# What is Statistics?

**Definition of Statistics**

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

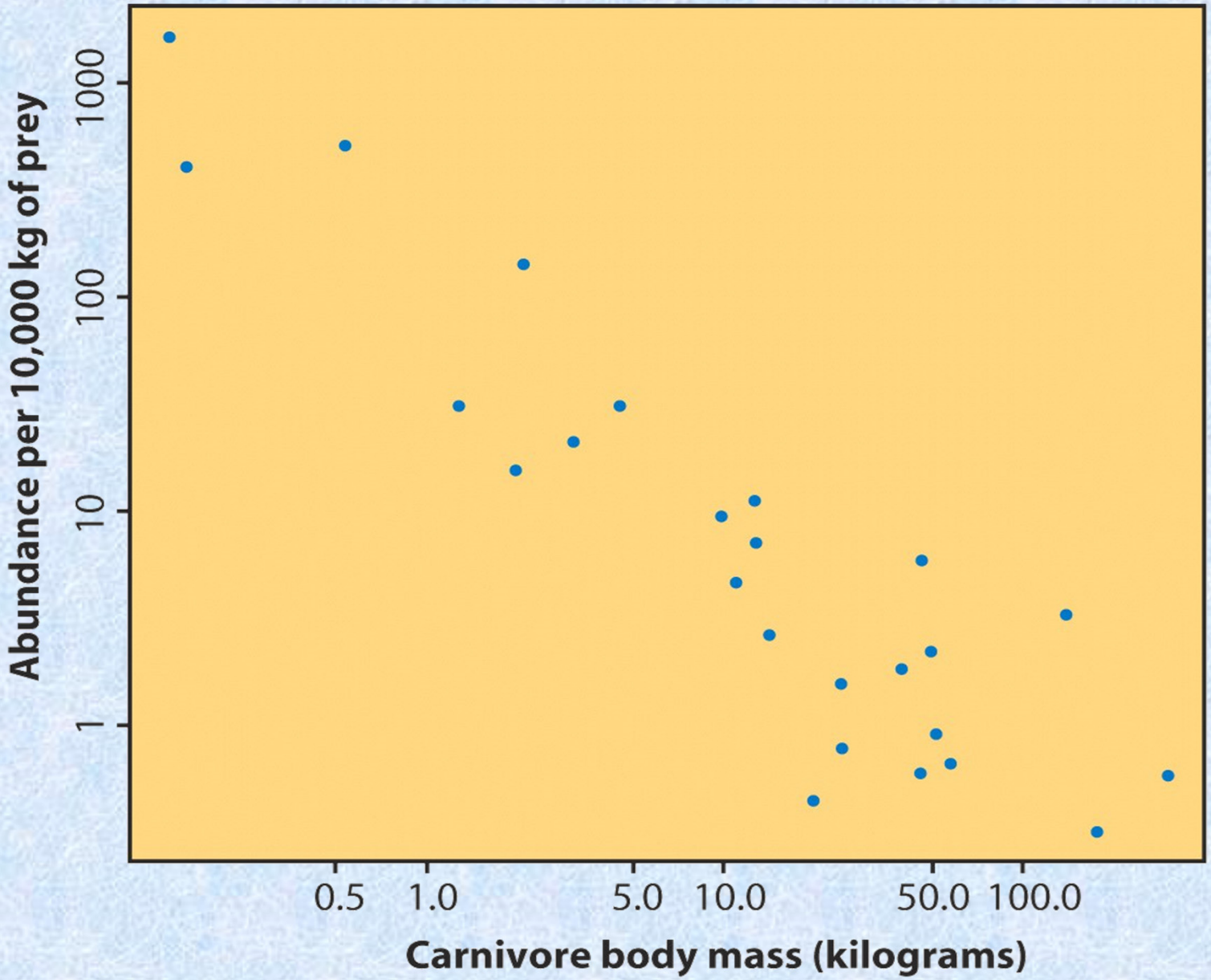
- **Branches of Statistics**
  - The study of statistics has two major branches – descriptive(exploratory) statistics and inferential statistics.
    - **Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data.
    - **Inferential statistics** 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


Abundance per 10,000 kg of prey
1000
100
10
1
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.

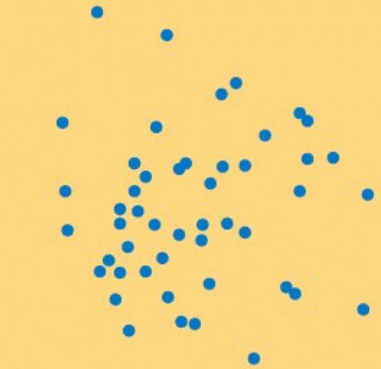
$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

# Facts about correlation

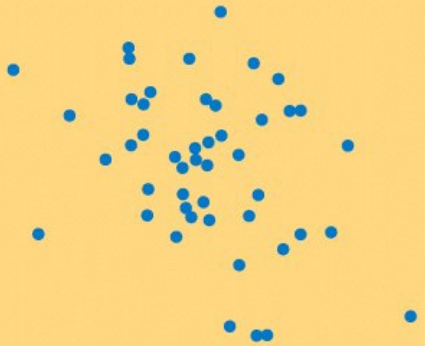
- What kind of variables do we use?
  - 1. No distinction between explanatory and response variables.
  - 2. Both variables should be quantitative
- Numerical properties
  - 1. $-1 \le r \le 1$
  - 2. r>0: positive association between variables
  - 3. r<0: negative association between variables
  - 4.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

$$\triangleleft -\textit{negative relationship}- \triangleright \triangleleft -\textit{positive relationship}- \triangleright$$
$$-1 \qquad\qquad 0 \qquad\qquad 1$$
$$\triangleleft ----\ \textit{stronger} \qquad\qquad \textit{stronger}\ ---- \triangleright$$

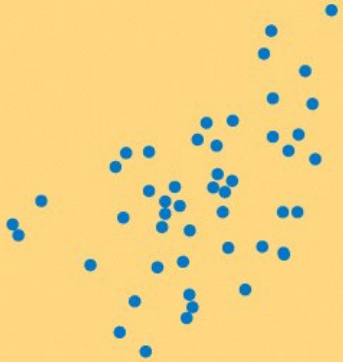
  - 6. Effected by a few outliers →not resistant.
  - 7. It doesn't describe curved relationships
  - 8. Not easy to guess the value of r from the appearance of a scatter plot



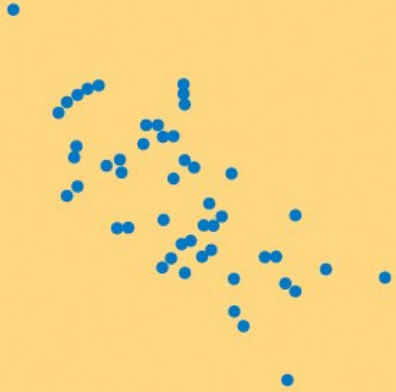
Correlation $r = 0$



Correlation $r = -0.3$

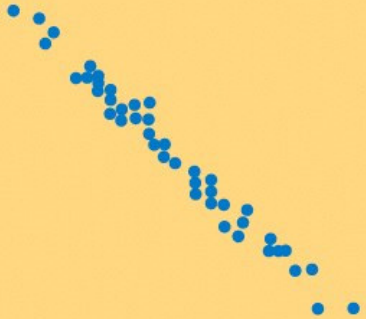


Correlation $r = 0.5$



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.**

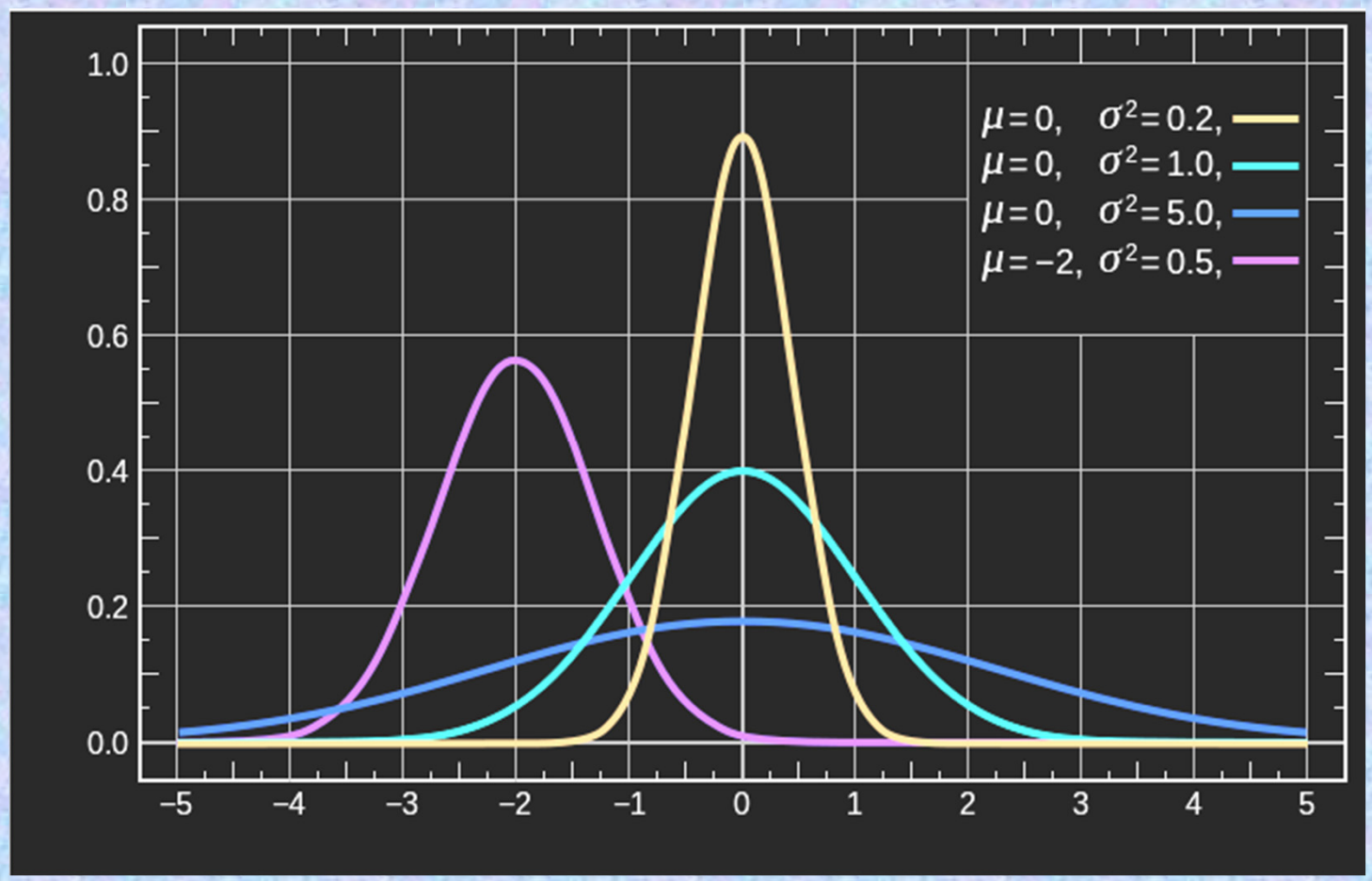
**Its also called a BELL function/curve.**

**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 $\sigma^2$.**



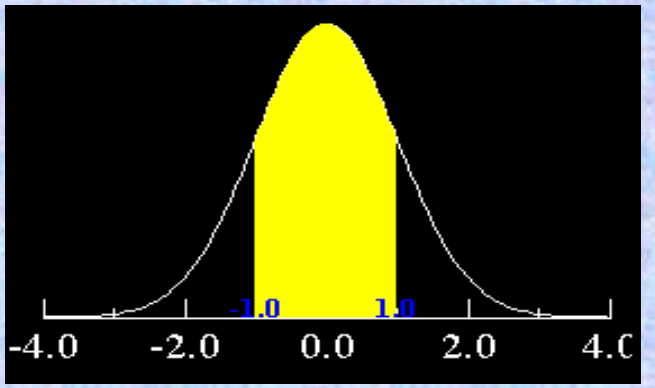

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$$

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

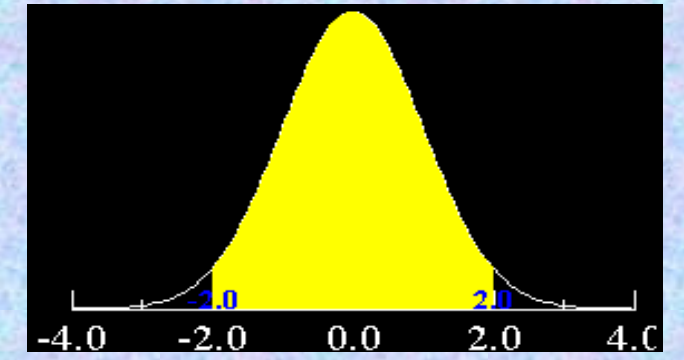
**The 68 – 95 - 99.7% Rule:**

**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 $(\mu+\sigma)$

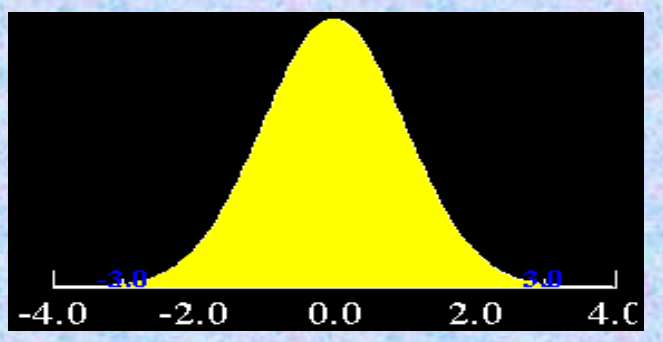


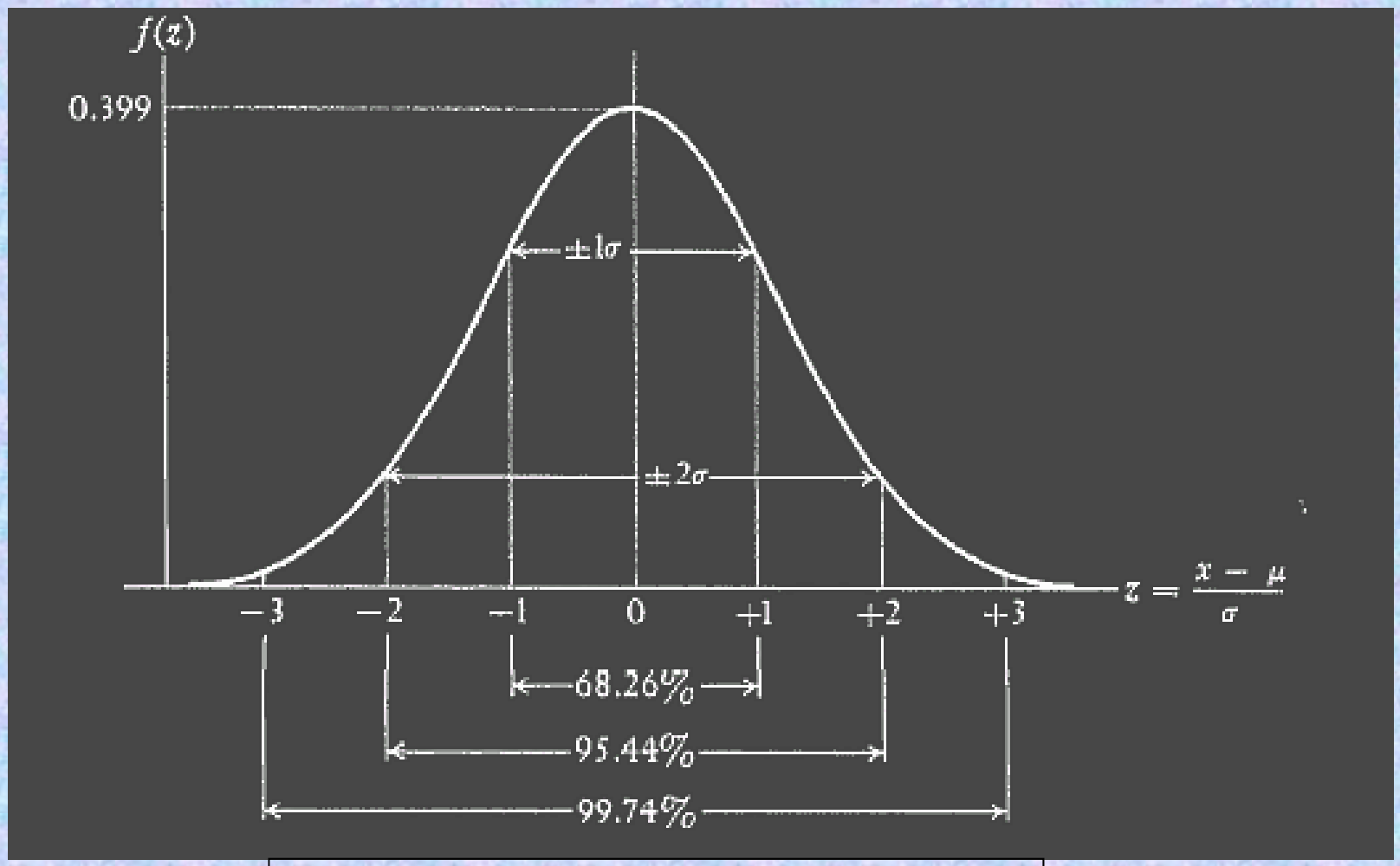

**- 95% of the observations fall within 2 standard deviations of the mean, that is, between** $(\mu-2\sigma)$ and $(\mu+2\sigma)$




**- 99.7% of the observations fall within 3 standard deviations of the mean, that is, between**

$(\mu-3\sigma)$ and $(\mu+3\sigma)$






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

**The normal distribution $p(x)$, with any mean $\mu$ and any positive deviation $\sigma$, has the following properties:**

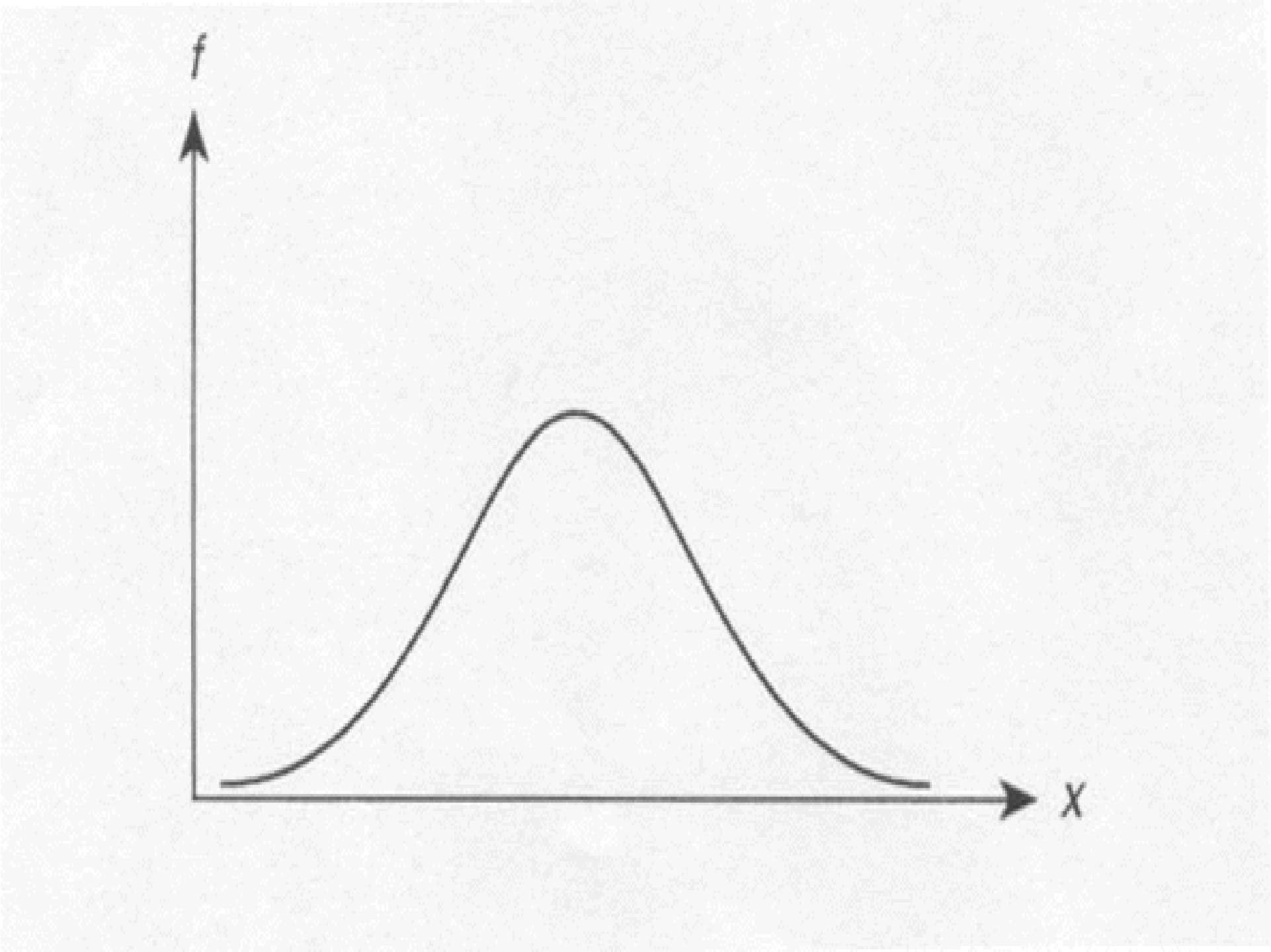
- **It is symmetric around the mean ($\mu$) of the distribution.**
- **It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$.**
- **It has two inflection points (where the second derivative of $f$ is zero and changes sign), located one standard 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 $n$.**


f
X

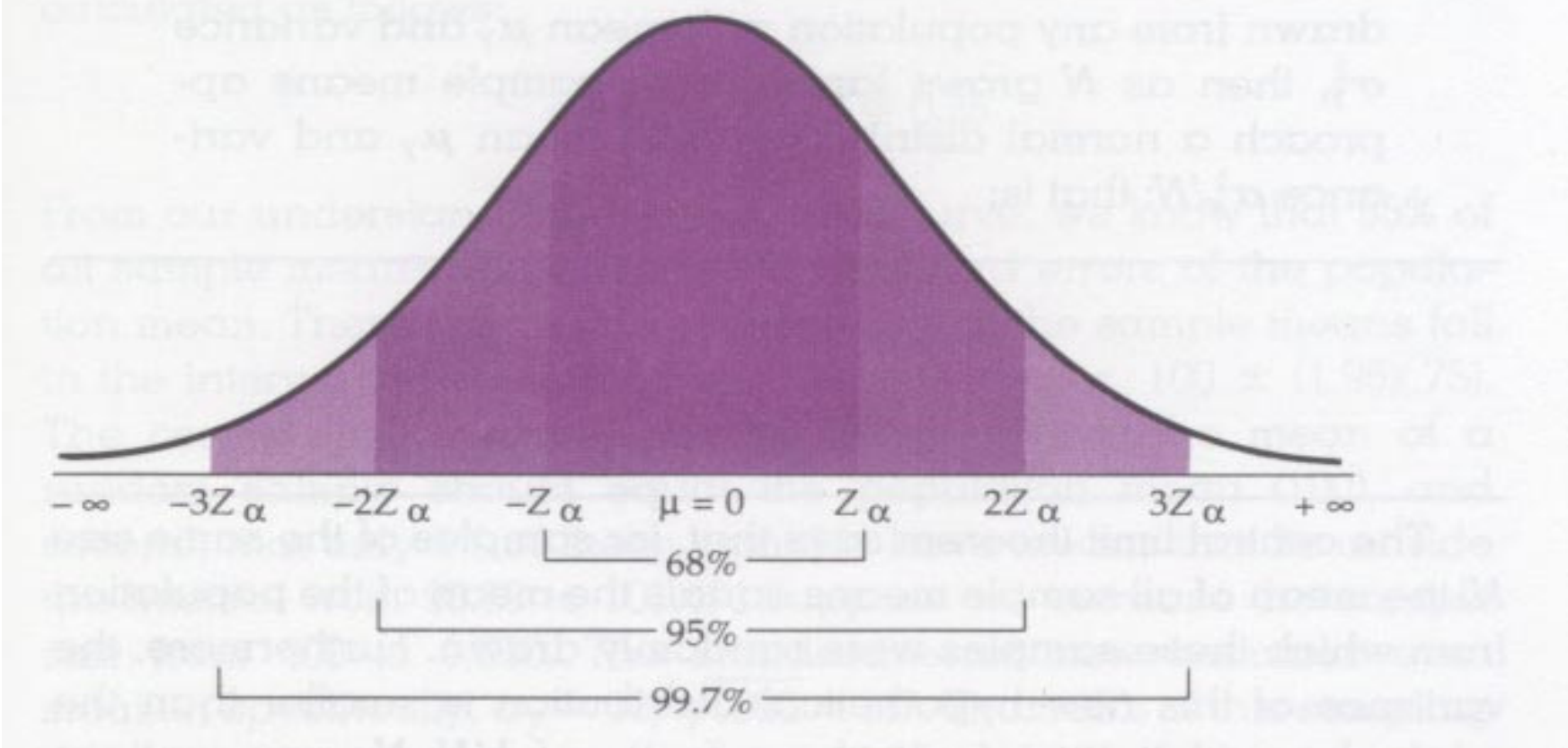
Mathematically,

$$p(Y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}}\,\varepsilon^{-(Y-\mu_Y)^2/2\sigma_Y^2}$$

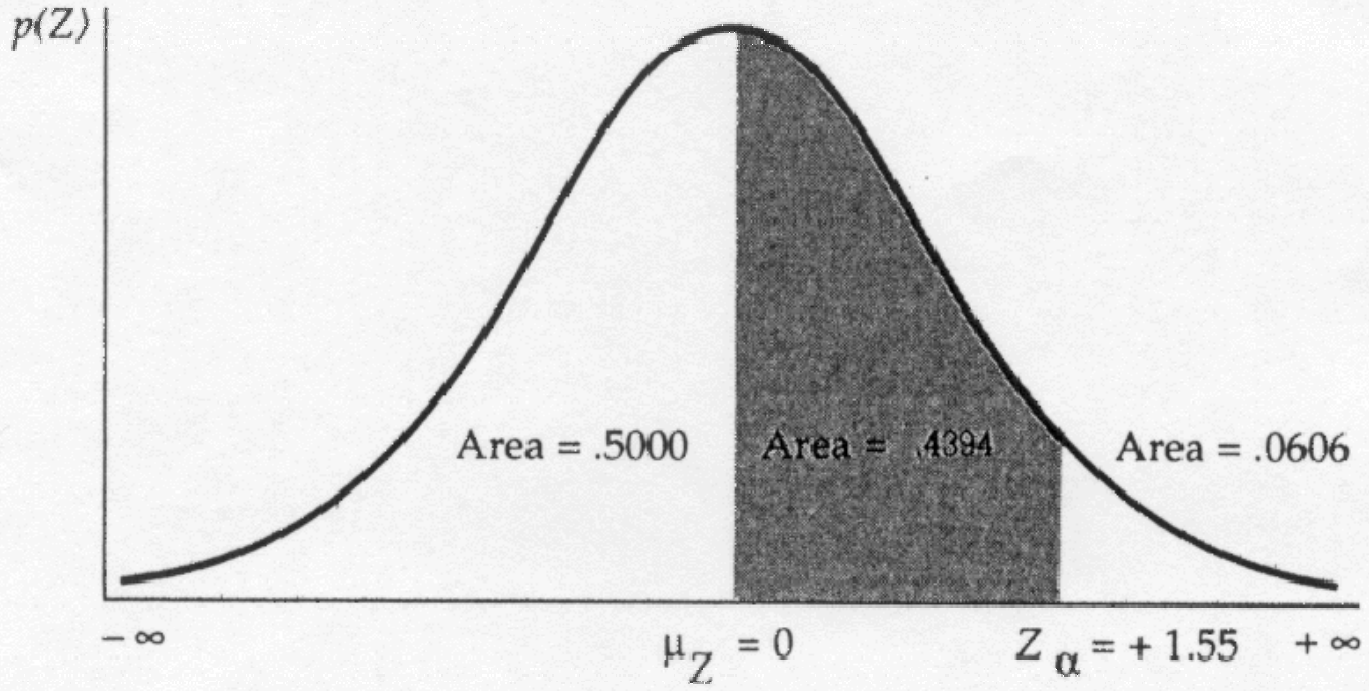
A normal distribution:

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



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

**Bivariate Normal Density:**

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

$\mu$ - Mean; $\sigma$ - S.D.; $\rho_{xy}$ - Correlation Coefficient

**Visualize ρ as equivalent to the orientation of the 2-D Gabor 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 $k^{th}$ moment is defined as:**

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

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

**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 $\sigma_x$.**

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

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

*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).**

**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.**

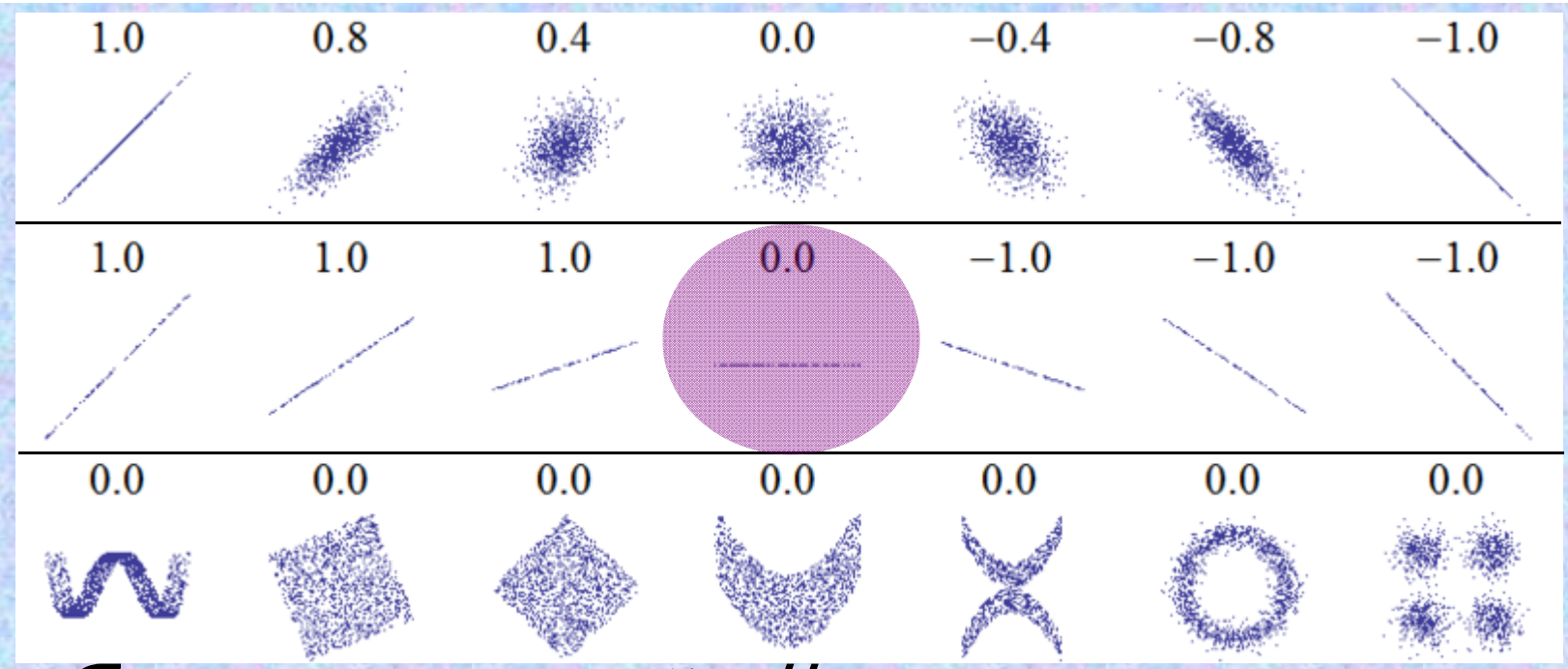
**Correlation between x and y:** $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E[(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})]$

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

**For Z = ax + by ;**

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

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




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

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

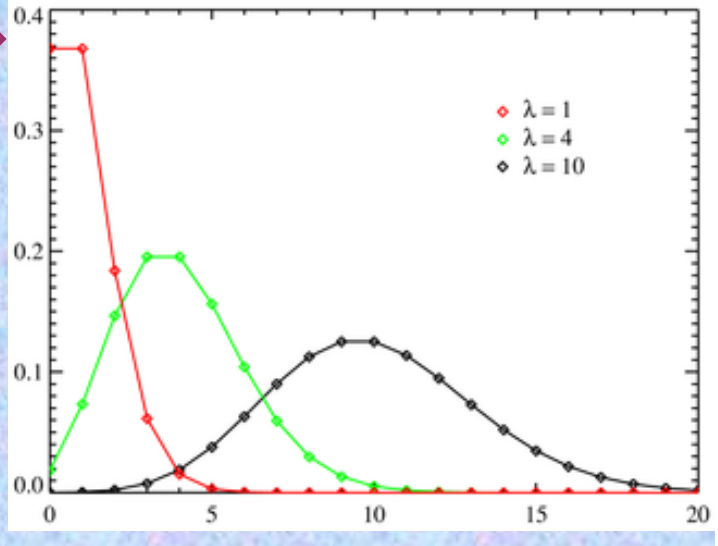
**The correlation coefficient can also be viewed as the cosine of the angle between the two vectors ($\mathbb{R}^D$) of samples drawn from the two random variables.**

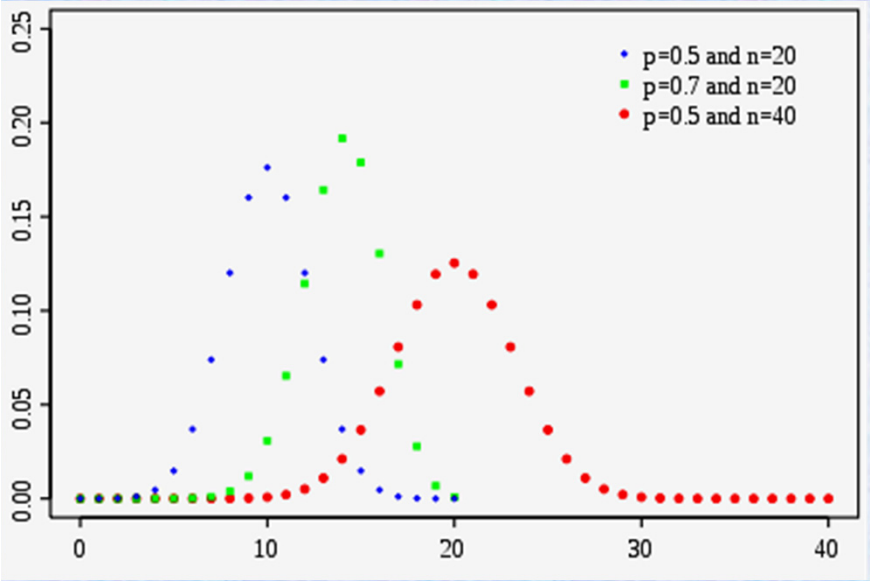
**This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero.**

## Other PDFs:

Poisson →

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



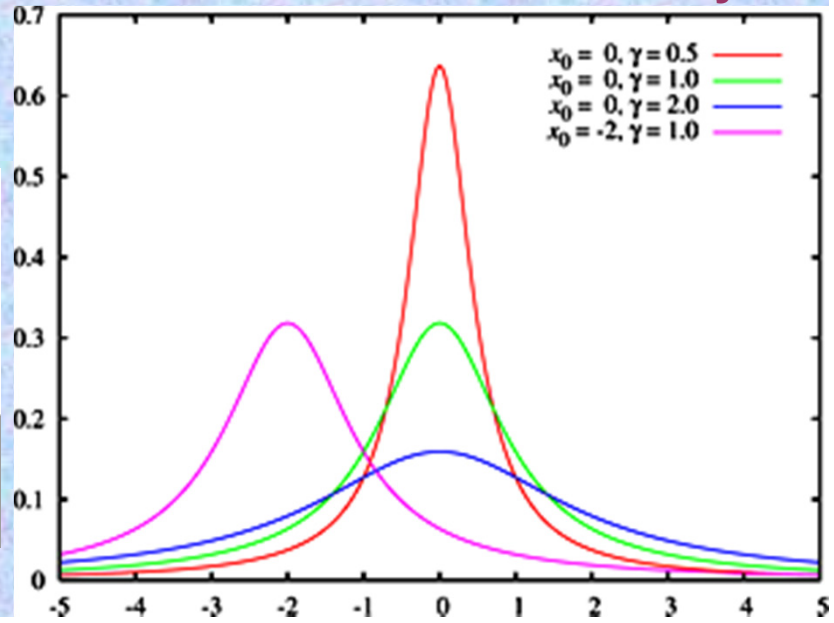

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

← Binomial

Cauchy

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

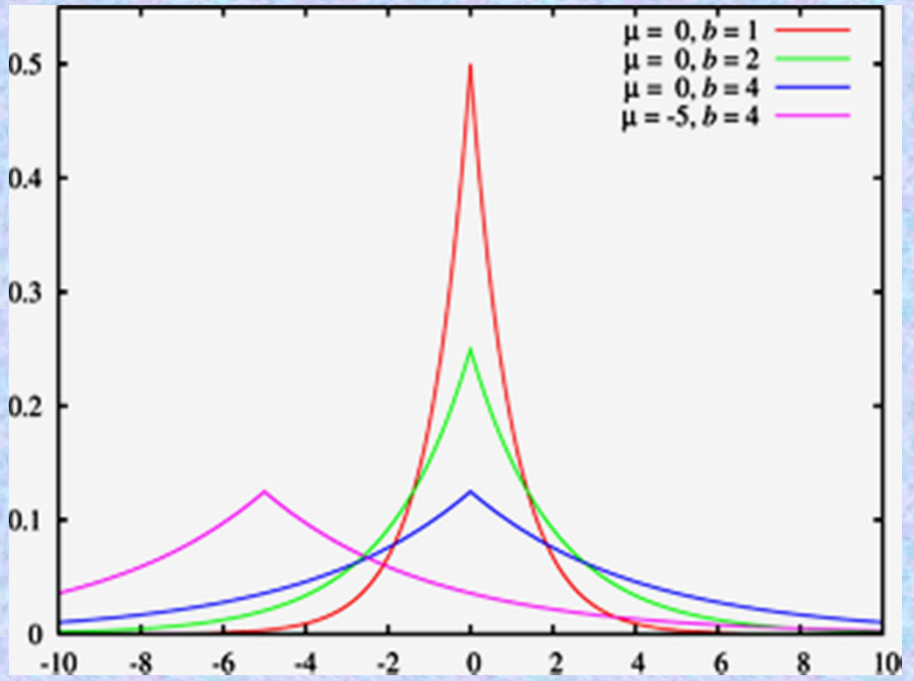
$$= \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$



## LAPLACE:

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

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




## Double Exponential Density:

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

Read about:

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

# 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:**

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

# 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

**What are the main types of sampling 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: $k = \mathrm{N/n}$ 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**.

# 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.

# 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, $\bar{X}$, is the most common estimator of the population mean, $\mu$.
- The sample variance, $s^2$, is the most common estimator of the population variance, $\sigma^2$.
- The sample standard deviation, $s$, is the most common estimator of the population standard deviation, $\sigma$.
- 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 |

$E(X) = 1(1/6) + 2(1/6) + 3(1/6) + \ldots\ldots\ldots\ldots\ldots\ldots = 3.5$

$V(X) = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots\ldots\ldots\ldots\ldots = 2.92$

# Throwing a dice twice – sampling distribution of sample mean

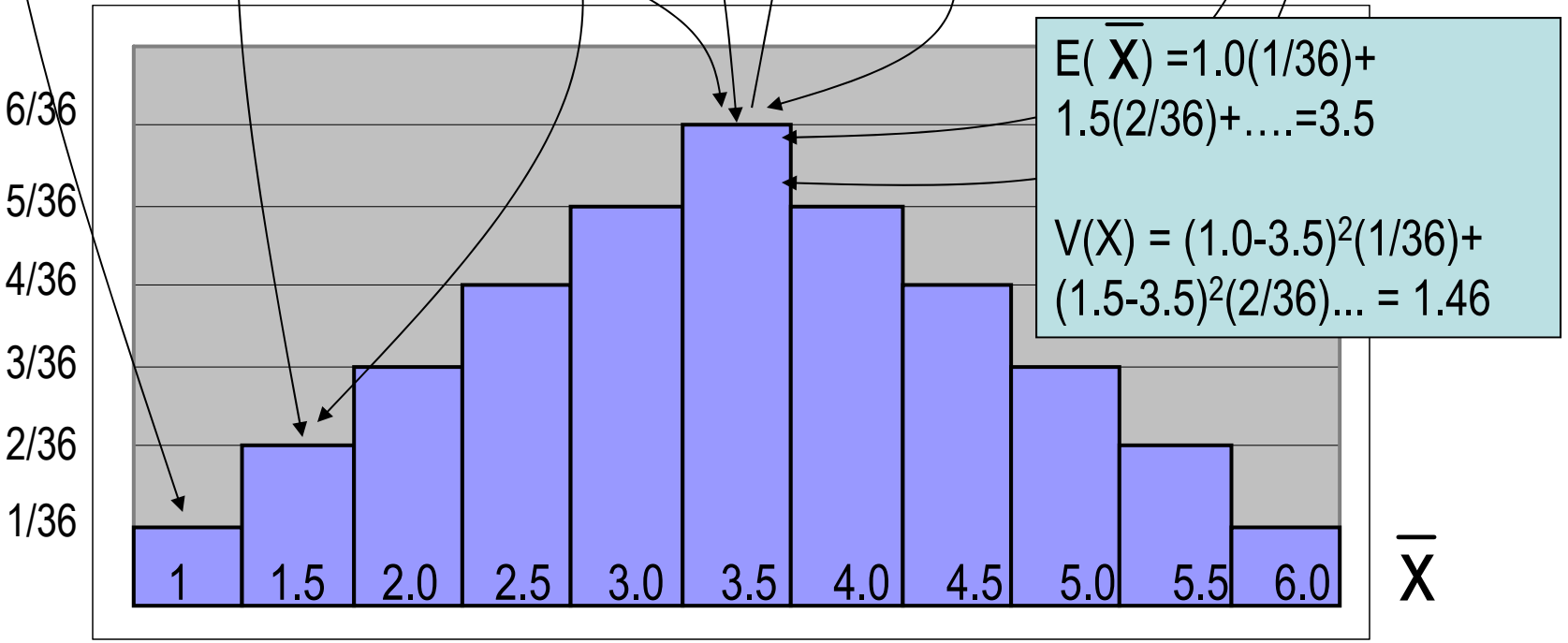
- Suppose we want to estimate $\mu$ from the mean $\bar{x}$ of a sample of size n = 2.
- What is the distribution of $\bar{x}$ ?

# 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 |
| 4 | | | | | | | | .5 |
| 5 | | | | | | | | 5 |
| 6 | | | | | | | | .5 |
| 7 | | | | | | | | .5 |
| 8 | | | | | | | | 4 |
| 9 | | | | | | | | .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

$n = 5$

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

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

$n = 10$

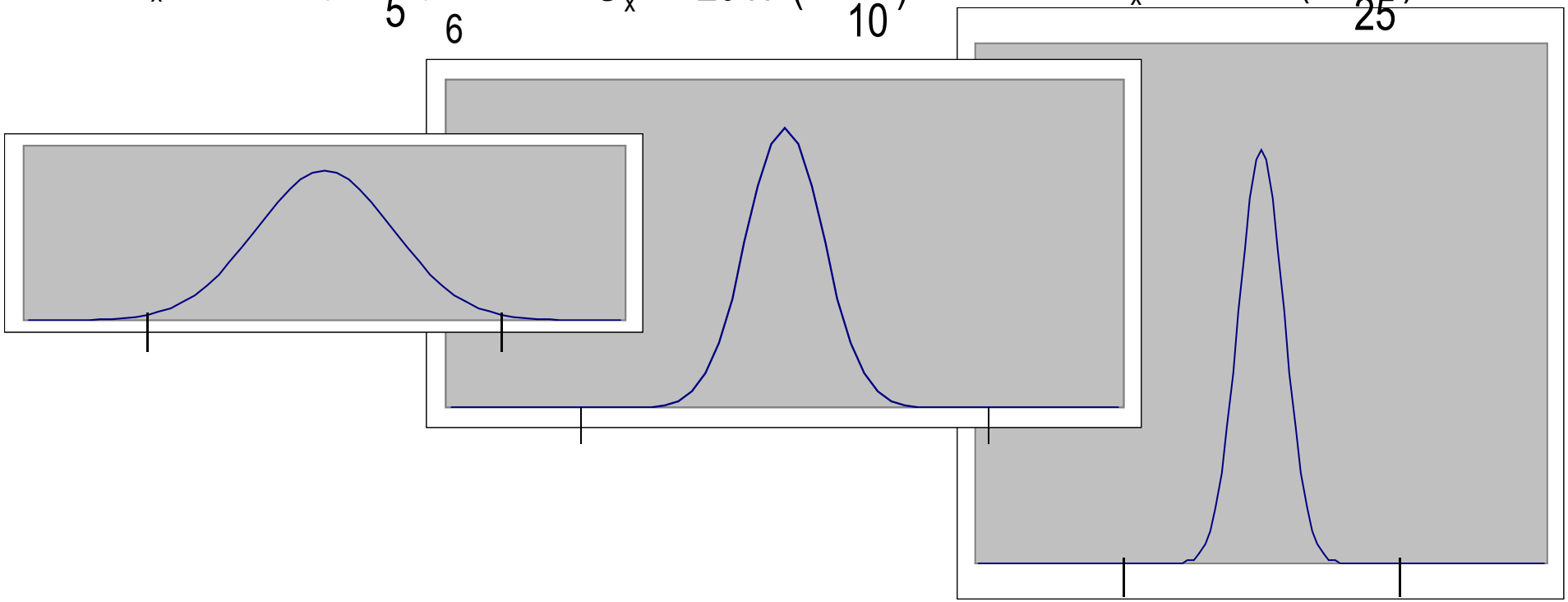
$\mu_{\bar{x}} = 3.5$

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

$n = 25$

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

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



# Sampling Distribution of the Mean

$n = 5$

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

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

$n = 10$

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

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

$n = 25$

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

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

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

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

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

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

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:

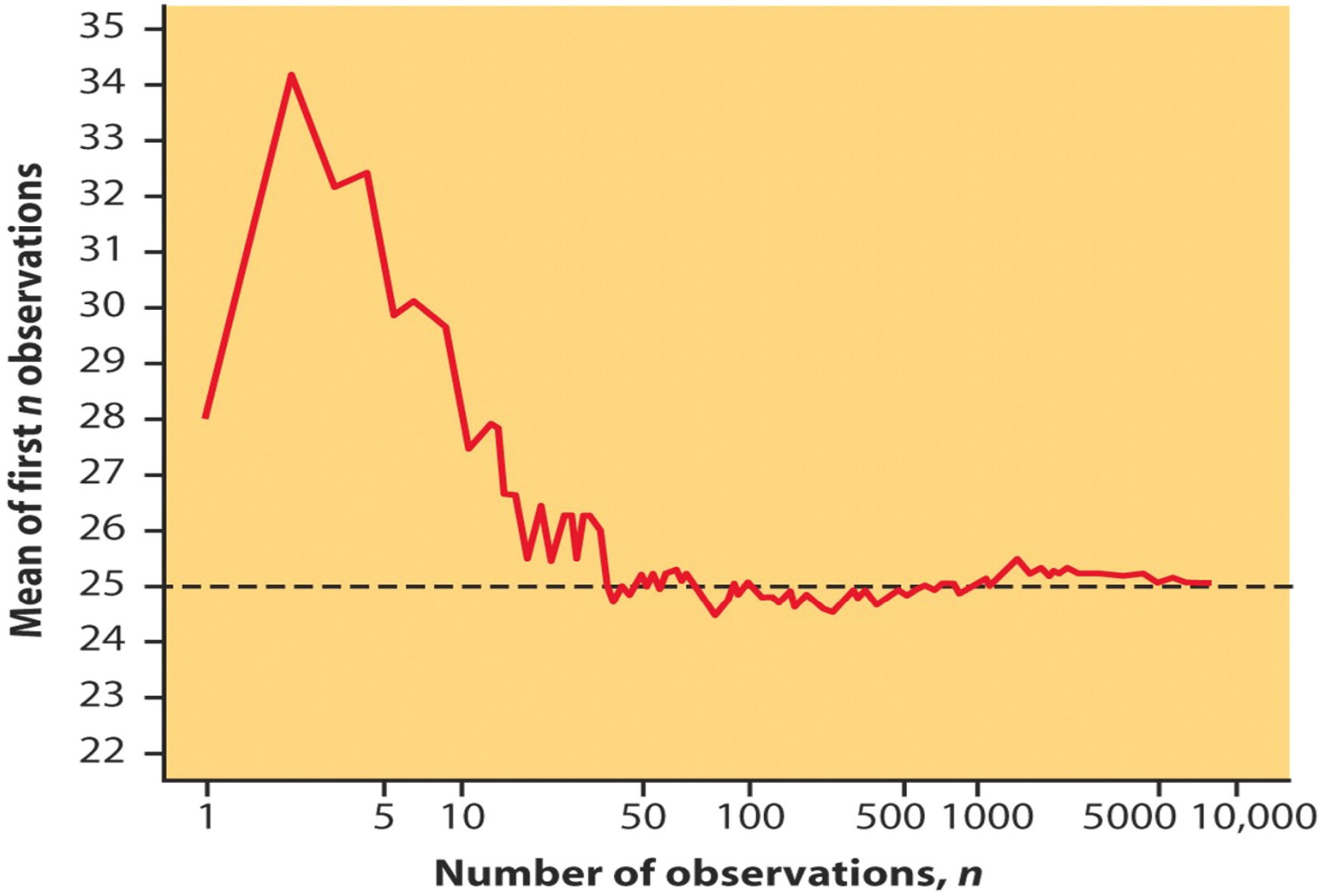
$$\text{s.e.} = SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

# Law of Large Number

**LAW OF LARGE NUMBERS**

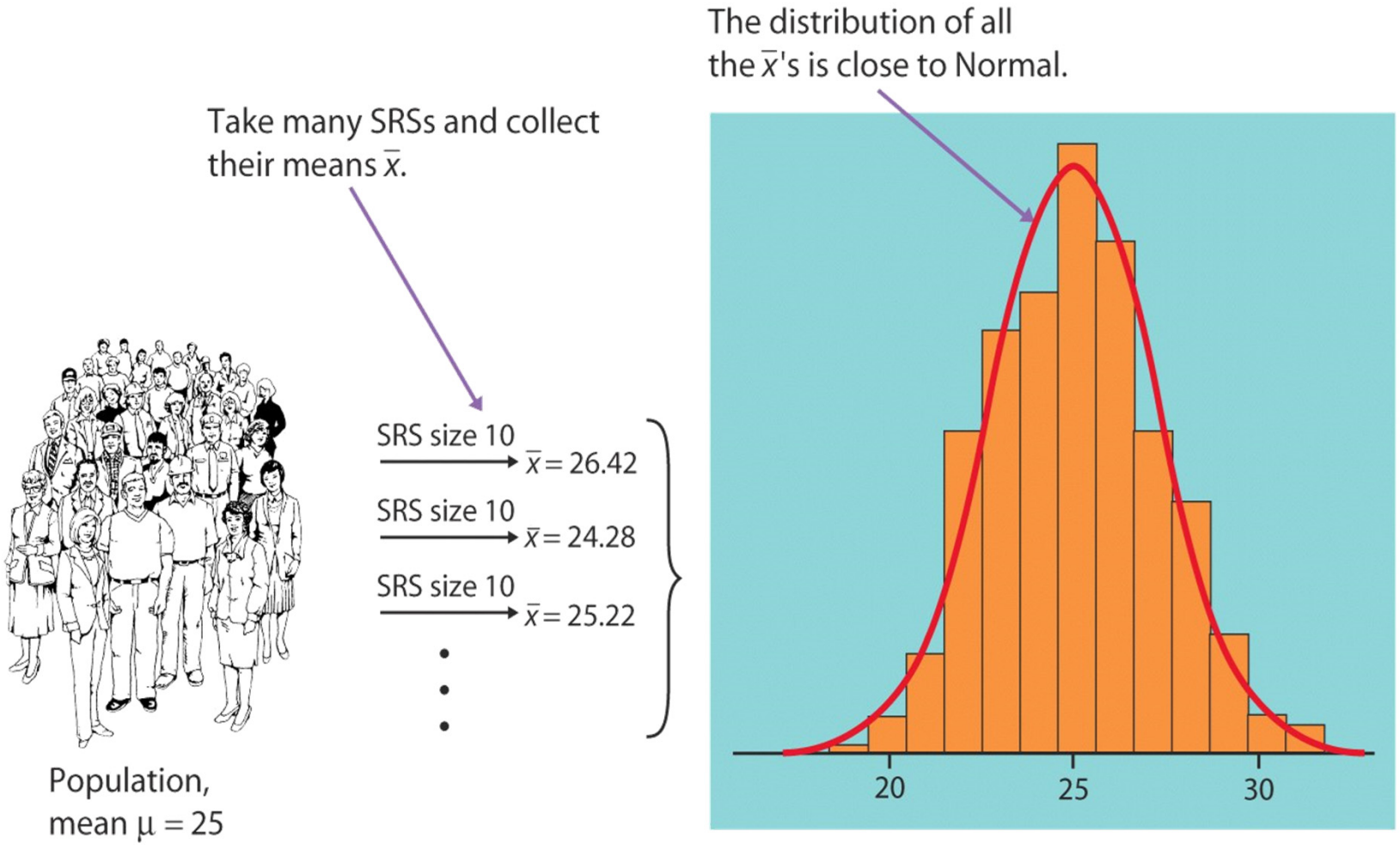
Draw observations at random from any population with finite mean $\mu$. As the number of observations drawn increases, the mean $\overline{x}$ of the observed values gets closer and closer to the mean $\mu$ of the population.

How sample means approach the population mean ($\mu$=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 $\bar{x}$

## First Property: The Mean

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

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

# 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.

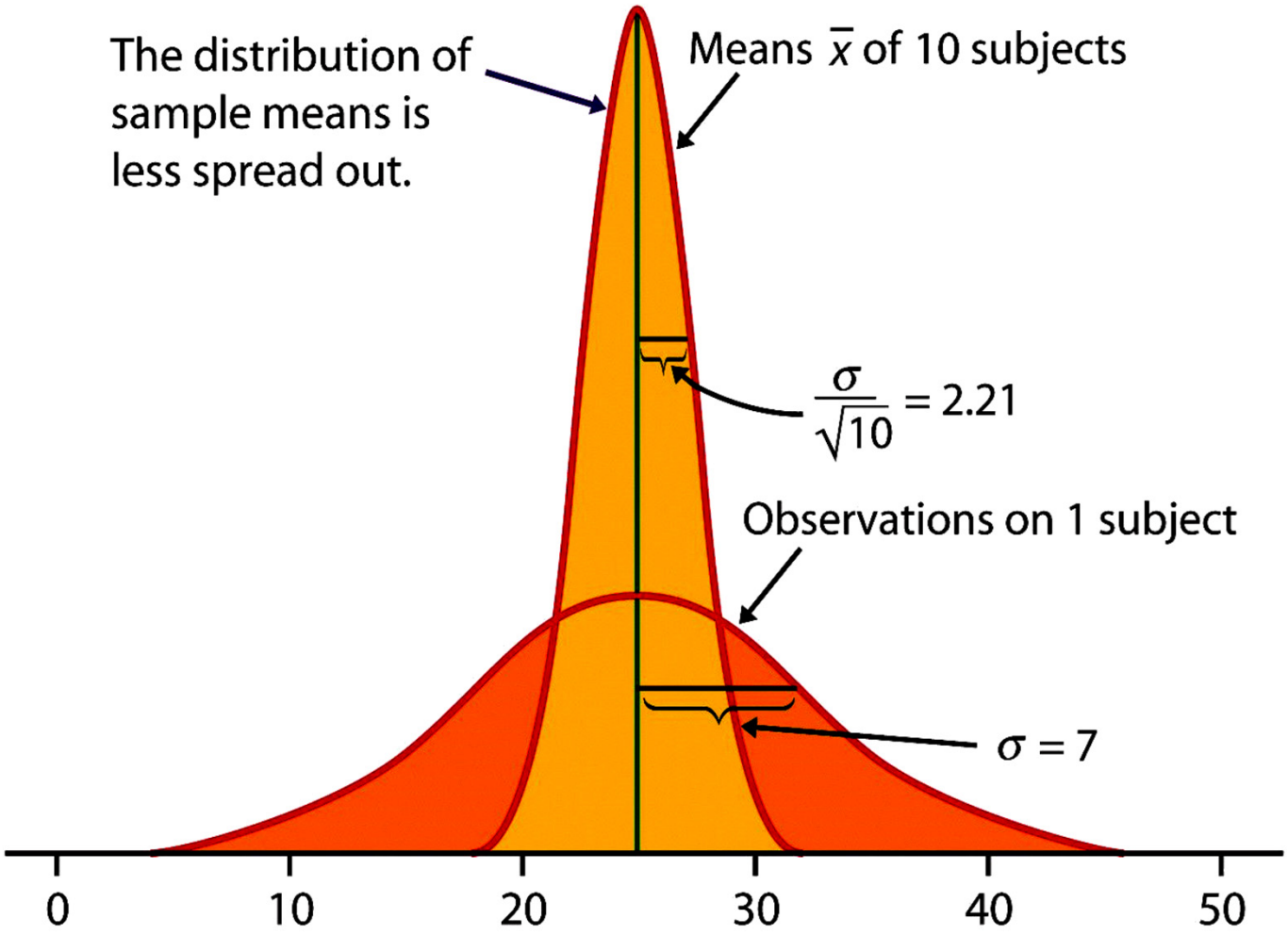
# Sampling distribution of the sample mean

- 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 $\bar{x}$ 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 $\mu$ and finite standard deviation $\sigma$. 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.

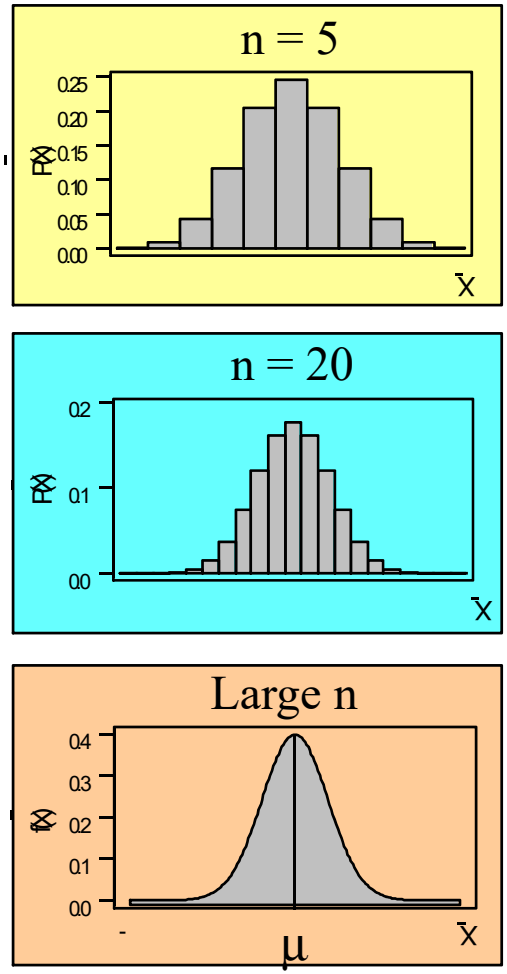
# Sampling distribution of the sample mean

- 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 $\bar{x}$ follows approximately Normal distribution as follows **N(μ, σ/√n ).**

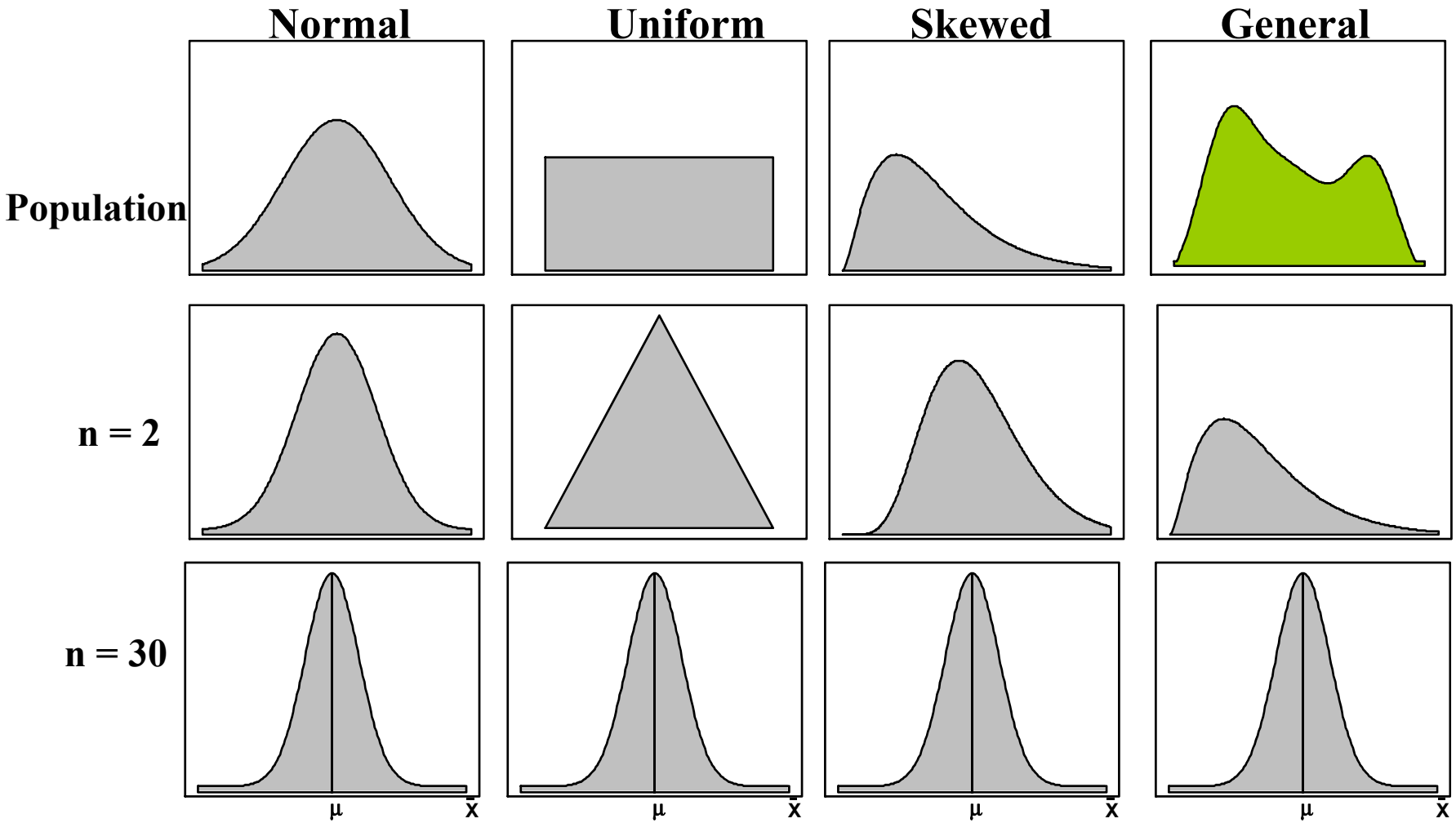
# The Central Limit Theorem

When sampling from a population with mean $\mu$ and finite standard deviation $\sigma$, the sampling distribution of the sample mean will tend to be a normal distribution with mean $\mu$ 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



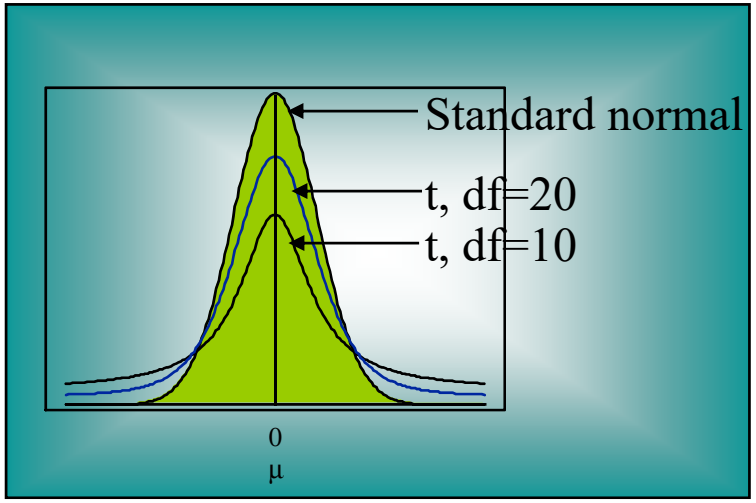
# Student's *t* Distribution

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

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

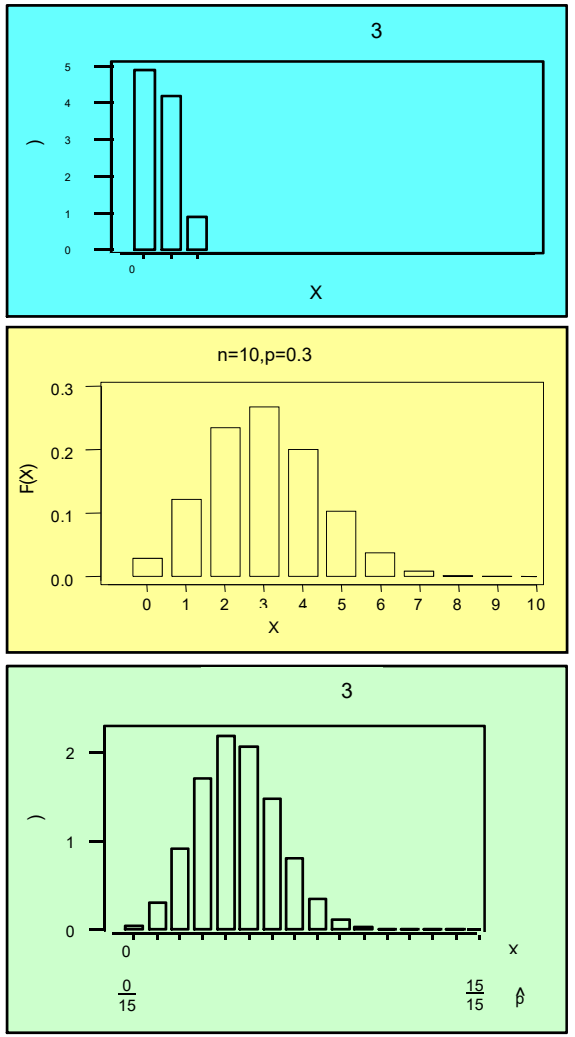
- A finite population is treated as being infinite if $n/N \leq .05$.
- $\sqrt{(N-n)/(N-1)}$ is the finite correction factor.
- $\sigma_{\bar{x}}$ is referred to as the standard error of the mean.

# 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, $n$.

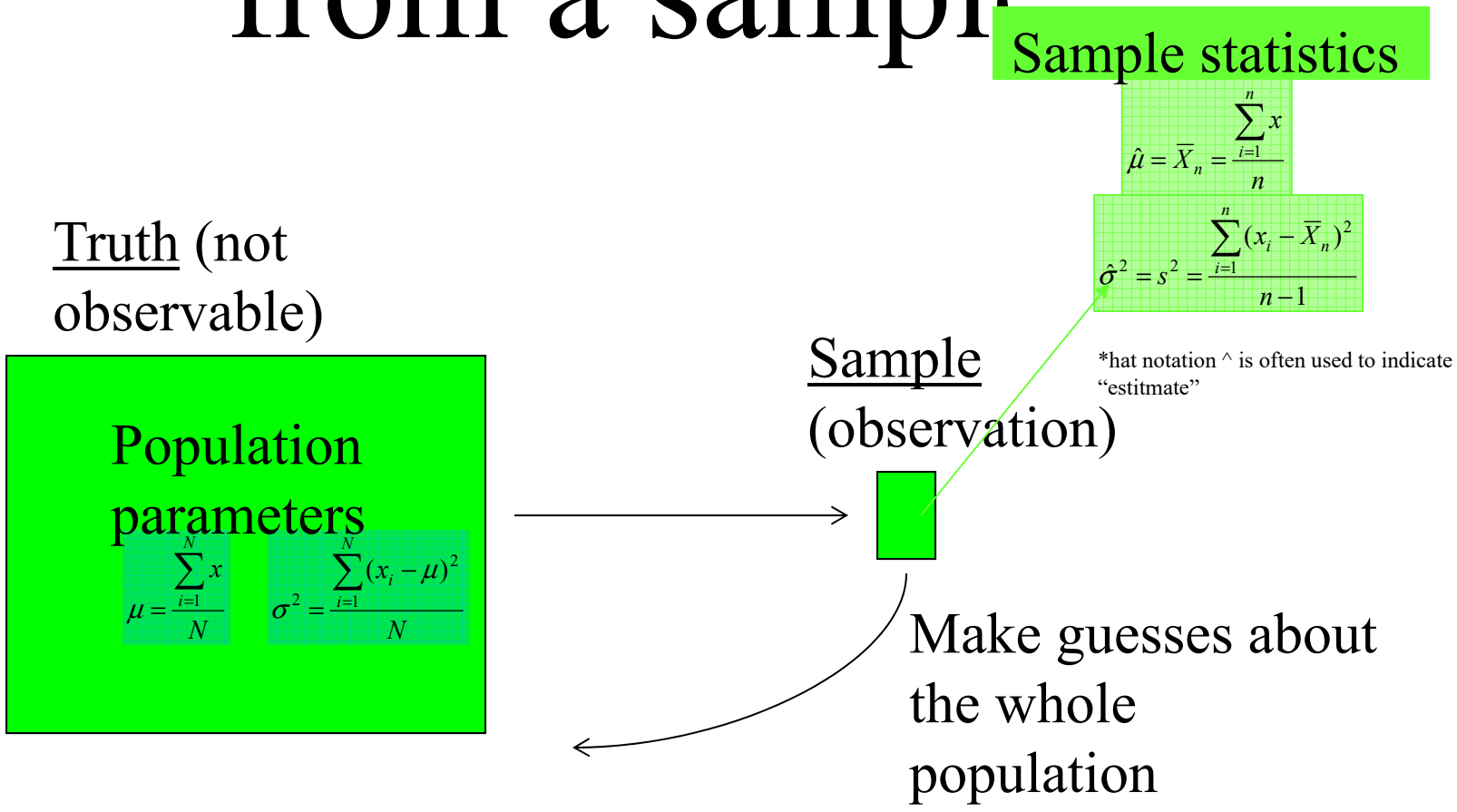
Sample proportion: $\hat{p} = \frac{X}{n}$

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



# Statistics vs. Parameters

- **Sample Statistic** – 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

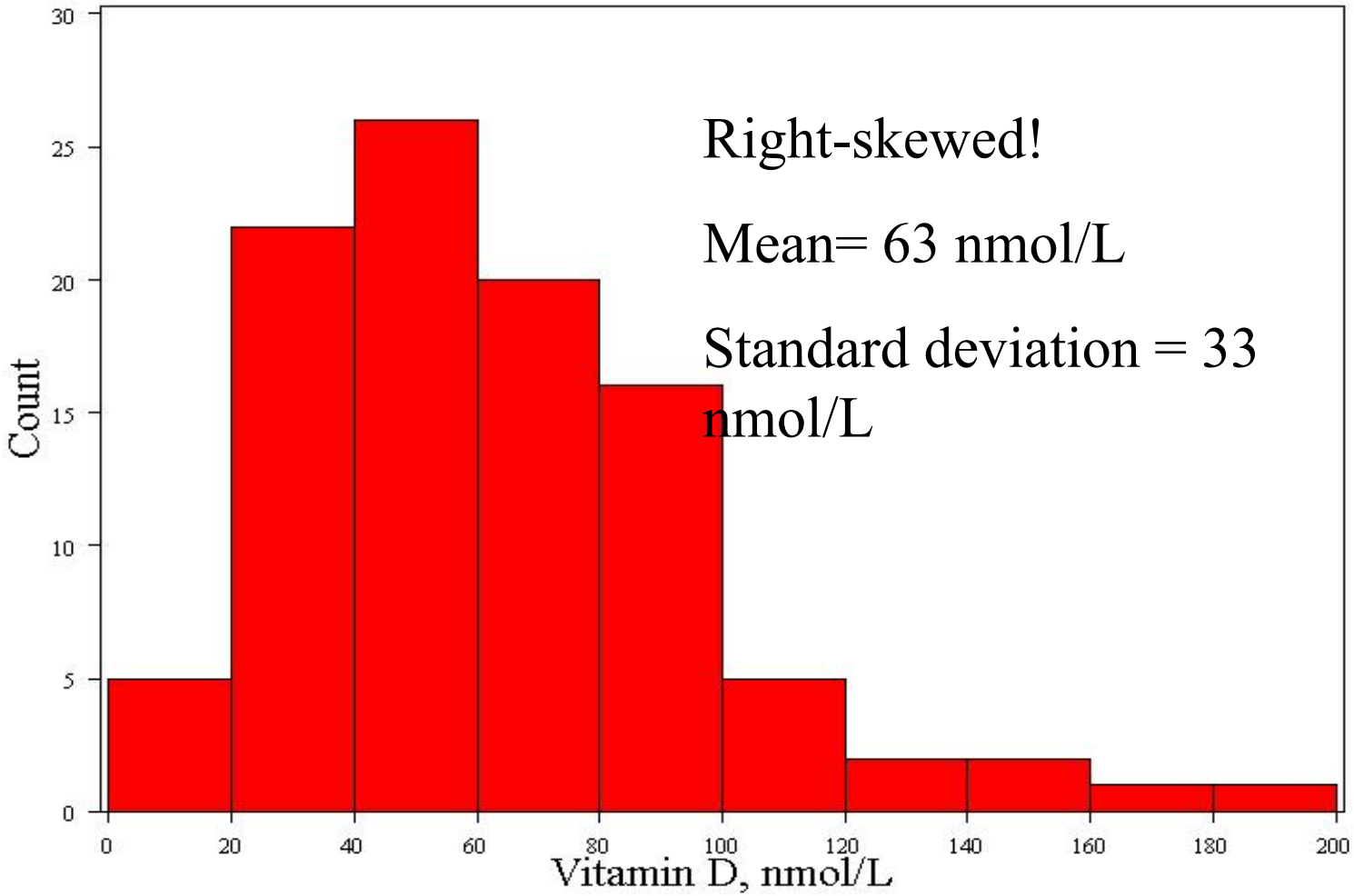
- **Population parameter** – the true value/true effect in the entire population of interest
  - E.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L
  - E.g., the true correlation between vitamin D and cognitive function in all middle-aged and older European men is 0.15

# Example 1: cognitive function and vitamin D

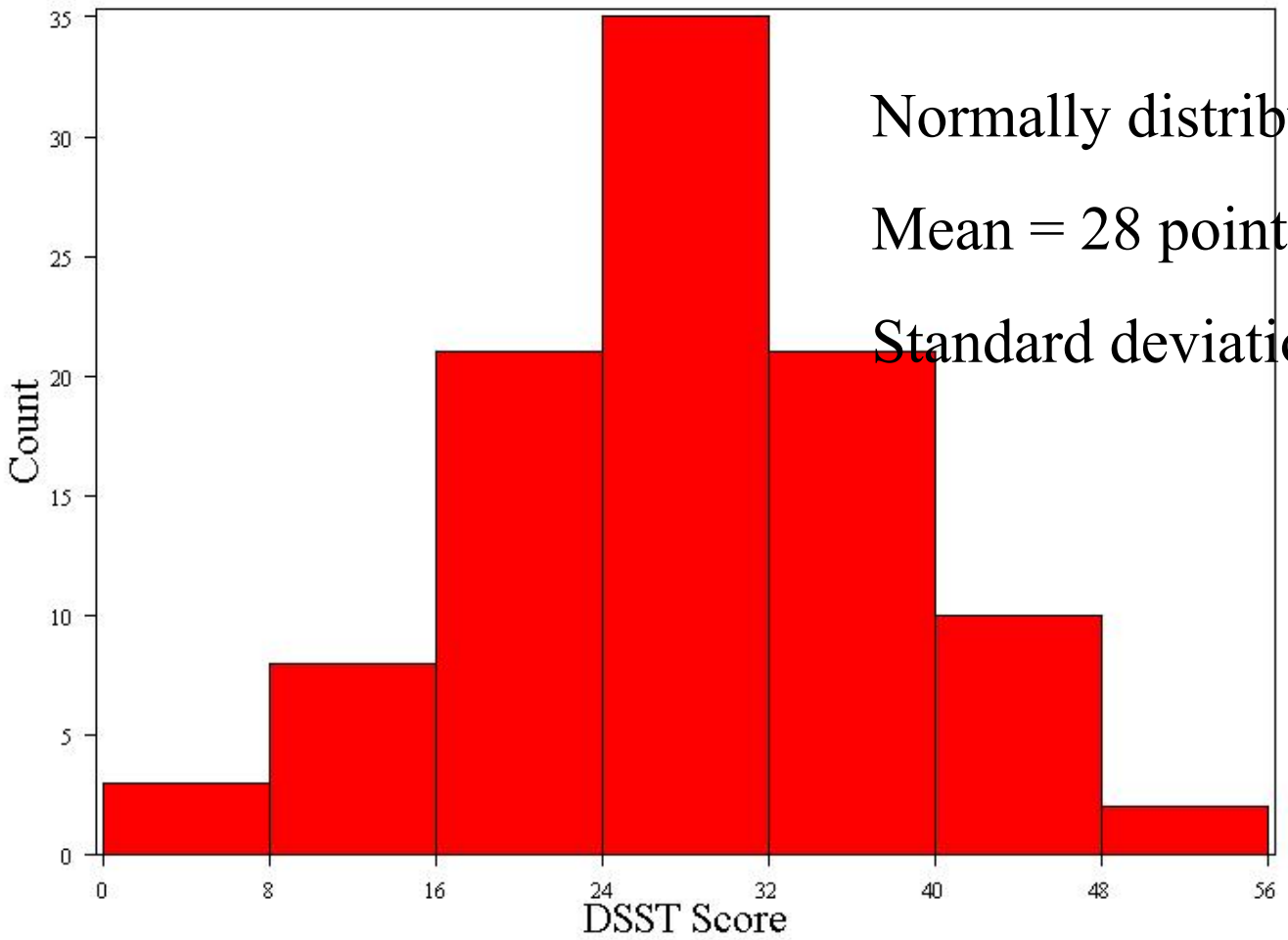
- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men.
- Estimation: What is the average serum vitamin D in middle-aged and older European men?
  - Sample statistic: mean vitamin D levels
- Hypothesis testing: Are vitamin D levels and cognitive function correlated?
  - Sample statistic: correlation coefficient between vitamin D and cognitive function, measured by the Digit Symbol Substitution Test (DSST).

1. Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

# Distribution of a trait: vitamin D



# Distribution of a trait: DSST




Normally distributed

Mean = 28 points

Standard deviation = 10 point

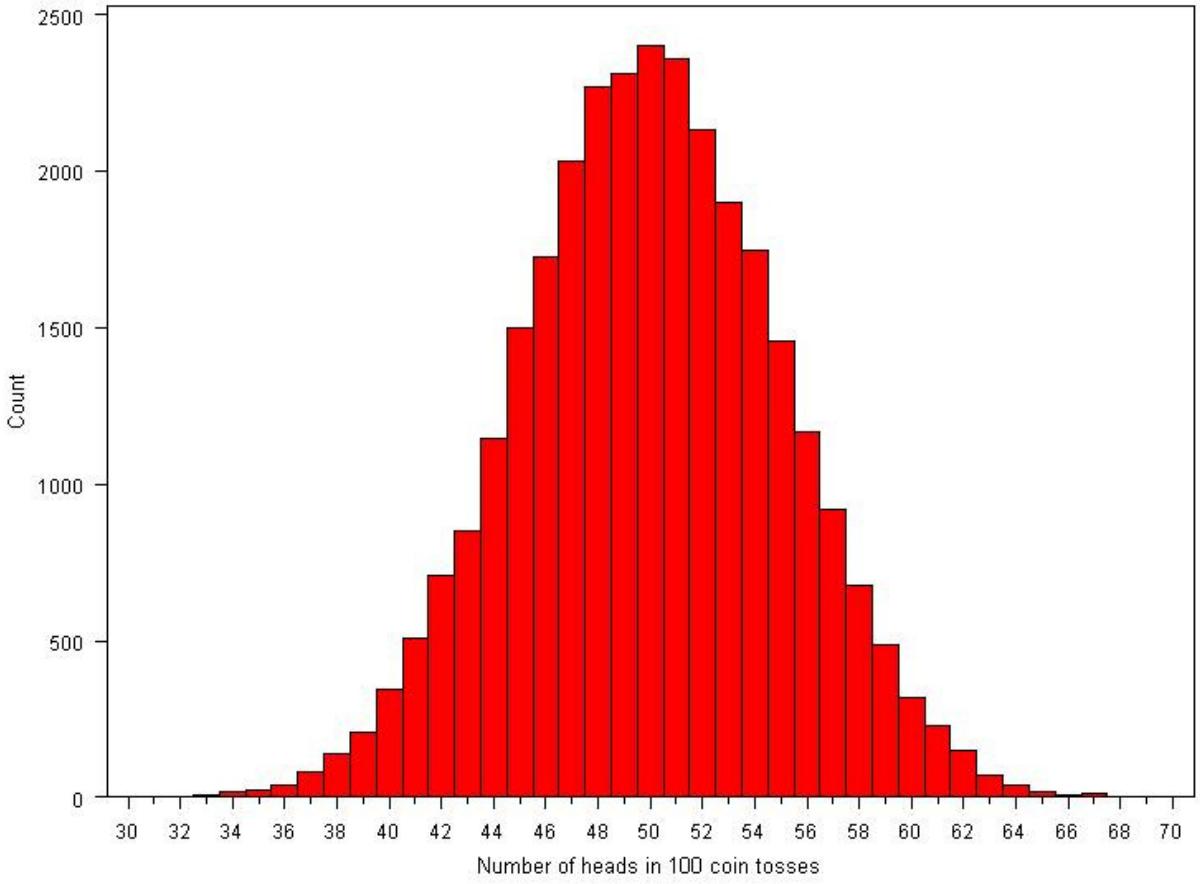
# Distribution of a statistic…

- 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:

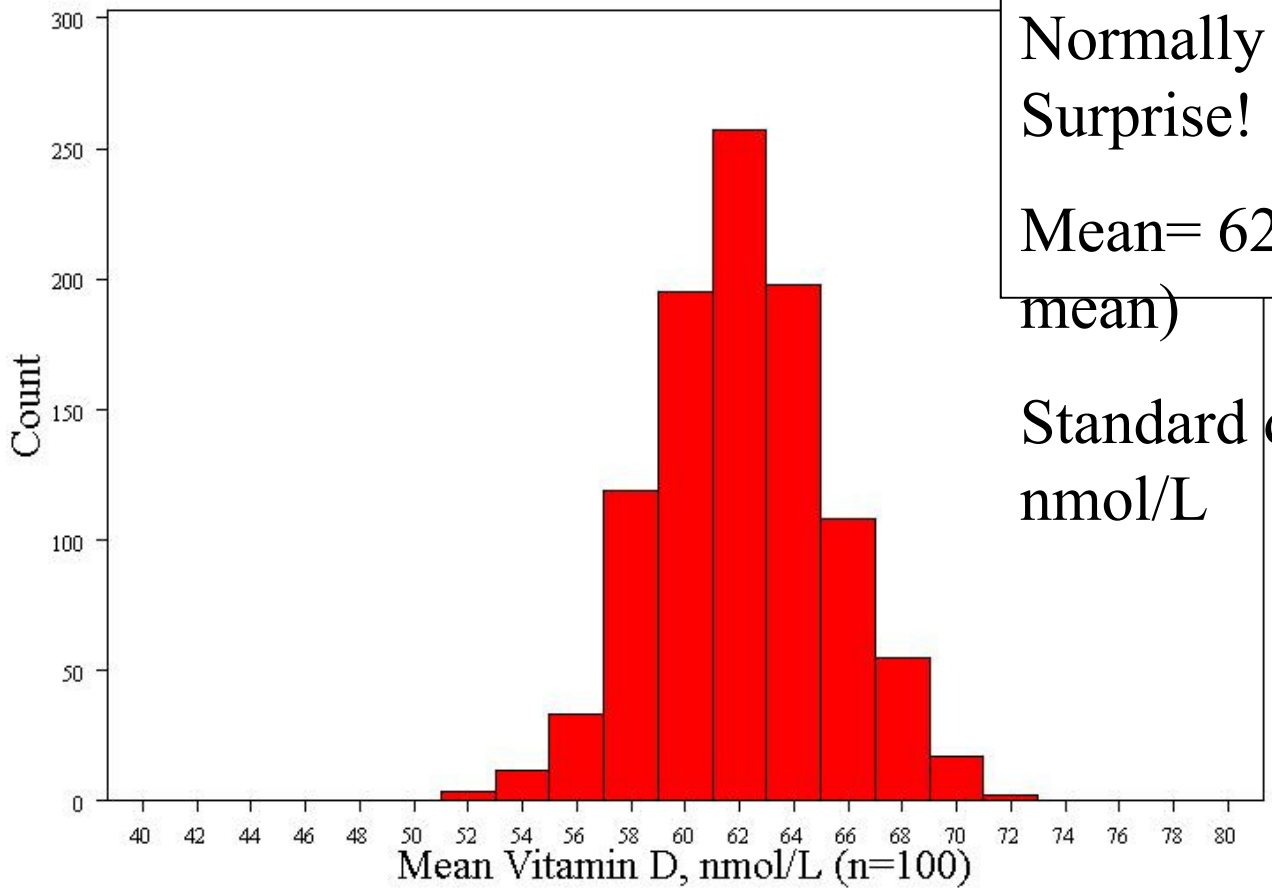
We usually get between 40 and 60 heads when we flip a coin 100 times.

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.

# Distribution of mean vitamin D (a sample statistic)




Normally distributed! Surprise!

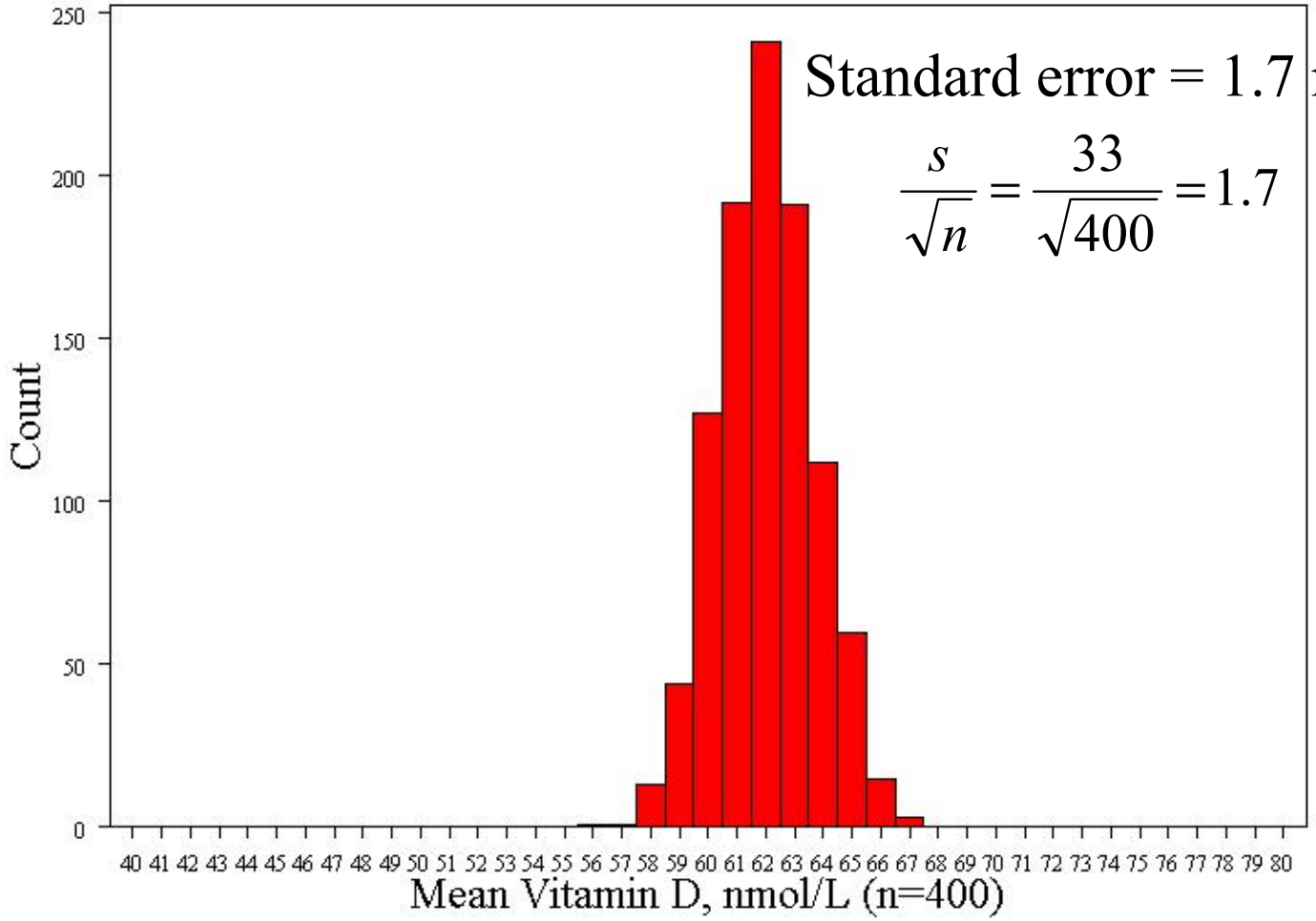
Mean= 62 nmol/L (the true mean)

Standard deviation = 3.3 nmol/L

# Distribution of mean vitamin D (a sample statistic)

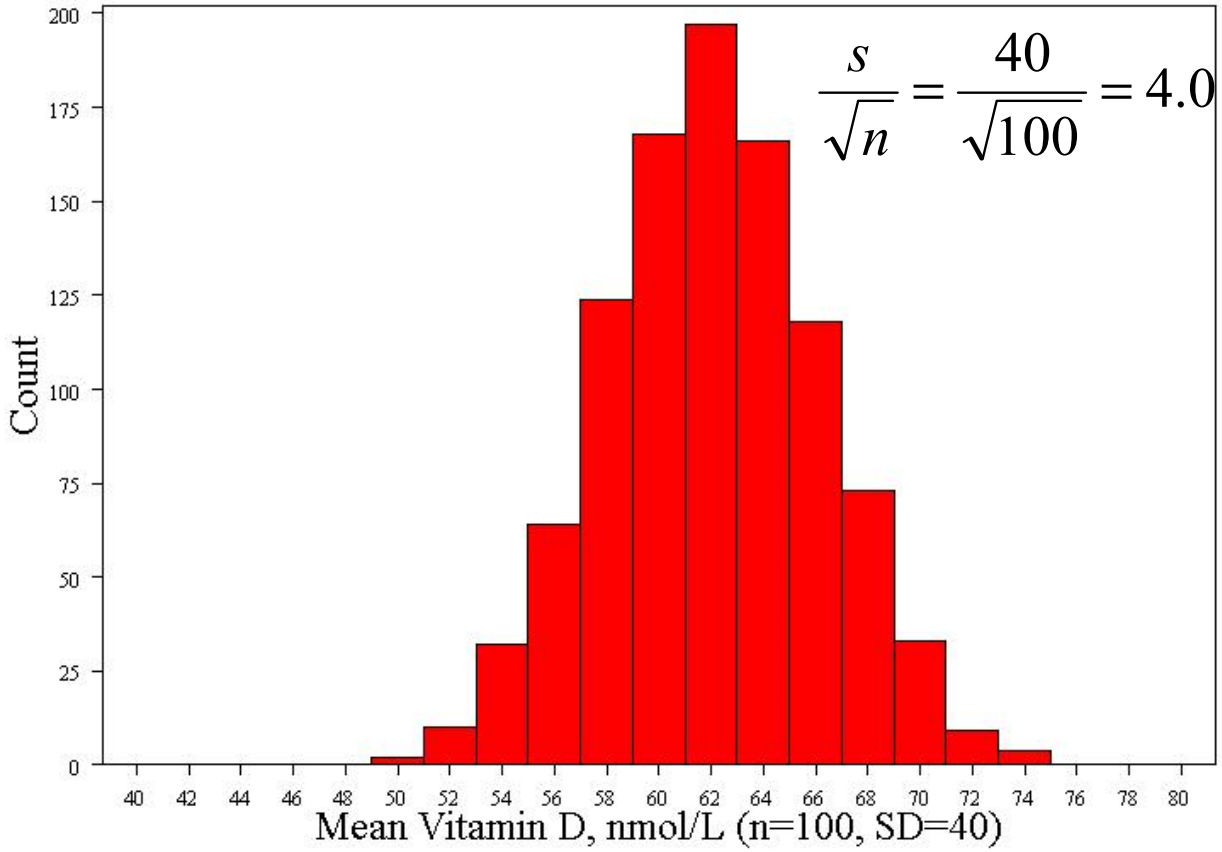
- Normally distributed (even though the trait is right-skewed!)
- Mean = true mean
- Standard deviation = 3.3 nmol/L
  - The standard deviation of a statistic is called a standard error
  - The standard error of a mean = $\frac{s}{\sqrt{n}}$

# If I increase the sample size to n=400…



# If I increase the variability of vitamin D (the trait) to SD=40…

Standard error = 4.0 nmol/L



# 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:

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

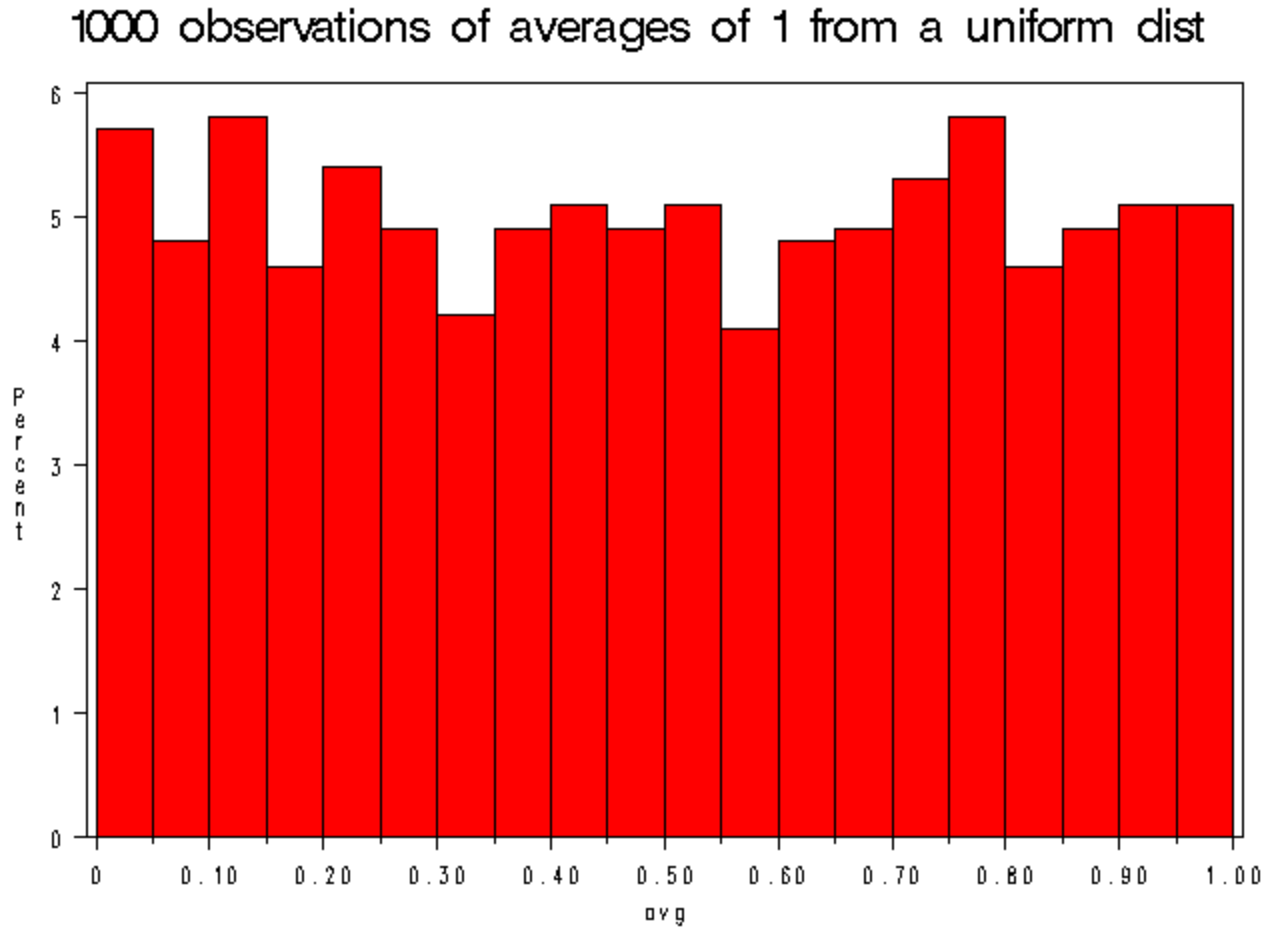
$$Var(\overline{X}_n) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} Var(x)}{n^2} = \frac{nVar(x)}{n^2} = \frac{Var(x)}{n}$$

# Computer simulation of the CLT:

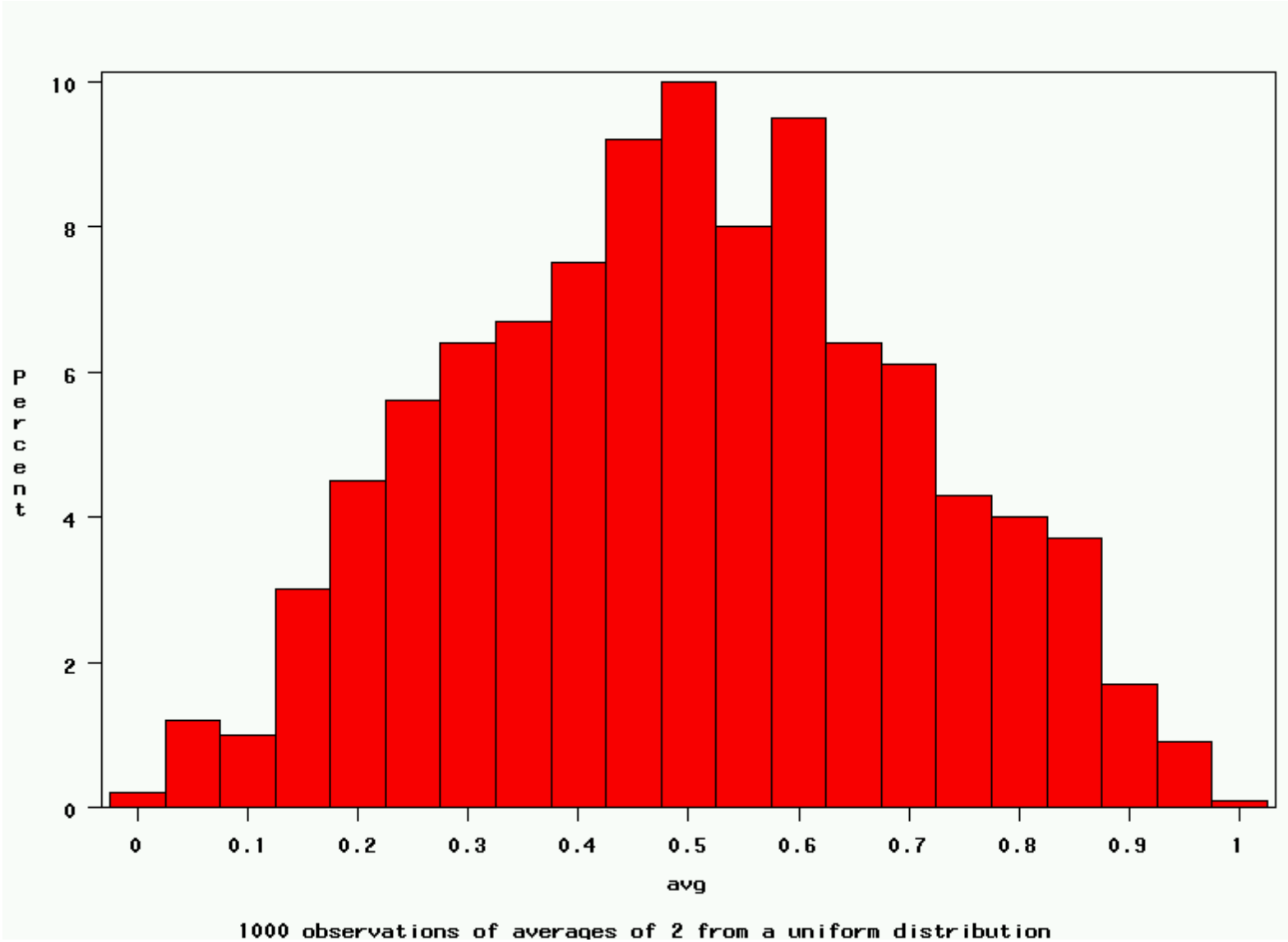
1. Pick any probability distribution and specify a mean and standard deviation.
2. Tell the computer to randomly generate 1000 observations from that probability distributions

   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.

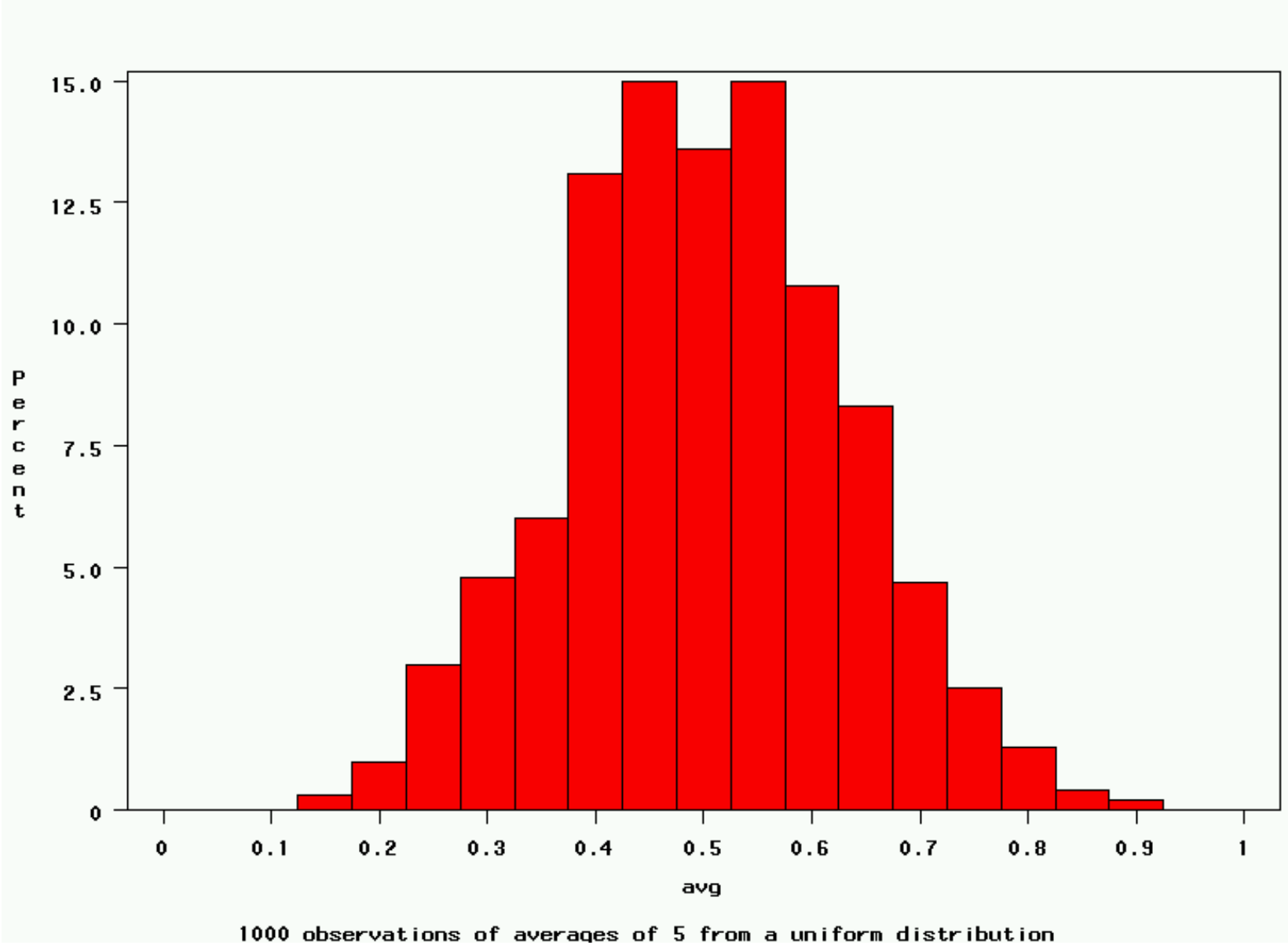
# Uniform on [0,1]: average of 1 (original distribution)



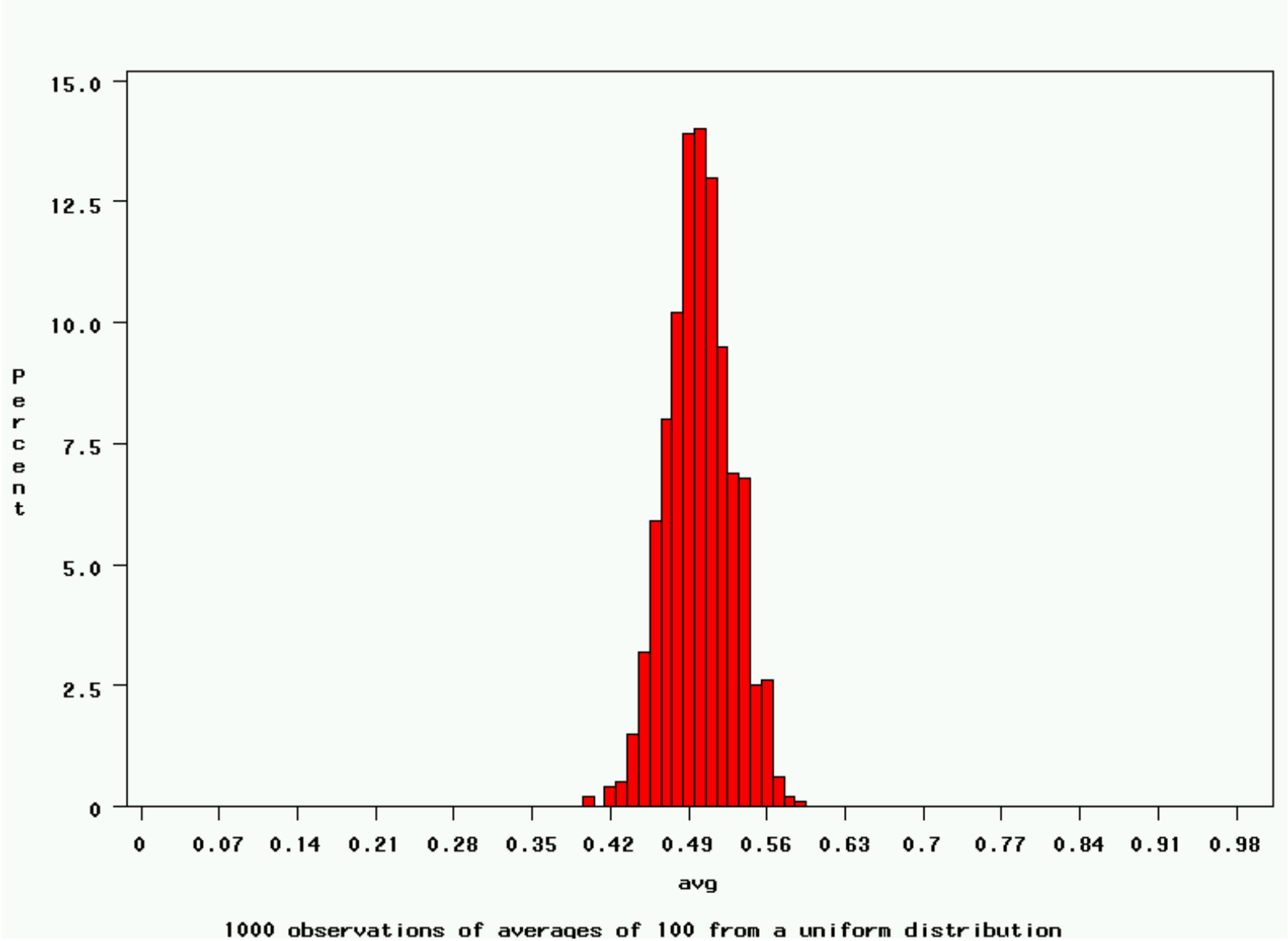
# Uniform: 1000 averages of 2



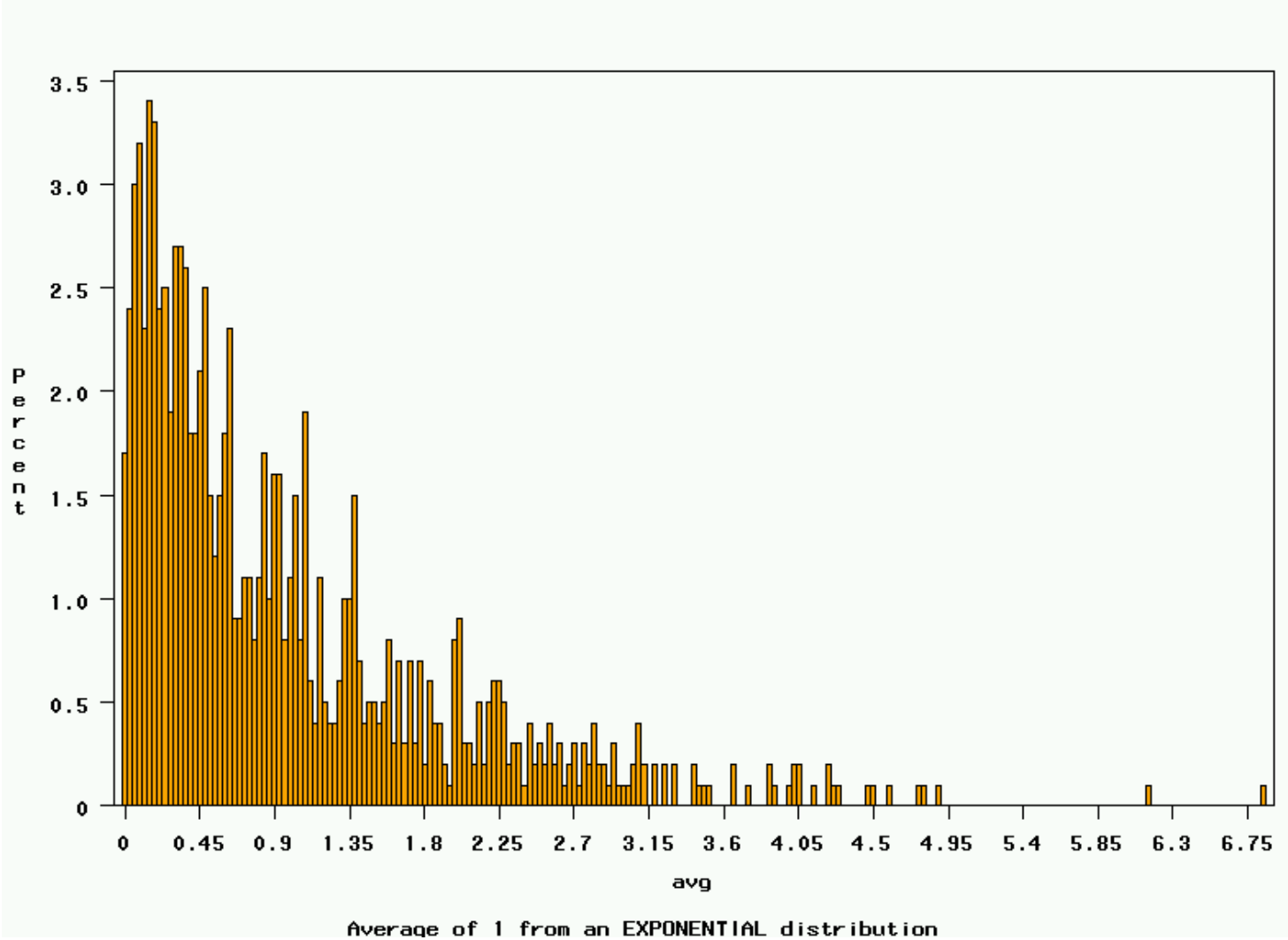
# Uniform: 1000 averages of 5



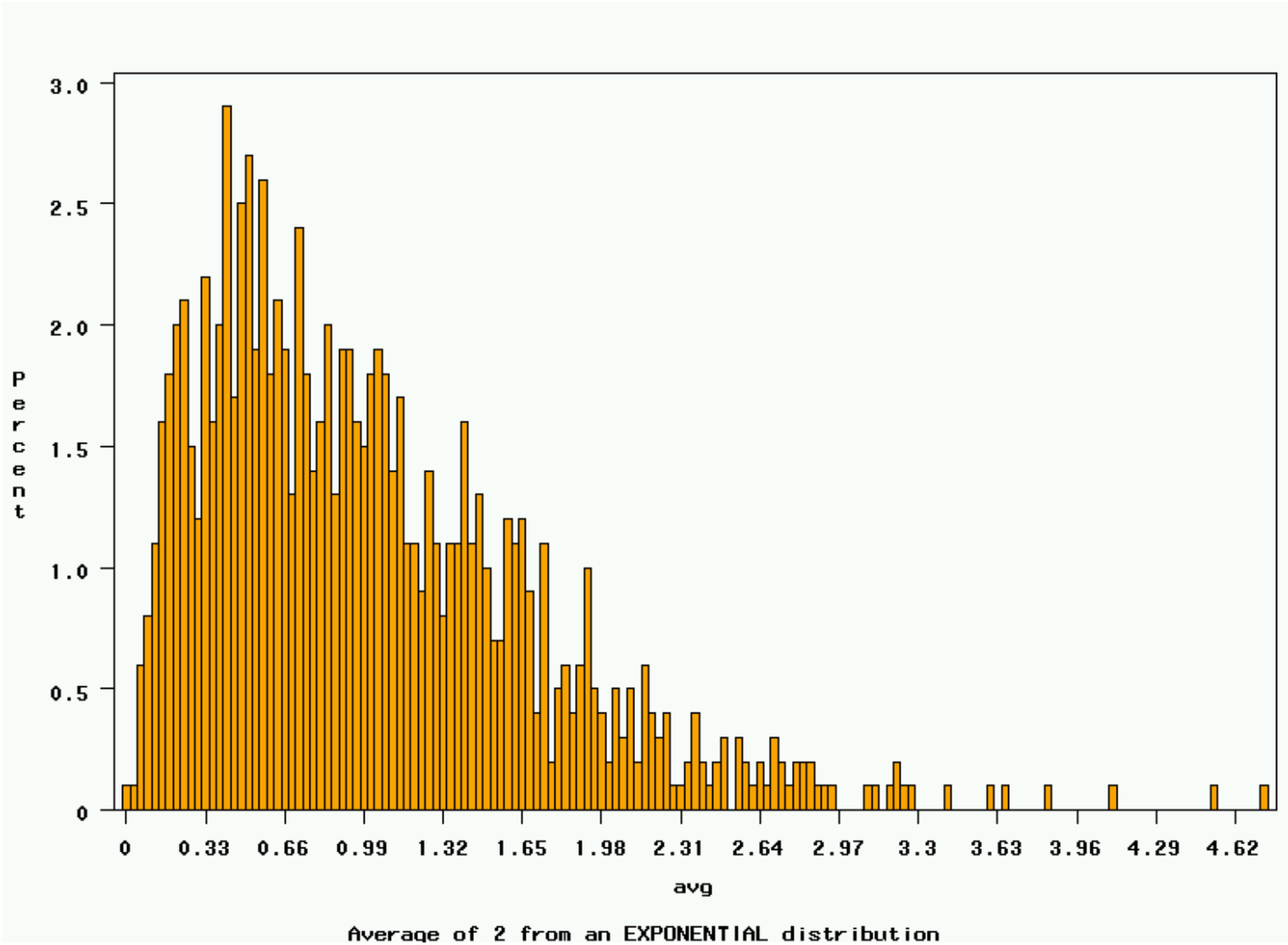
# Uniform: 1000 averages of 100



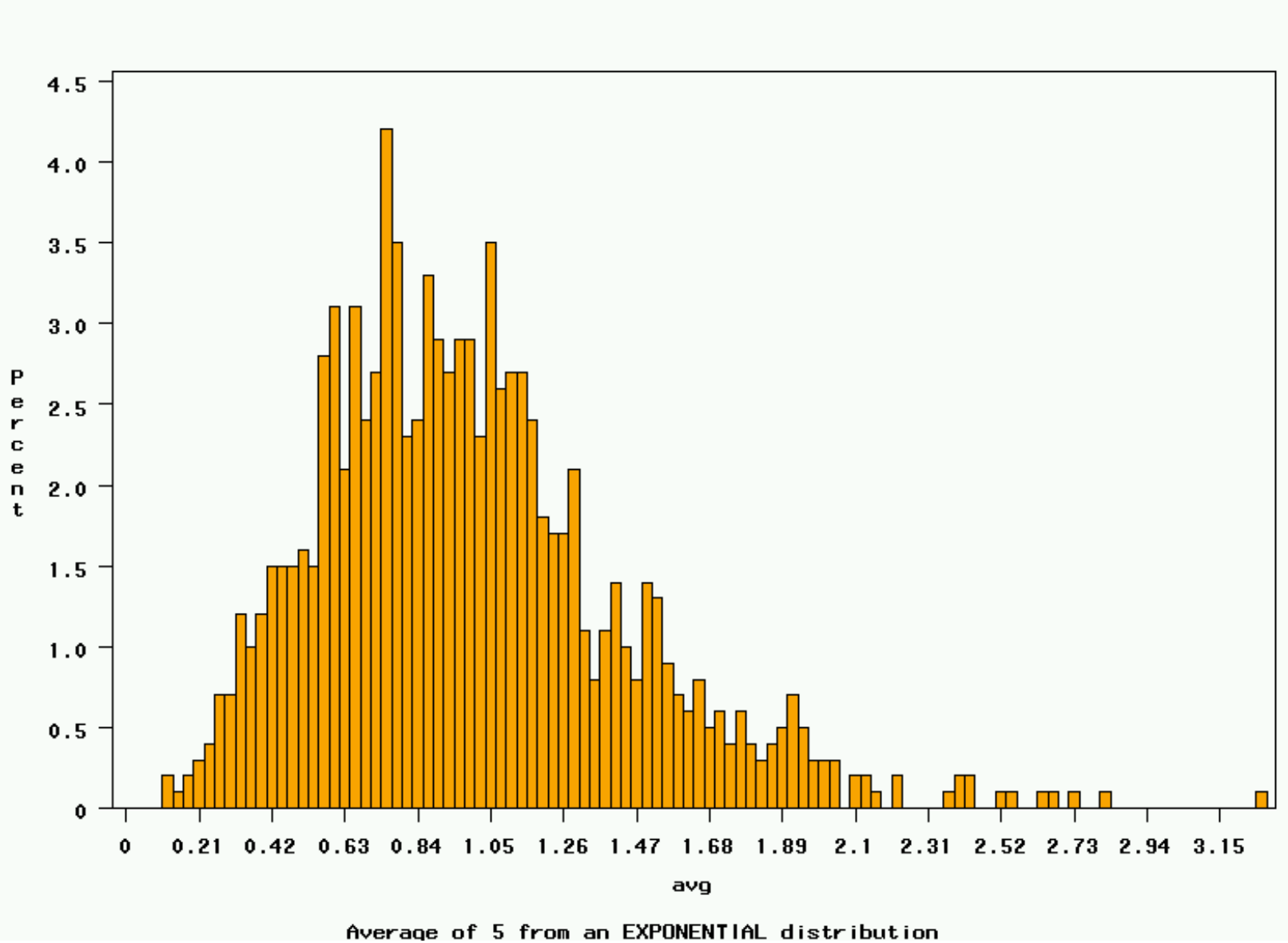
# ~Exp(1): average of 1 (original distribution)



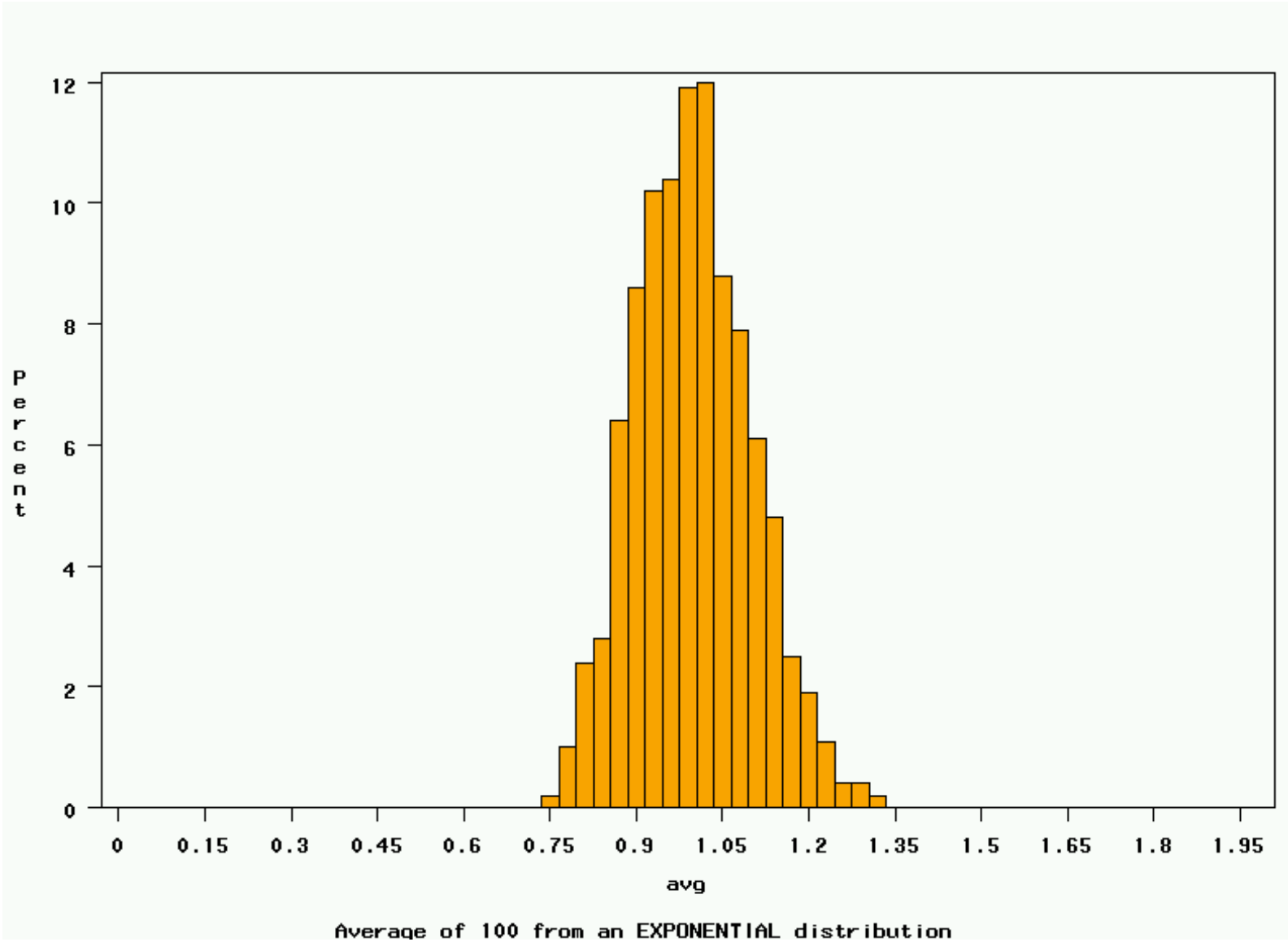
# ~Exp(1): 1000 averages of 2



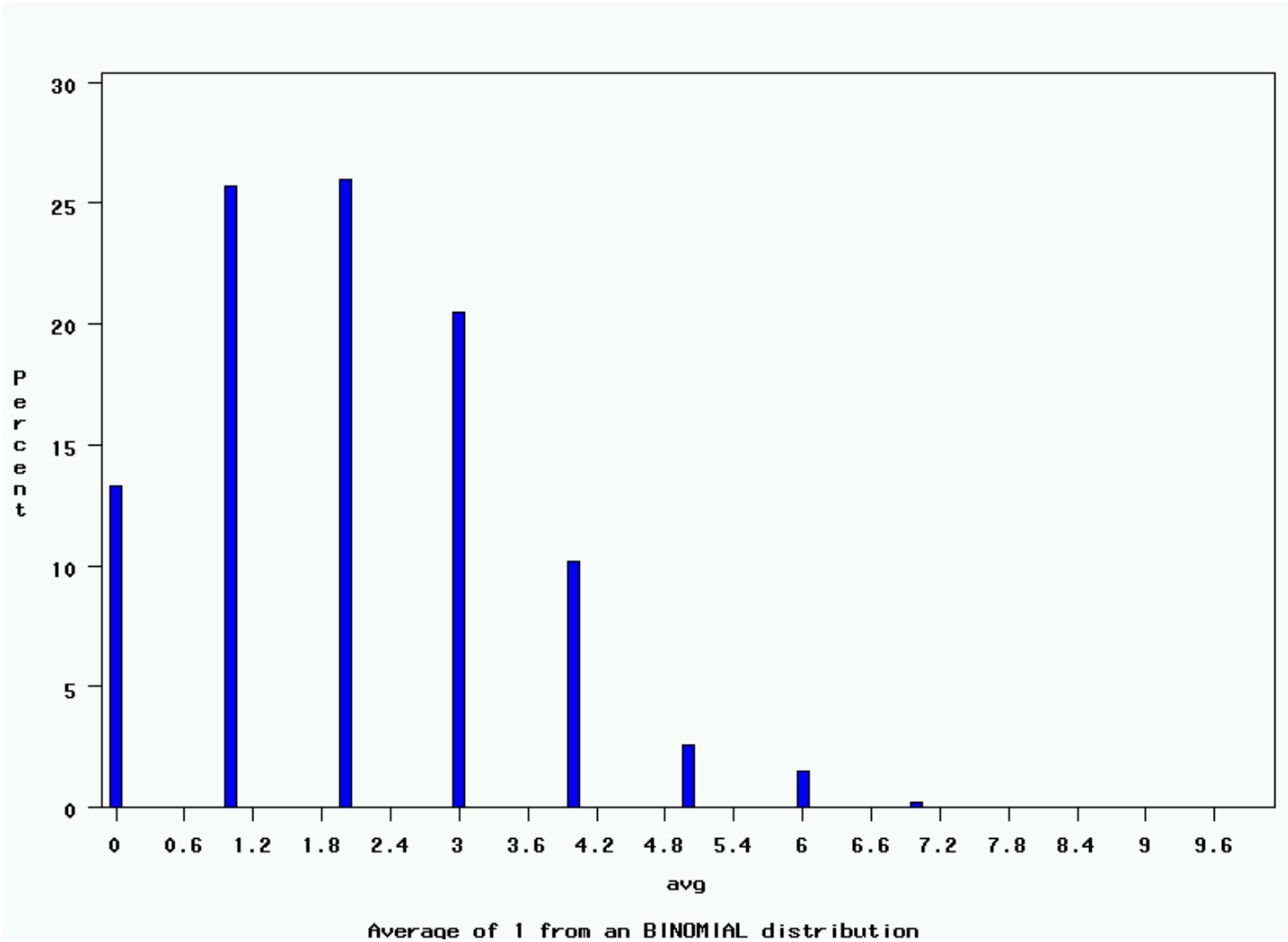
# ~Exp(1): 1000 averages of 5



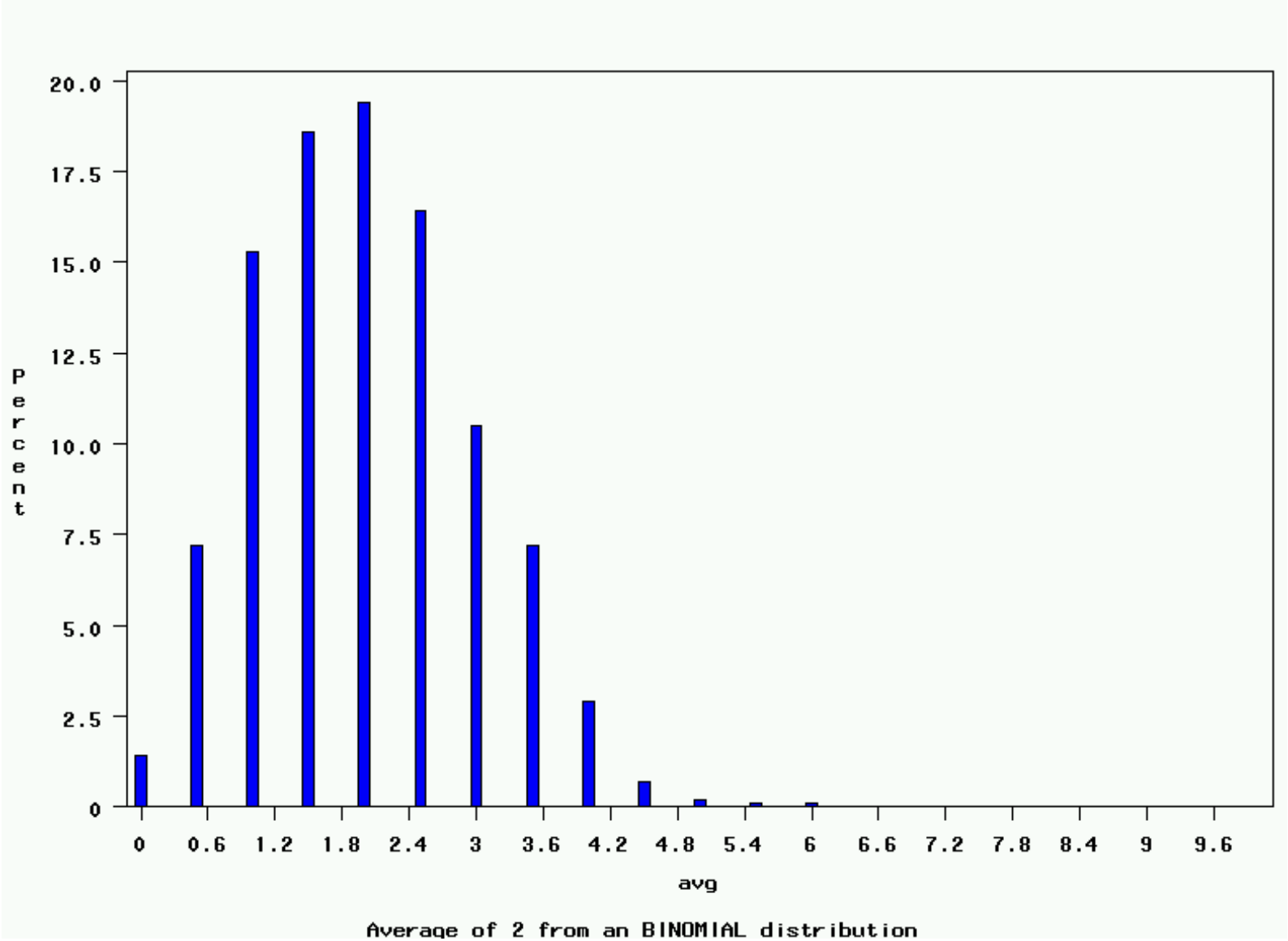
# ~Exp(1): 1000 averages of 100



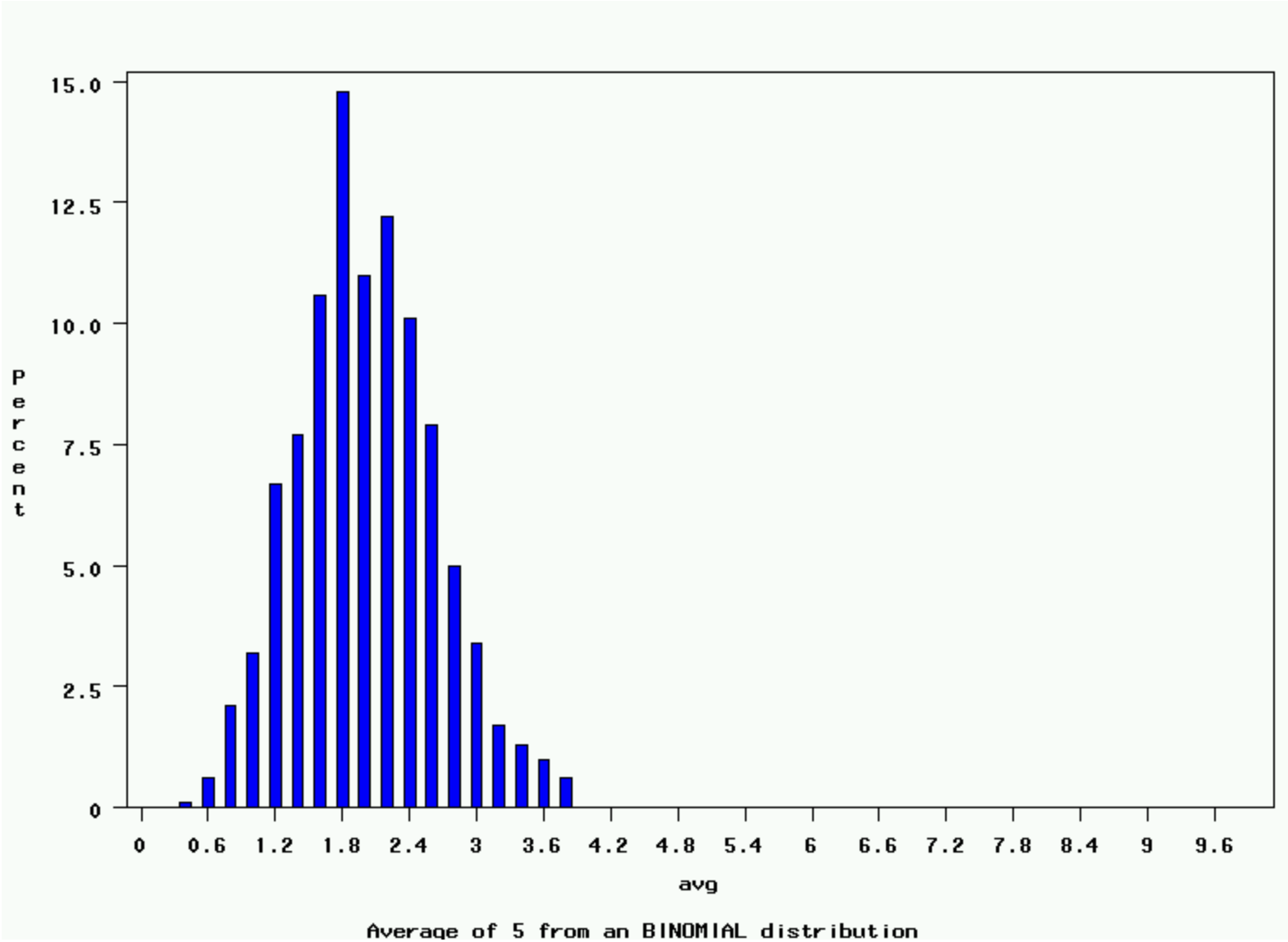
# ~Bin(40, .05): average of 1 (original distribution)



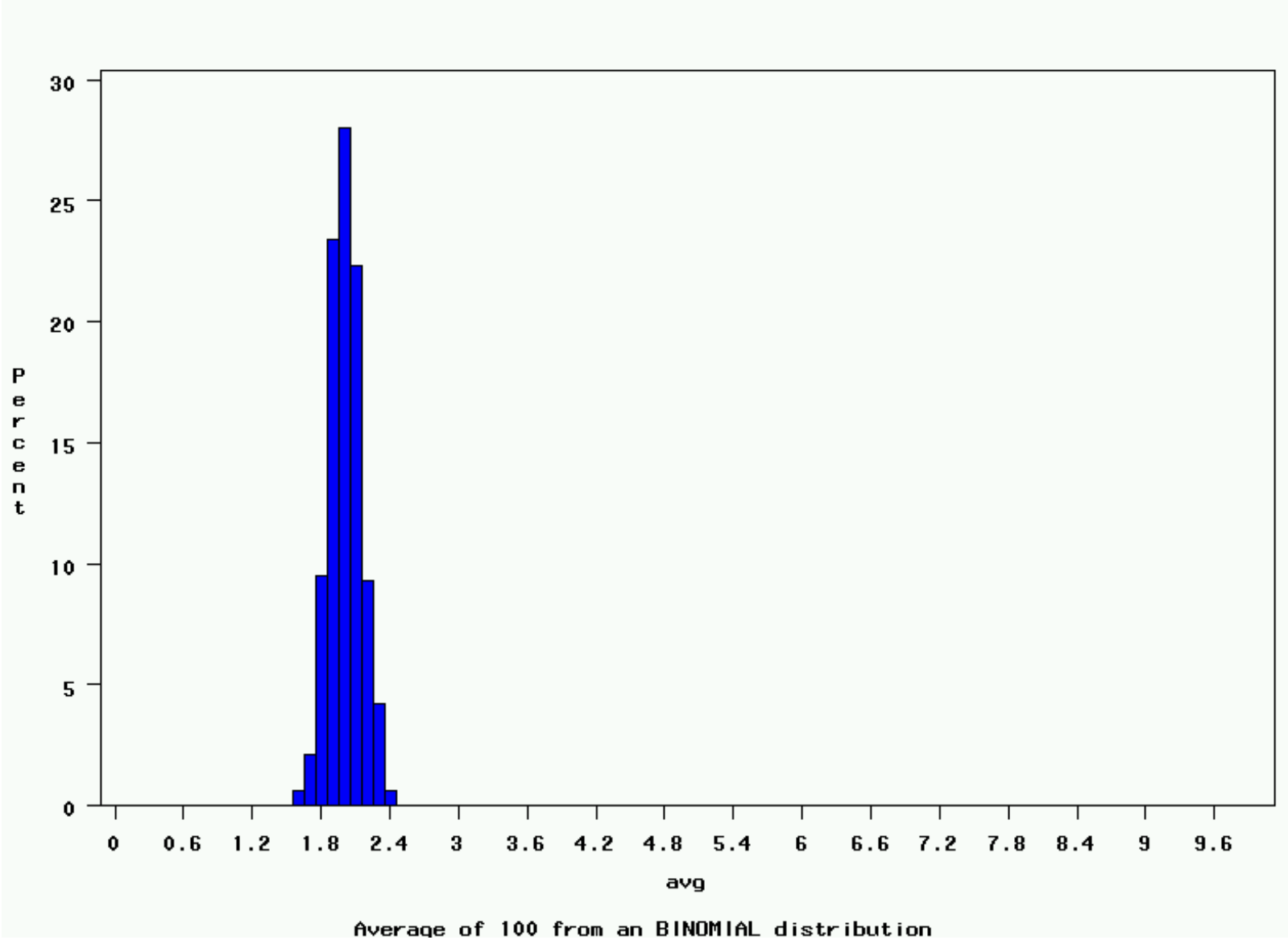
# ~Bin(40, .05): 1000 averages of 2



# ~Bin(40, .05): 1000 averages of 5



# ~Bin(40, .05): 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 $\mu$ and standard deviation $\sigma$, the sample mean $\overline{X}$ has the normal distribution with mean $\mu$ 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 $\mu$ and standard deviation $\sigma$, when $n$ is large, the distribution of $\overline{X}$ is approximately normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$. Consequently,

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

# 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 $\mu = 82$ and $\sigma = 12$. Since $n = 64$ is large, the central limit theorem tells us that the distribution of the sample mean is approximately normal with

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

Converting to the standard normal variable:

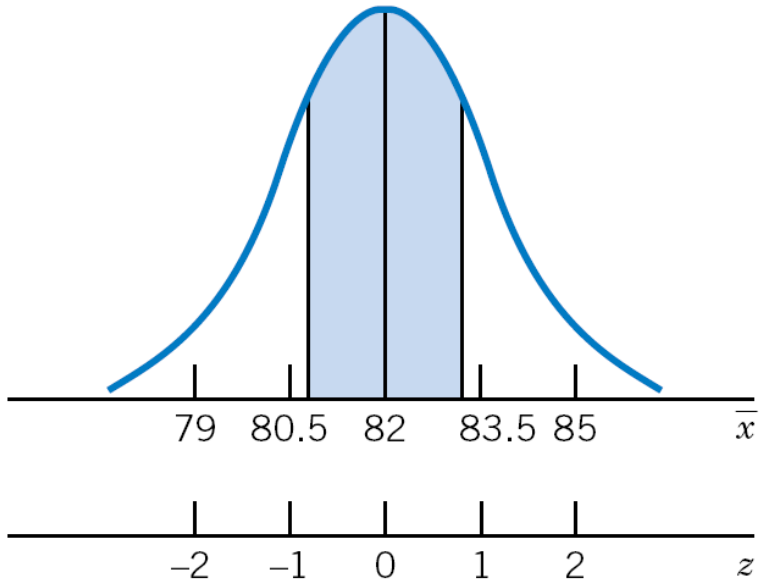
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 82}{1.5}$$

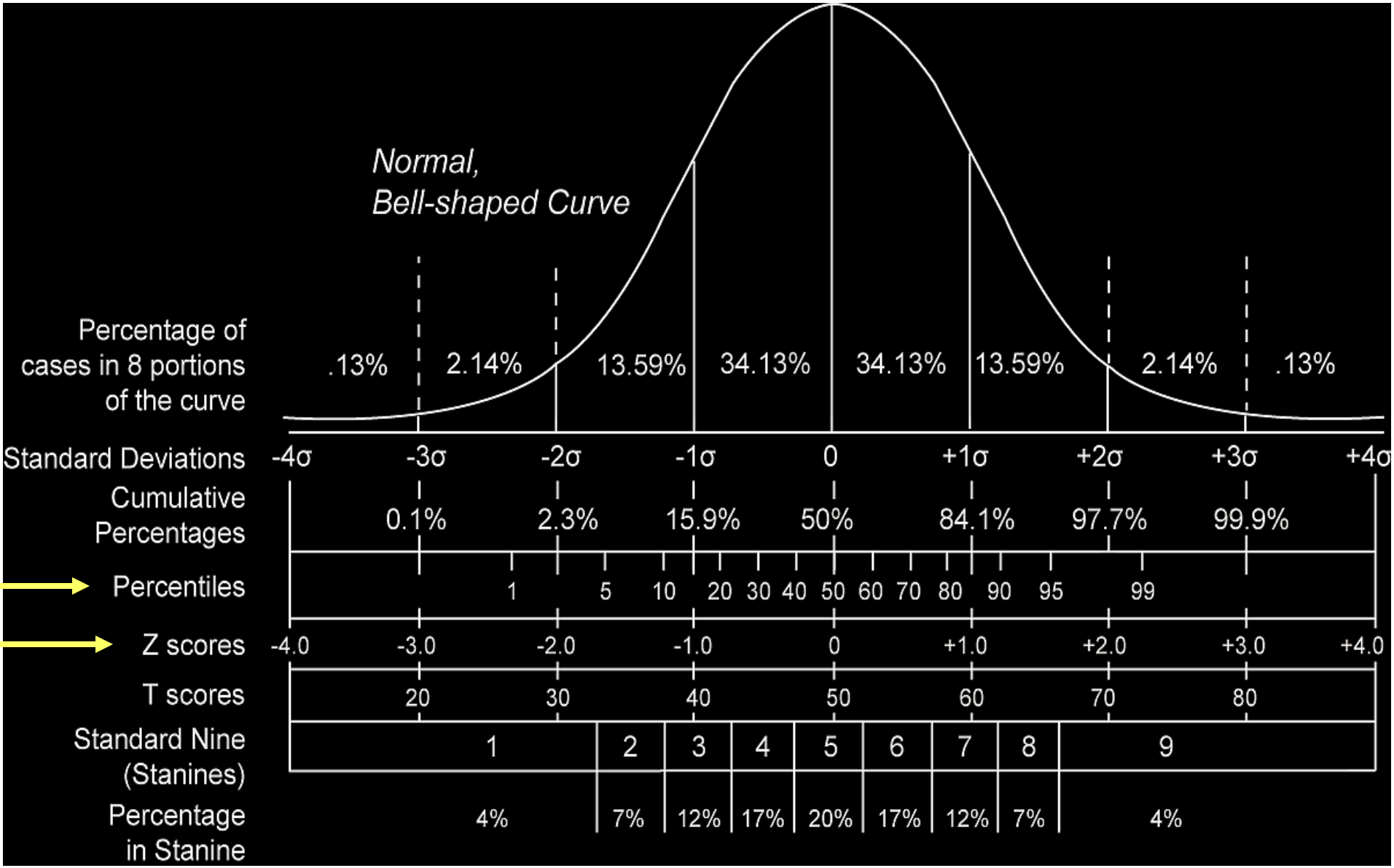
Thus,

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

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

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




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%

## 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 $\gamma$, 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:

## Standardizing in mathematical statistics [edit]

*Further information: Normalization (statistics)*

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

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

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

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

then the standardized version is

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

## 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]

# 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 $\mu$ and a standard deviation $\sigma$, then the sampling distribution of all theoretically possible sample means will be a normal distribution having a mean $\mu$ and a standard deviation $\hat{\sigma}$ given by $\frac{s_Y}{\sqrt{n}}$

[Sirkin (1999), p. 239]

| | *Mean* | *Standard Deviation* | *Variance* |
|---|---|---|---|
| *Universe* | $\mu_Y$ | $\sigma_Y$ | $\sigma_Y^2$ |
| *Sampling Distribution* | $\mu_Y$ | $\hat{\sigma}_Y$ | $\hat{\sigma}_Y^2$ |
| *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:

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

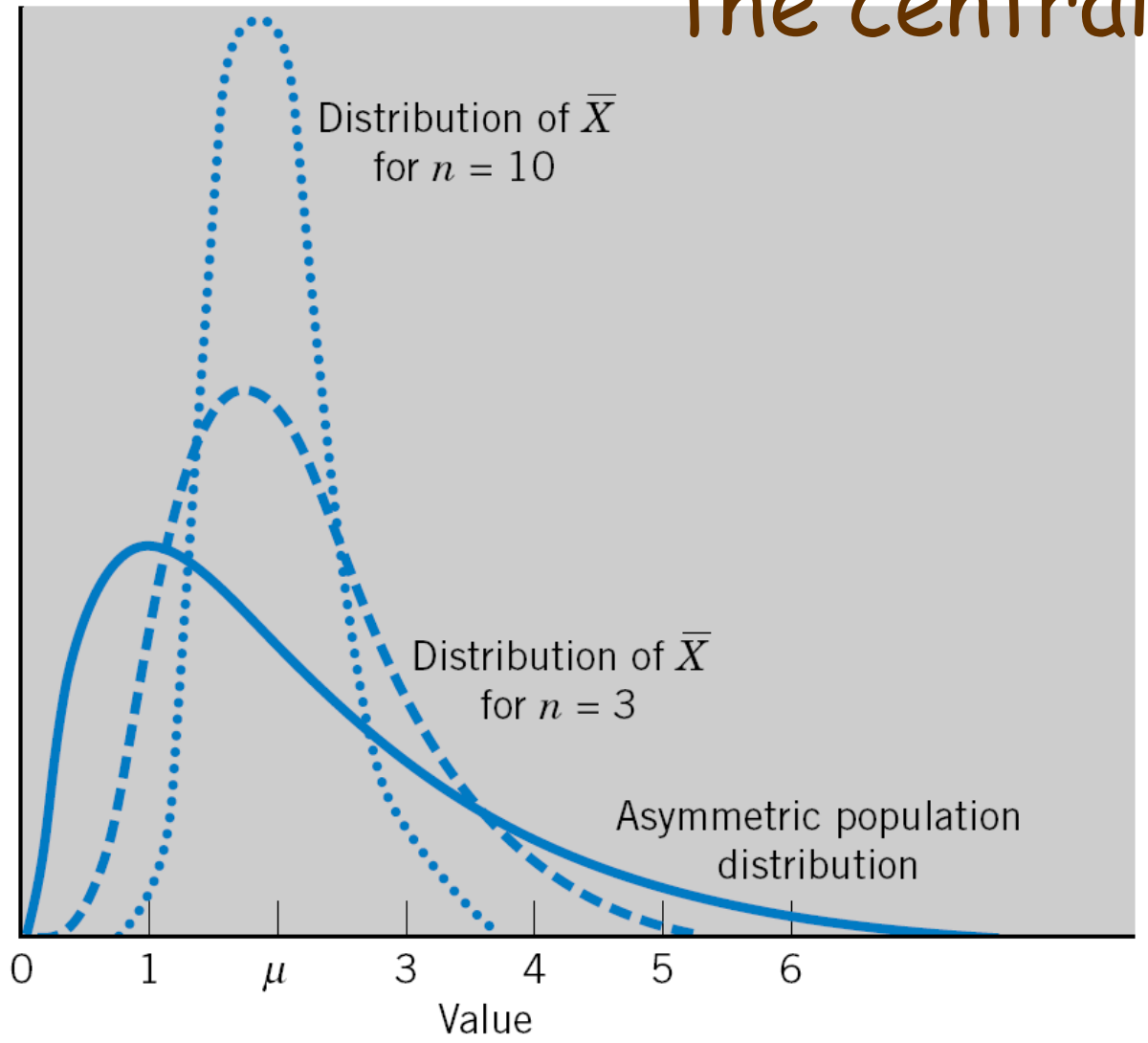
The **standard error** is estimated by ***dividing*** the **standard deviation** of the sample by the ***square root*** of the **size of the sample**. In our example,

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

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

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

# 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*** $\mu$ is equal to the ***mean*** in the ***universe*** (also $\mu$).
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.

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

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 $\mu$, population mean, they are still calculated the same way: $\sum \frac{x_i}{n}$ .
2. Identify whether the standard deviation is known, $\sigma$, 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 |

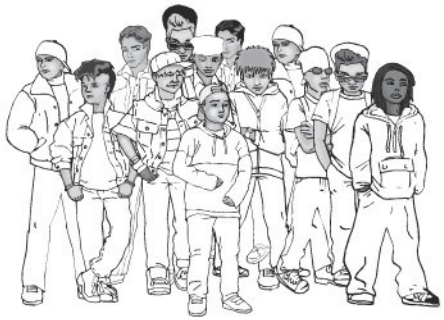
   - If the standard deviation is unknown then t* is used as the critical value. This value is dependent on the confidence level (C) for the test and degrees of freedom. The degrees of freedom is found by subtracting one from the number of observations, $n - 1$. The critical value is found from the t-distribution 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: $\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$

## STATISTICAL INFERENCE

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

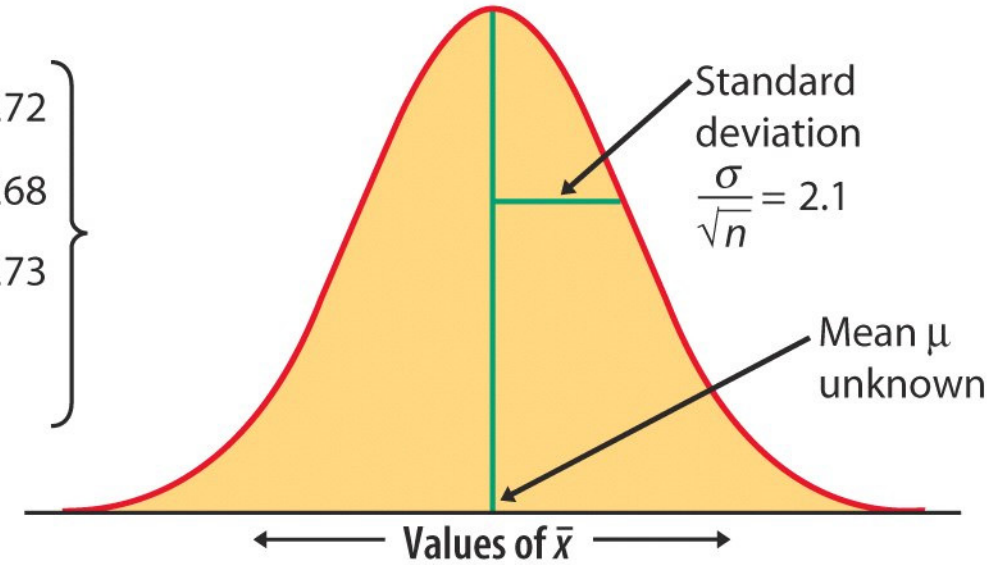
## INFERENCE ABOUT A MEAN: SIMPLE CONDITIONS

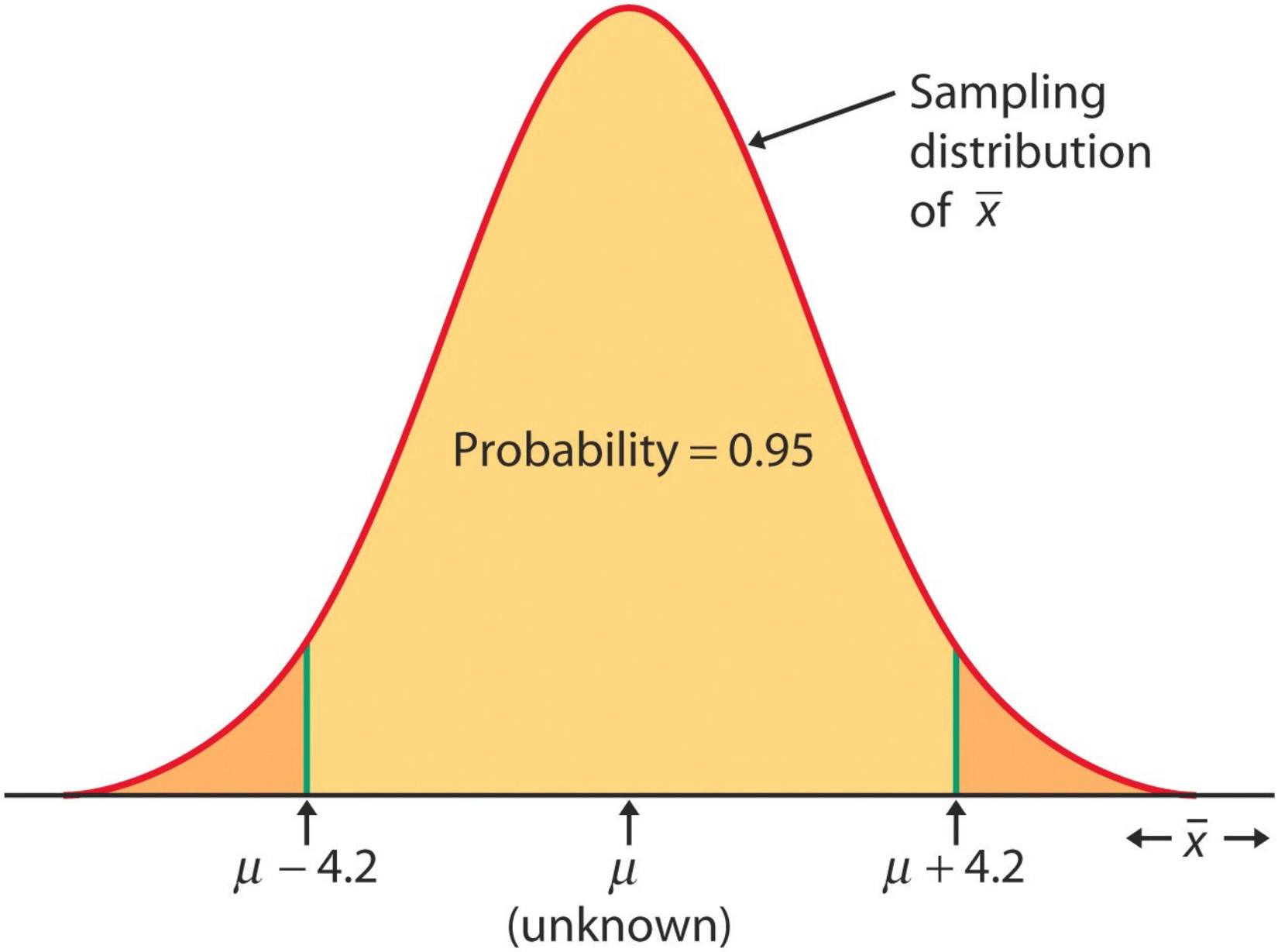
1. We have an SRS from the population of interest.
2. The variable we measure has a perfectly Normal distribution $N(\mu, \sigma)$ in the population.
3. We don't know the population mean $\mu$. Our task is to infer something about $\mu$ from the sample data. But we do know the population standard deviation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

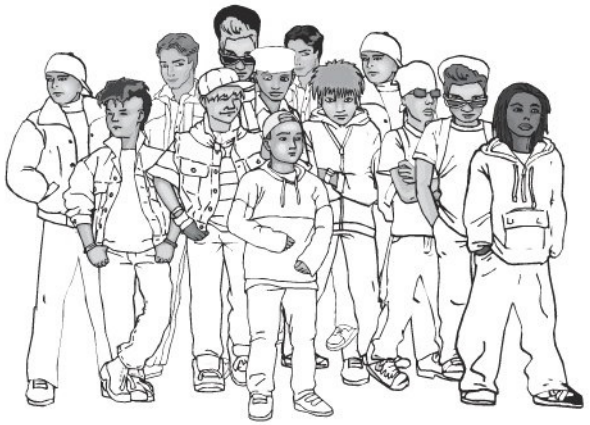


Population
$\mu = ?$
$\sigma = 60$

SRS $n = 840$ → $\bar{x} = 272$
SRS $n = 840$ → $\bar{x} = 268$
SRS $n = 840$ → $\bar{x} = 273$
•
•
•




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



Population
$\mu = ?$
$\sigma = 60$

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 $\mu$ of the population.

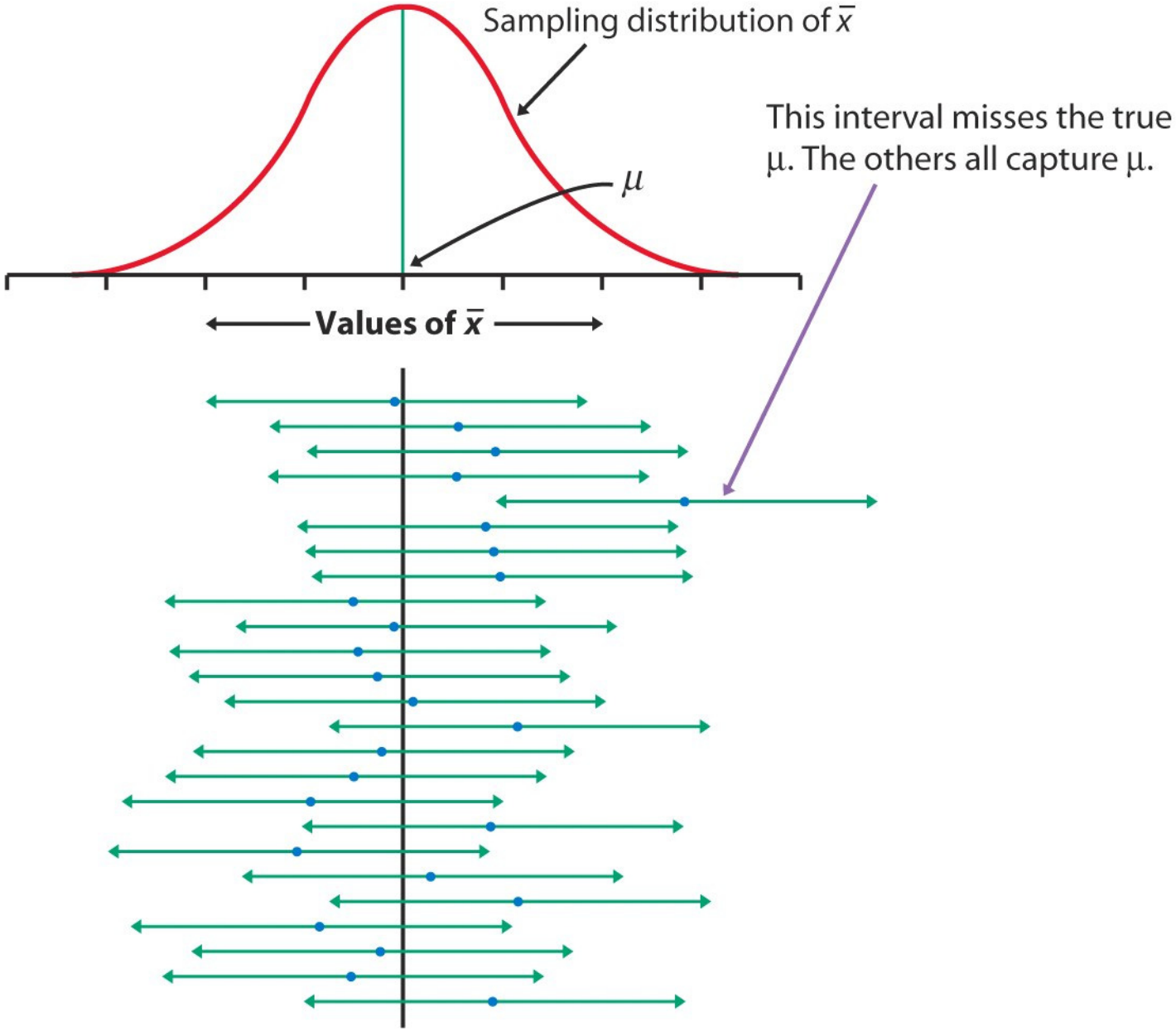
## CONFIDENCE INTERVAL

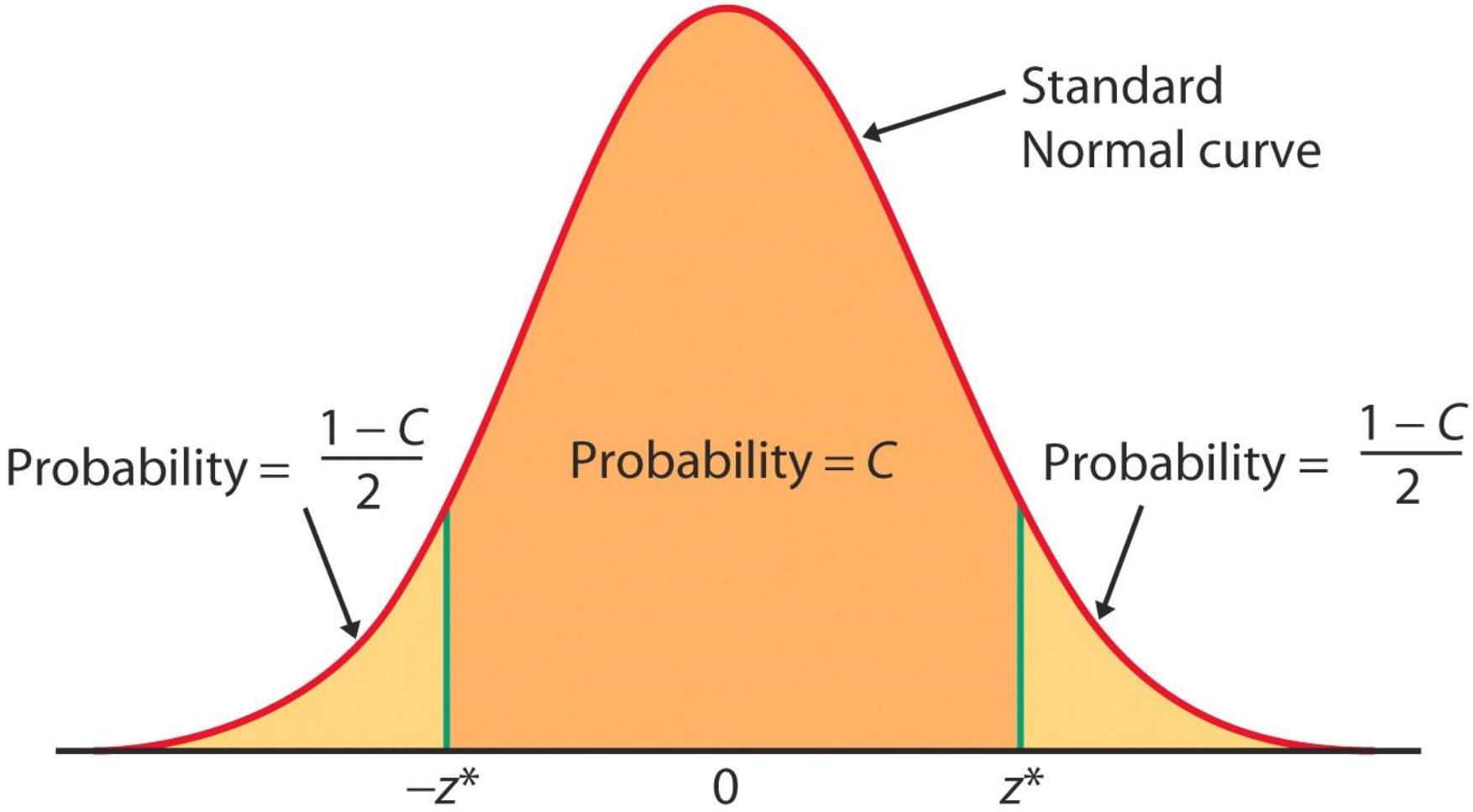
A **level C confidence interval** for a parameter has two parts:

- 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 $\bar{x}$

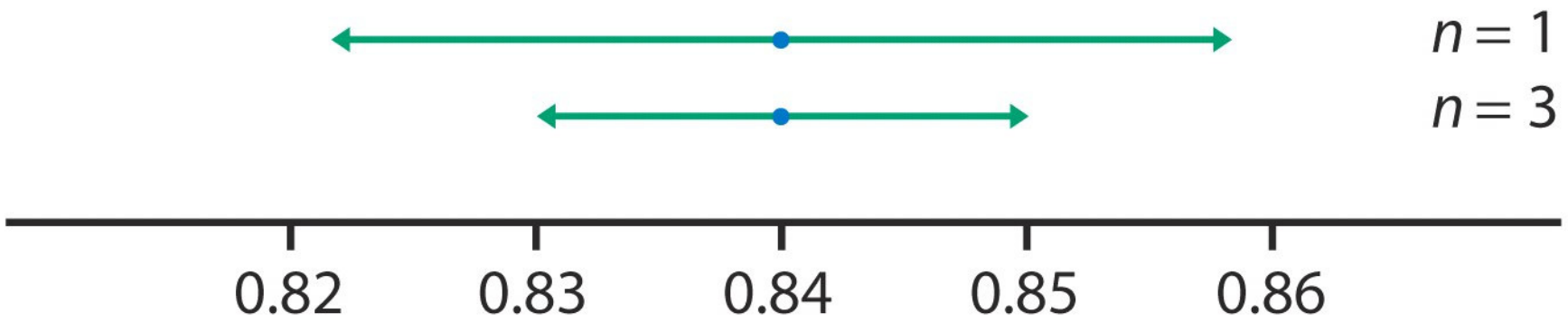

Standard
Normal curve
Probability = $\frac{1-C}{2}$
Probability = $C$
Probability = $\frac{1-C}{2}$
$-z^*$
0
$z^*$

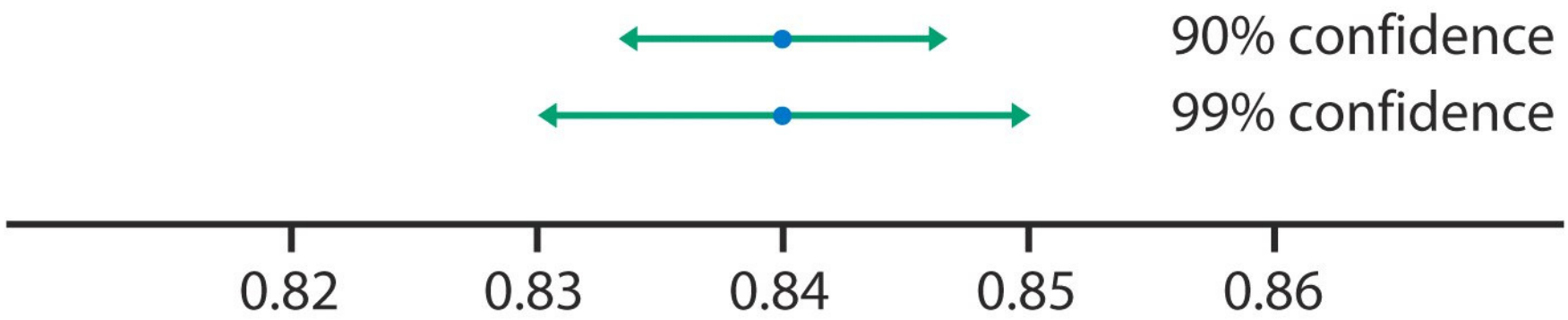
## CONFIDENCE INTERVAL FOR THE MEAN OF A NORMAL POPULATION

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

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

The critical value $z^*$ is illustrated in Figure 13.5 and found in Table C.


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 $\mu$ and standard deviation $\sigma$. In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both $\mu$ and $\sigma$ are unknown parameters.

## STANDARD ERROR

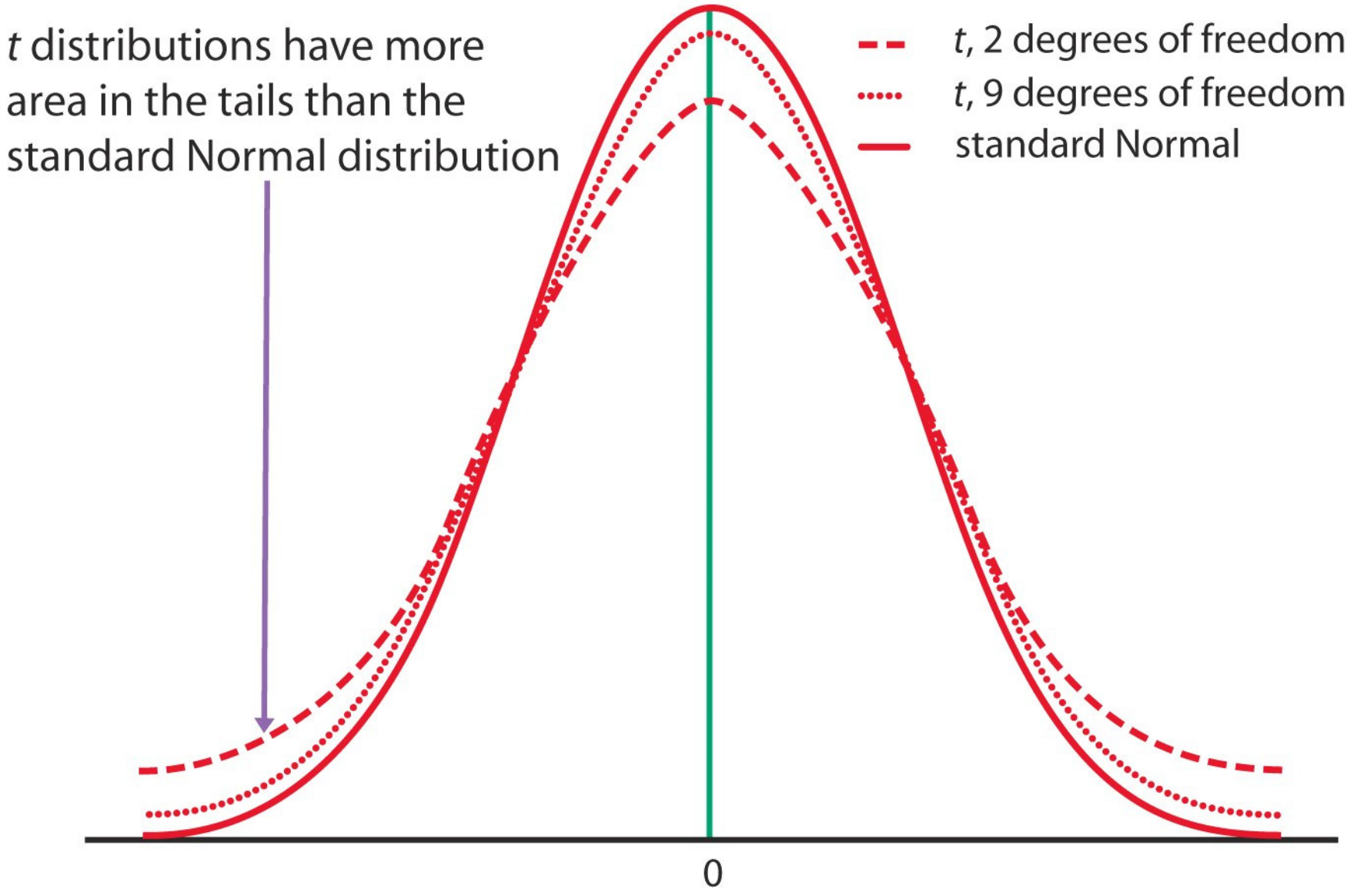
When the standard deviation of a statistic is estimated from data, the result is called the **standard error** of the statistic. The standard error of the sample mean $\overline{x}$ is $s/\sqrt{n}$.

## 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 $\mu$ and standard deviation $\sigma$. 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 $\mu$. A level C confidence interval for $\mu$ 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.

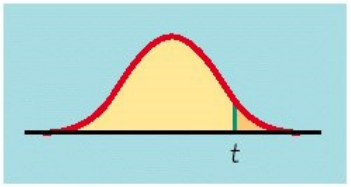
## THE ONE-SAMPLE $t$ TEST

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

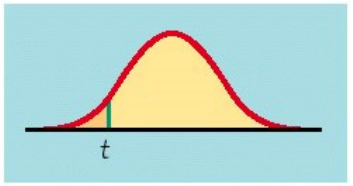
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

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

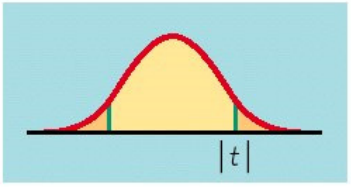
$H_a$: $\mu > \mu_0$ is $P(T \geq t)$




$H_a$: $\mu < \mu_0$ 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.

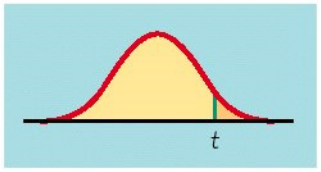
## THE ONE-SAMPLE $t$ TEST

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

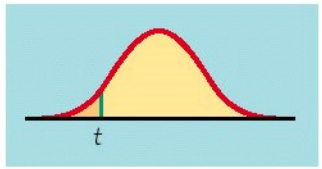
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

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

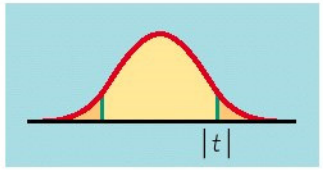
$H_a$: $\mu > \mu_0$ is $P(T \geq t)$




$H_a$: $\mu < \mu_0$ 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.

# The Chi-Square Test

# Significance Test

- If the distributions of the second variable are nearly the same given the category of the first variable, then we say that there is not an association between the two variables.
- If there are significant differences in the distributions, then we say that there is an association between the two variables.
- Significance test is needed to draw a conclusion.

# Hypothesis Test

- Hypotheses:
  - Null: the percentages for one variable are the same for every level of the other variable
    (no difference in conditional distributions).
    (No real relationship).
  - Alt: the percentages for one variable vary over levels of the other variable. (Is a real relationship).

Null hypothesis:
The percentages for one variable are the same for every level of the other variable.
(No real relationship).

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 24% | 25% |
| Somewhat better | 23% | 23% |
| About the same | 31% | 36% |
| Somewhat worse | 16% | 13% |
| Much worse | 6% | 3% |
| Total | 100% | 100% |

For example, could look at differences in percentages between Canada and U.S. for each level of "Quality of life":

24% vs. 25% for those who felt '*Much better*',

23% vs. 23% for '*Somewhat better*', etc.

Problem of *multiple comparisons*!

# Hypothesis Test

- **$H_0$**: no real relationship between the two categorical variables that make up the rows and columns of a two-way table
- **To test $H_0$, compare the *observed counts* in the table (the original data) with the *expected counts* (the counts we would expect if $H_0$ were true)**
  - if the observed counts are far from the expected counts, that is evidence against $H_0$ in favor of a real relationship between the two variables

# 3. Expected Counts

- The expected count in any cell of a two-way table (when $H_0$ is true) is

$$\text{expected count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

For the observed data to the right, find the expected value for each cell:

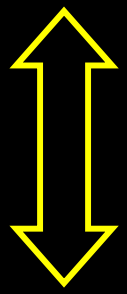
| Quality of life | Canada | United States | Total |
|---|---|---|---|
| Much better | 75 | 541 | 616 |
| Somewhat better | 71 | 498 | 569 |
| About the same | 96 | 779 | 875 |
| Somewhat worse | 50 | 282 | 332 |
| Much worse | 19 | 65 | 84 |
| Total | 311 | 2165 | 2476 |

For the expected count of *Canadians* who feel '*Much better*' (expected count for Row 1, Column 1):

$$\text{expected count} = \frac{(\text{row1 total}) \times (\text{column1 total})}{\text{table total}} = \frac{616 \times 311}{2476} = 77.37$$

150

Observed counts:



Compare to see if the data support the null hypothesis

Expected counts:

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 75 | 541 |
| Somewhat better | 71 | 498 |
| About the same | 96 | 779 |
| Somewhat worse | 50 | 282 |
| Much worse | 19 | 65 |

| Quality of life | Canada | United States |
|---|---|---|
| Much better | 77.37 | 538.63 |
| Somewhat better | 71.47 | 497.53 |
| About the same | 109.91 | 765.09 |
| Somewhat worse | 41.70 | 290.30 |
| Much worse | 10.55 | 73.45 |

# 4. Chi-Square Statistic

- To determine if the differences between the observed counts and expected counts are statistically significant (to show a real relationship between the two categorical variables), we use the ***chi-square statistic***:

$$X^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

where the sum is over all cells in the table.

# Chi-Square Statistic

- The chi-square statistic is a measure of the distance of the observed counts from the expected counts
  - is always zero or positive
  - is only zero when the observed counts are exactly equal to the expected counts
  - large values of $X^2$ are evidence against $H_0$ because these would show that the observed counts are far from what would be expected if $H_0$ were true

| | Observed counts | | Expected counts | |
|---|---|---|---|---|
| Quality of life | Canada | United States | Canada | United States |
| Much better | 75 | 541 | 77.37 | 538.63 |
| Somewhat better | 71 | 498 | 71.47 | 497.53 |
| About the same | 96 | 779 | 109.91 | 765.09 |
| Somewhat worse | 50 | 282 | 41.70 | 290.30 |
| Much worse | 19 | 65 | 10.55 | 73.45 |

$$X^2 = \sum\left[\frac{(75-77.37)^2}{77.37} + \frac{(541-538.63)^2}{538.63} + \cdots\right]$$

$$= 0.073 + 0.010 + \cdots$$

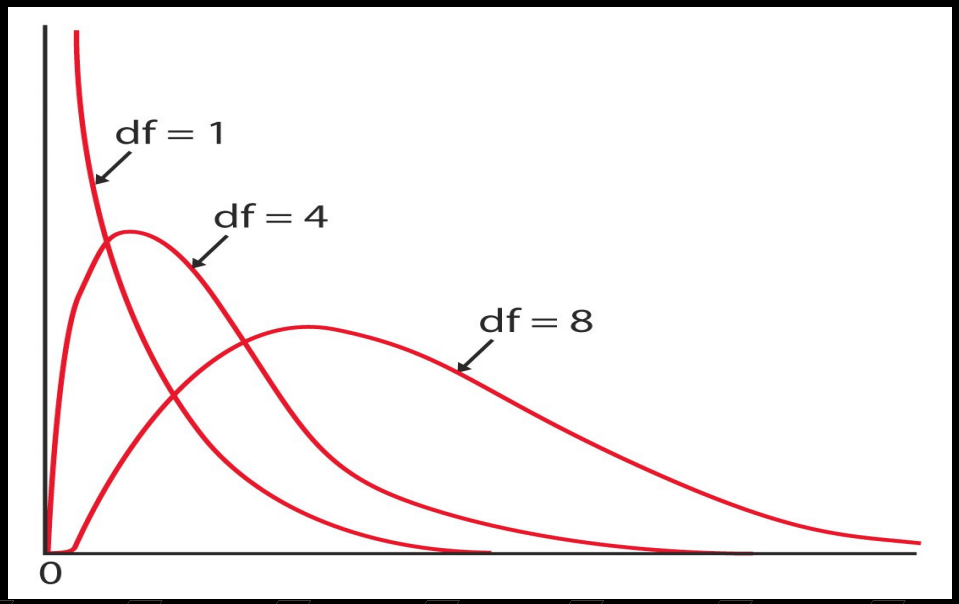
$$= 11.725$$

# 5. Chi-Square Test

- Calculate value of chi-square statistic
- Find *P*-value in order to reject or fail to reject $H_0$
  - use **chi-square table** for **chi-square distribution** (next few slides)
  - from computer output

# Chi-Square Distributions

- Family of distributions that take only positive values and are skewed to the right
- Specific chi-square distribution is specified by giving its *degrees of freedom* (similar to *t* dist.)



# Chi-Square Test

- Chi-square test for a two-way table with *r* rows and *c* columns uses critical values from a chi-square distribution with **(*r* – 1)×(*c* – 1)** degrees of freedom
- *P*-value is the area to the right of $X^2$ under the density curve of the chi-square distribution
  - use *chi-square table*
  - *P-value* = $P(X^2 >= X_{obs}^2)$

# 6. Uses of the Chi-Square Test

- Tests the null hypothesis

  *$H_0$: no relationship between two categorical variables*

  when you have a two-way table from either of these situations:

  - Independent SRSs from each of several populations, with each individual classified according to one categorical variable
    [Example: Health Care case study: two samples (Canadians & Americans); each individual classified according to "Quality of life"]
  - A single SRS with each individual classified according to both of two categorical variables
    [Example: Sample of 8235 subjects, with each classified according to their "Job Grade" (1, 2, 3, or 4) and their "Marital Status" (Single, Married, Divorced, or Widowed)]

# Chi-Square Test: Requirements

- The chi-square test is an approximate method, and becomes more accurate as the counts in the cells of the table get larger
- The following must be satisfied for the approximation to be accurate:
  - No more than 20% of the expected counts are less than 5
  - All individual expected counts are 1 or greater
  - In particular, all four expected counts in a 2×2 table should be 5 or greater
- If these requirements fail, then two or more groups must be combined to form a new ('smaller') two-way table

# Summary: steps to do chi-square test

1. Find row total, col total, grand total.
2. Find expected count for each cell.
3. Find test statistic $X^2$: df = (r-1)(c-1)

$$X^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

4. Use Table E to find P-value:

   P-value = $P(X^2 >= X_{obs}^2)$

5. Compare P-value with significance level and draw conclusion.