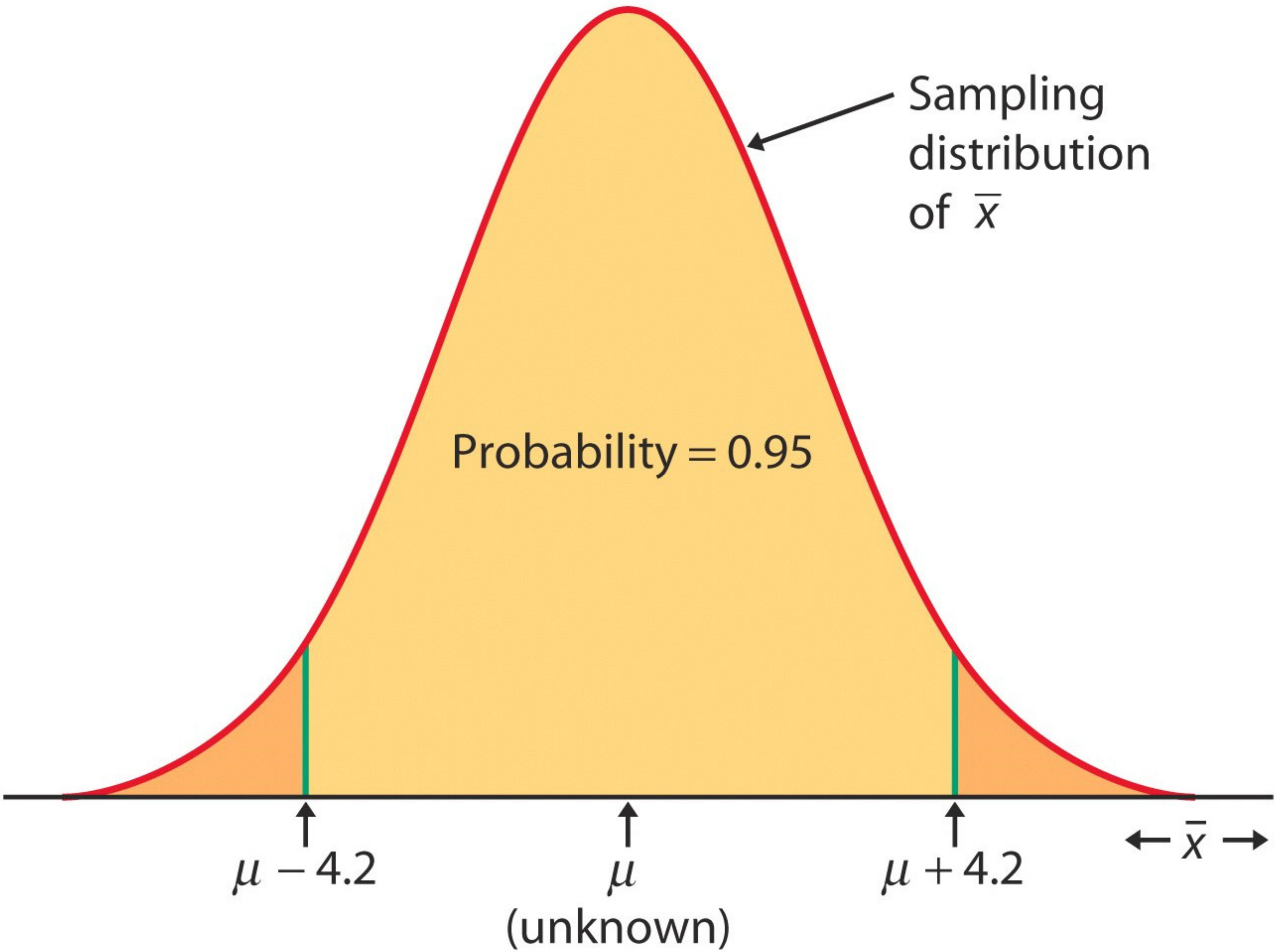
STATISTICAL INFERENCE

Statistical inference provides methods for drawing conclusions about a population from sample data.

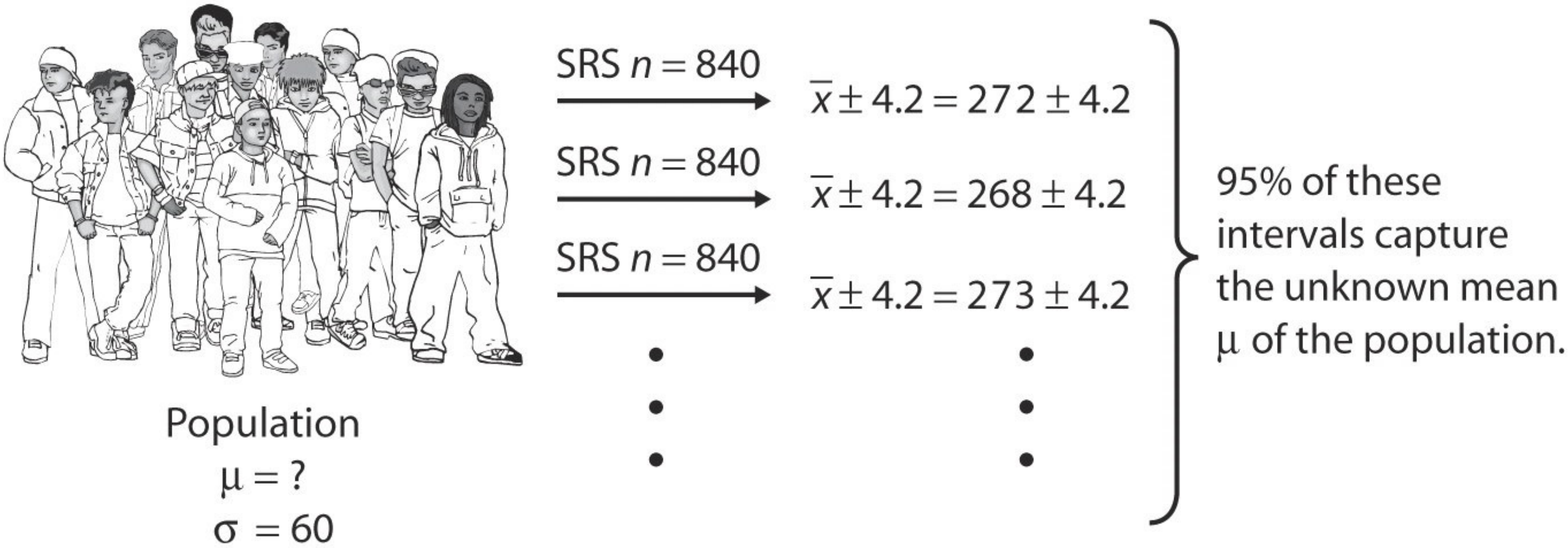
INFERENCE ABOUT A MEAN: SIMPLE CONDITIONS

1. We have an SRS from the population of interest.
2. The variable we measure has a perfectly Normal distribution $N(\mu, \sigma)$ in the population.
3. We don't know the population mean μ. Our task is to infer something about μ from the sample data. But we do know the population standard deviation σ.


SRS $n = 840$
$\bar{x} = 272$
SRS $n = 840$
$\bar{x} = 268$
SRS $n = 840$
$\bar{x} = 273$
Population
$\mu = ?$
$\sigma = 60$
Standard deviation
$\frac{\sigma}{\sqrt{n}} = 2.1$
Mean μ unknown
Values of $\bar{x}$



Sampling distribution of $\bar{x}$
Probability = 0.95
$\mu - 4.2$
μ
(unknown)
$\mu + 4.2$
$\leftarrow \bar{x} \rightarrow$


SRS $n = 840$
$\bar{x} \pm 4.2 = 272 \pm 4.2$
SRS $n = 840$
$\bar{x} \pm 4.2 = 268 \pm 4.2$
SRS $n = 840$
$\bar{x} \pm 4.2 = 273 \pm 4.2$
95% of these intervals capture the unknown mean μ of the population.
Population
$\mu = ?$
$\sigma = 60$

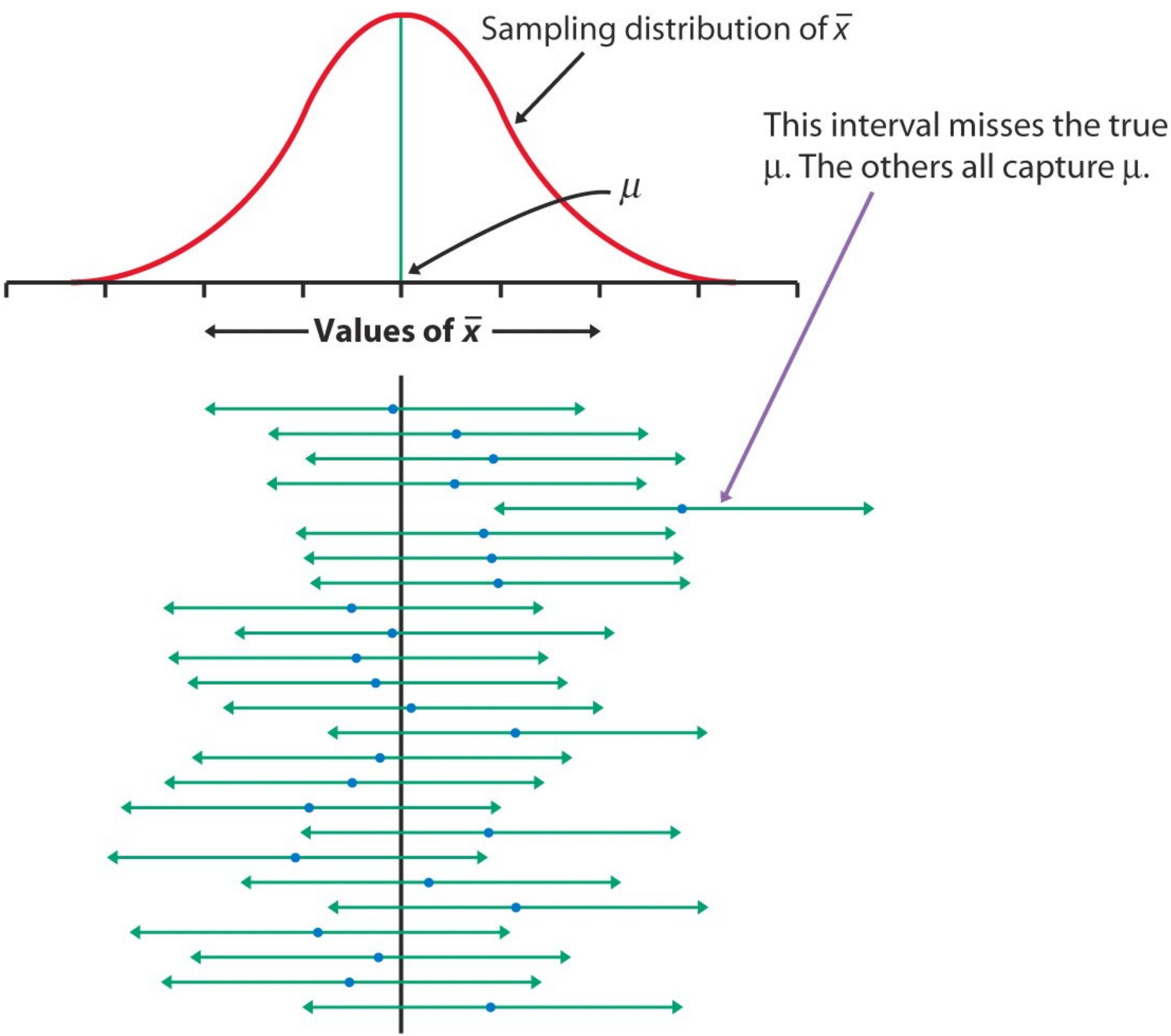
CONFIDENCE INTERVAL

A **level C confidence interval** for a parameter has two parts:

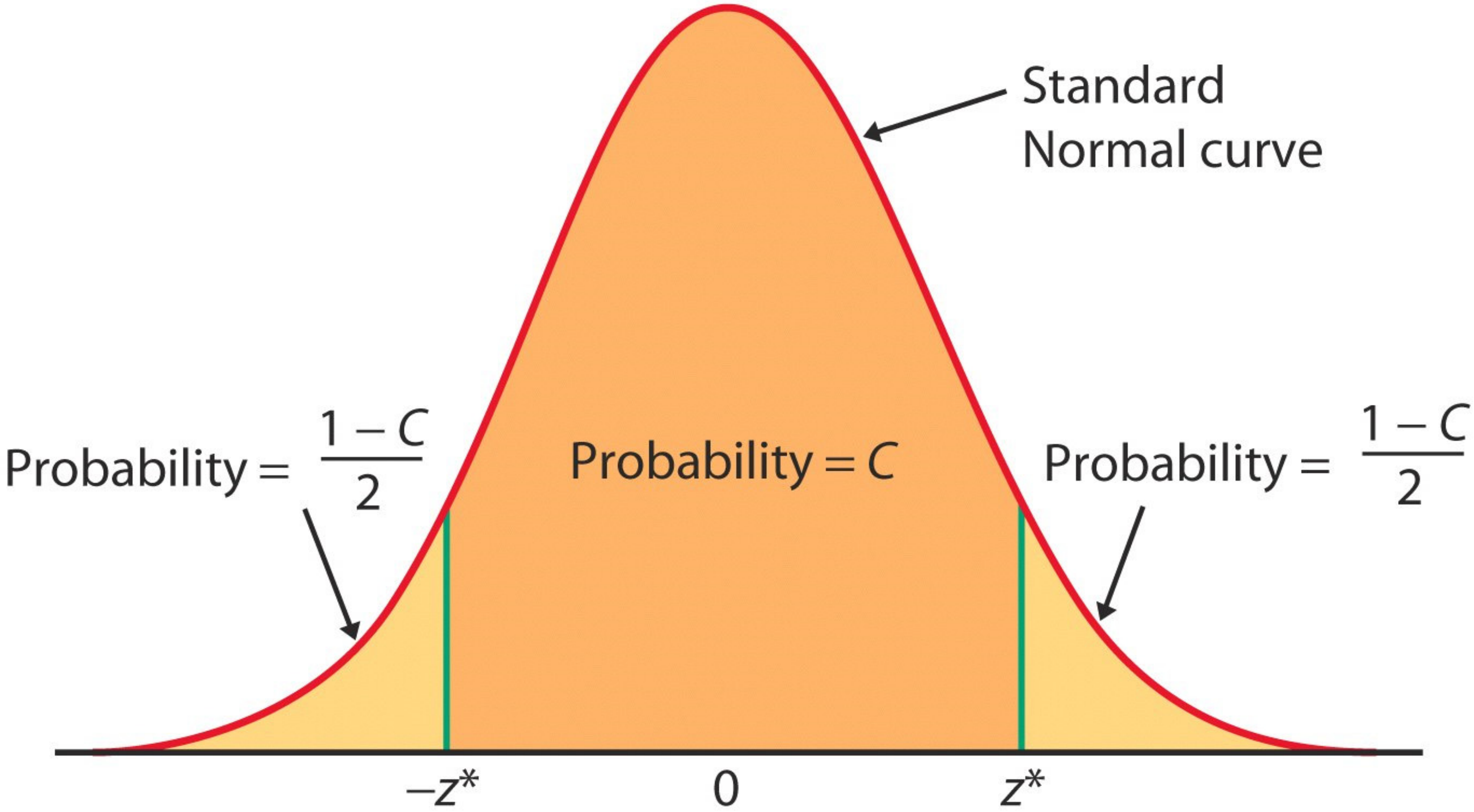
- An interval calculated from the data, usually of the form

$$\text{estimate} \pm \text{margin of error}$$

- A **confidence level** C, which gives the probability that the interval will capture the true parameter value in repeated samples. That is, the confidence level is the success rate for the method.



Sampling distribution of $\bar{x}$
This interval misses the true μ. The others all capture μ.
μ
Values of $\bar{x}$



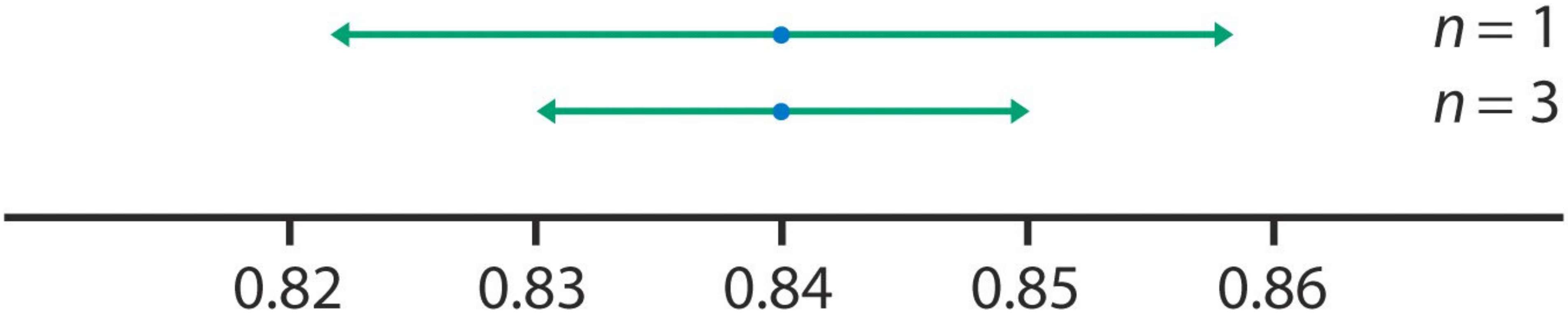
Standard
Normal curve
Probability = $\frac{1-C}{2}$
Probability = C
Probability = $\frac{1-C}{2}$
$-z^*$
0
z^*

CONFIDENCE INTERVAL FOR THE MEAN OF A NORMAL POPULATION

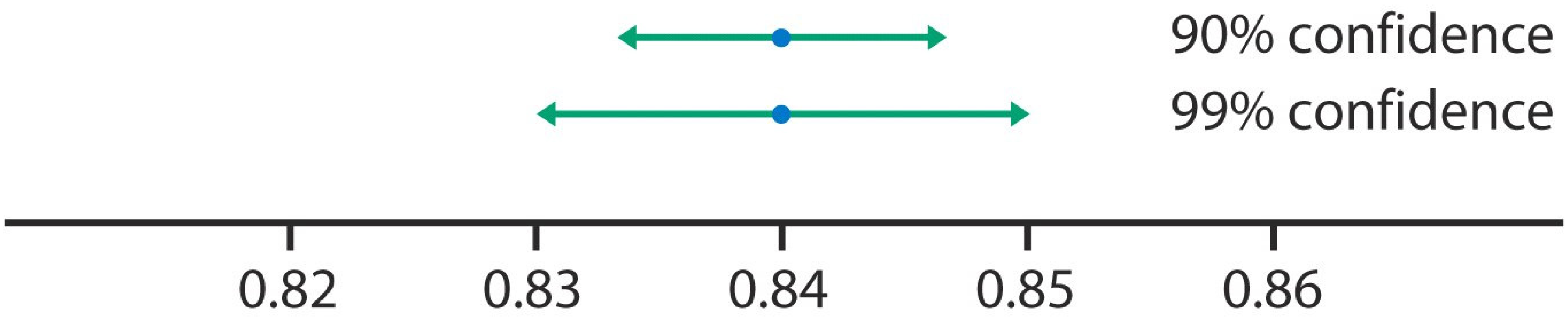
Draw an SRS of size n from a Normal population having unknown mean μ and known standard deviation σ. A level C confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

The critical value z^*



$n = 1$
$n = 3$
0.82
0.83
0.84
0.85
0.86


90% confidence
99% confidence
0.82
0.83
0.84
0.85
0.86

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for the mean of a Normal population will have a specified margin of error m when the sample size is

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

Inference about a Population Mean

CONDITIONS FOR INFERENCE ABOUT A MEAN

- Our data are a **simple random sample** (SRS) of size n from the population. This condition is very important.
- Observations from the population have a **Normal distribution** with mean μ and standard deviation σ. In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both μ and σ are unknown parameters.

STANDARD ERROR

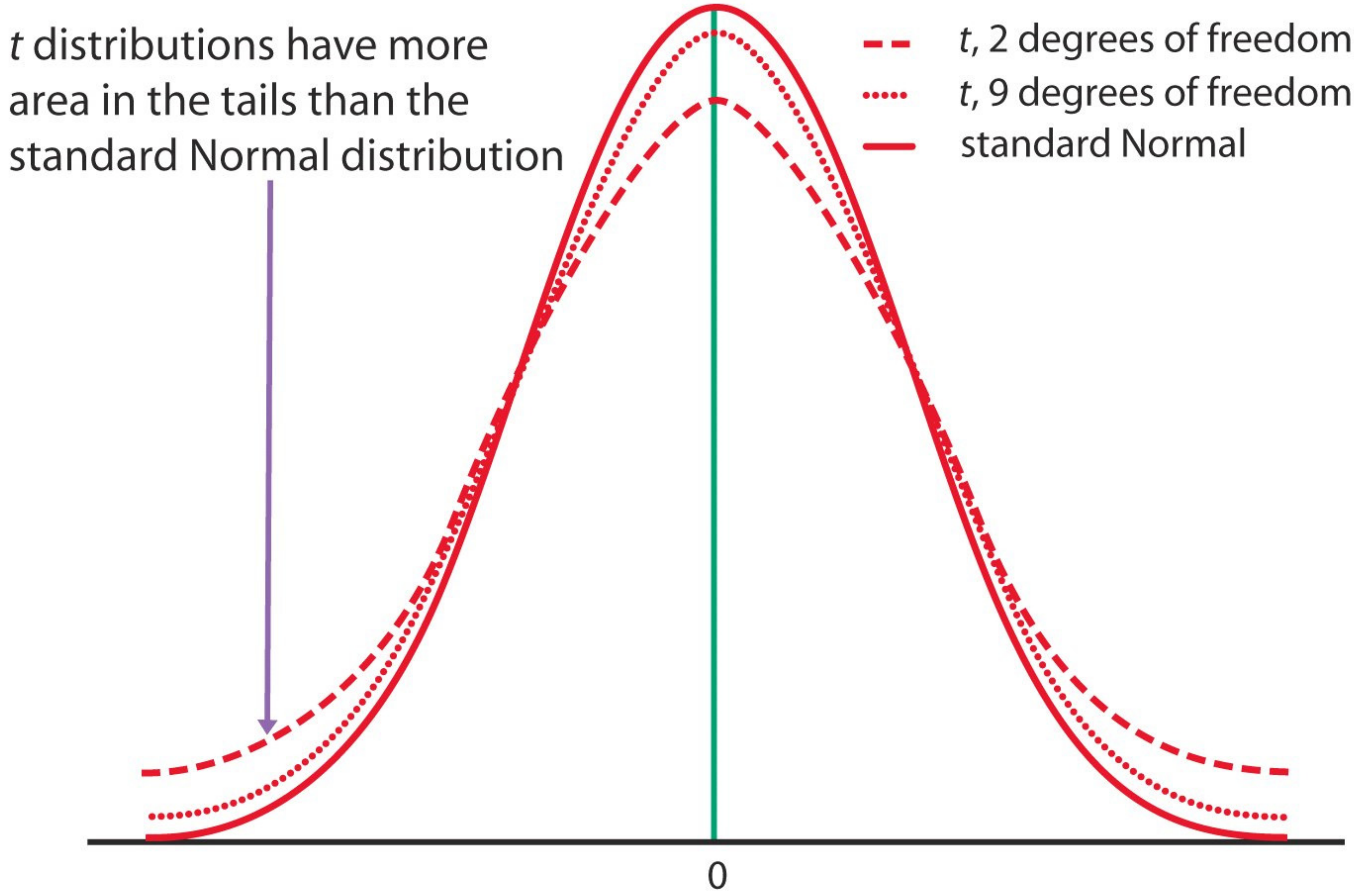
When the standard deviation of a statistic is estimated from data, the result is called the **standard error** of the statistic. The standard error of the sample mean $\overline{x}$ is $s/\sqrt{n}$.

THE ONE-SAMPLE *t* STATISTIC AND THE *t* DISTRIBUTIONS

Draw an SRS of size n from a population that has the Normal distribution with mean μ and standard deviation σ. The **one-sample *t* statistic**

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

has the ***t* distribution** with $n - 1$ degrees of freedom.



t distributions have more area in the tails than the standard Normal distribution
t, 2 degrees of freedom
t, 9 degrees of freedom
standard Normal
0

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

Draw an SRS of size n from a population having unknown mean μ. A level C confidence interval for μ is

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

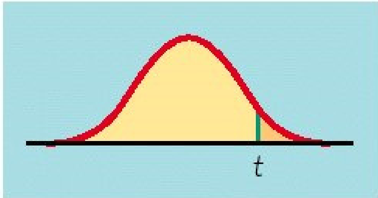
THE ONE-SAMPLE t TEST

Draw an SRS of size n from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic:

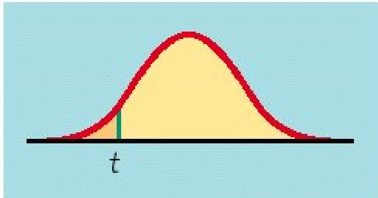
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

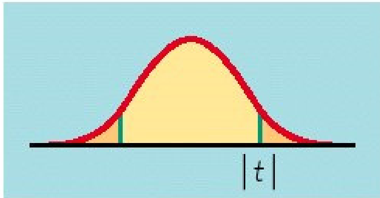
H_a: $\mu > \mu_0$ is $P(T \geq t)$



H_a: $\mu < \mu_0$ is $P(T \leq t)$



H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

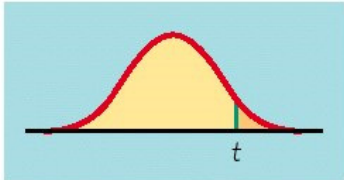
THE ONE-SAMPLE t TEST

Draw an SRS of size n from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic:

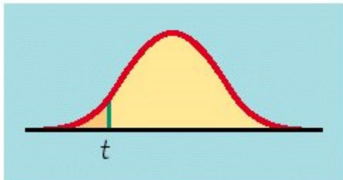
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

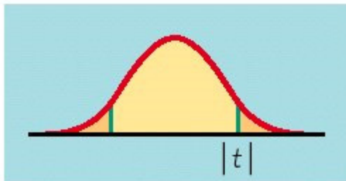
H_a: $\mu > \mu_0$ is $P(T \geq t)$



H_a: $\mu < \mu_0$ is $P(T \leq t)$



H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

ROBUST PROCEDURES

A confidence interval or significance test is called **robust** if the confidence level or *P*-value does not change very much when the conditions for use of the procedure are violated.

Recapitulation

1. Statistical inference involves **generalizing** from a **sample** to a (statistical) **universe**.
2. Statistical inference is **only possible** with **random samples**.
3. Statistical inference estimates the **probability** that a sample result could be **due to chance** (*in the selection of the sample*).
4. Sampling distributions are the keys that connect (known) **sample statistics** and (unknown) **universe parameters**.
5. **Alpha** (significance) **levels** are used to identify **critical values** on sampling distributions.