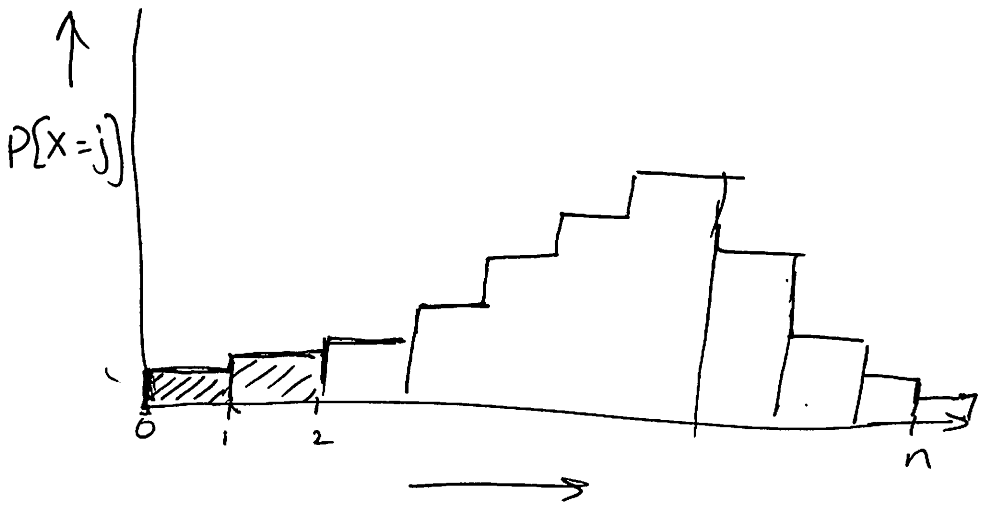
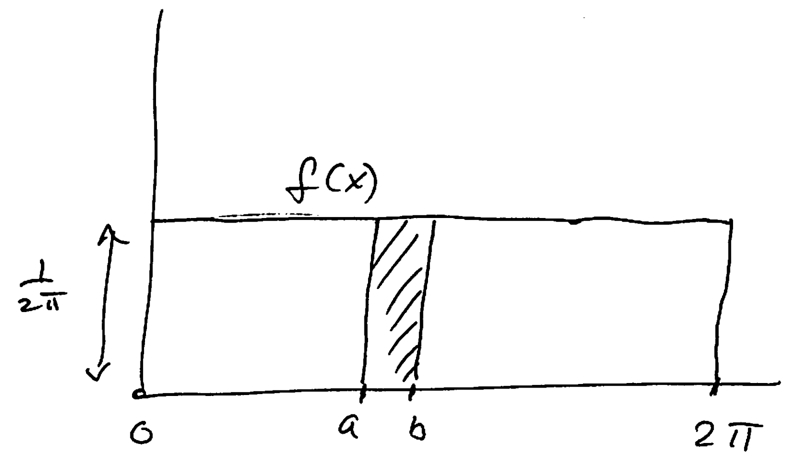
CS70 Fall 2013
Lecture 26

Normal Distribution + How to lie with Statistics

Flip coin n times:
$X = \#$ H's.

Pick a random pt in a unit circle.

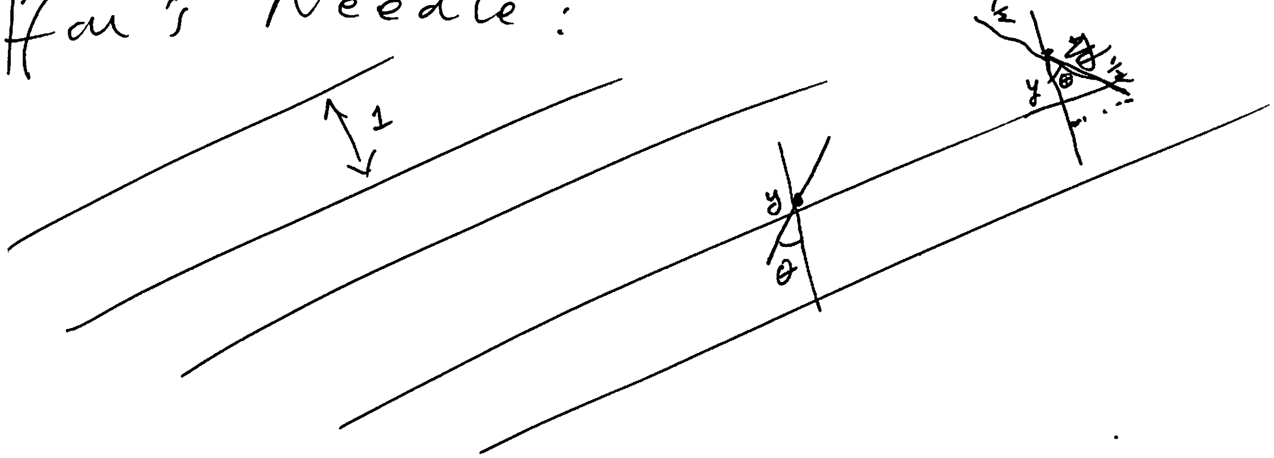


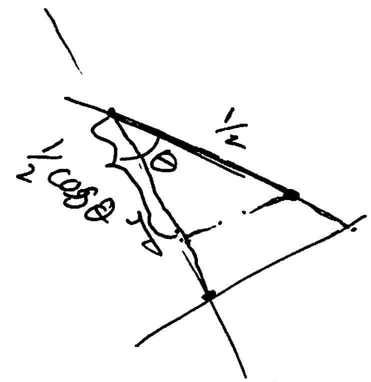


$\sum_j P[X=j] = 1 \iff$ area under curve $= 1$.

$P[a \le X \le b] = \int_a^b f(x)\,dx$.

Buffon's Needle:





$y \in [0, \frac{1}{2}]$. $\theta \in [-\pi/2, \pi/2]$.

Intersects $= \{ \frac{1}{2}\cos\theta \geq y \}$.

P[intersects] = ?

Intersects $= \{ (y,\theta) : y \leq \frac{1}{2}\cos\theta \}$

$f(y,\theta) = \frac{2}{\pi}$ if $0 \leq y \leq 1/2$

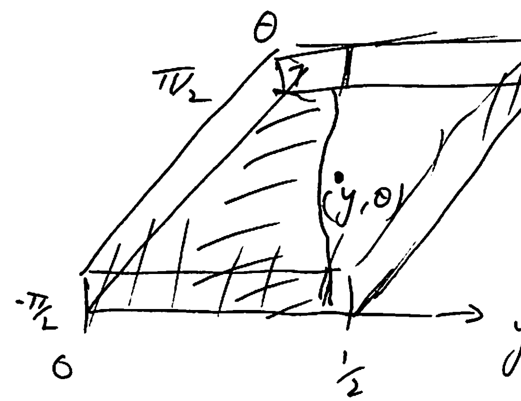
$-\pi/2 \leq \theta \leq \pi/2$.

$$P[\text{intersect}] = \int_{-\pi/2}^{\pi/2} \int_0^{\frac{1}{2}\cos\theta} f(y,\theta)\, dy\, d\theta.$$

$$= \int_{-\pi/2}^{\pi/2} \int_0^{\frac{1}{2}\cos\theta} \frac{2}{\pi}\, dy\, d\theta.$$

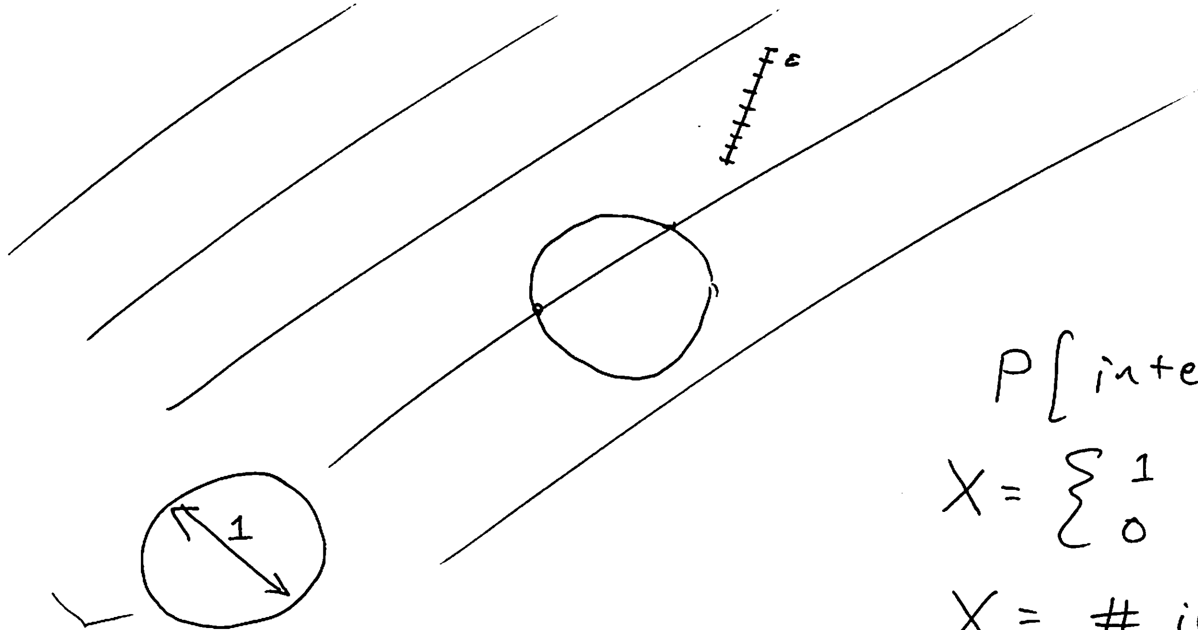
$$= \frac{2}{\pi} \int_{-\pi/2}^{\pi/2} d\theta \int_0^{\frac{1}{2}\cos\theta} dy = \frac{2}{\pi} \int_{-\pi/2}^{\pi/2} \frac{1}{2}\cos\theta$$

$$= \frac{2}{\pi} \cdot \frac{1}{2} \sin\theta \Big|_{-\pi/2}^{\pi/2} = 2/\pi.$$



$\frac{1}{2} \times \pi\, h = 1$.

$h = \frac{2}{\pi}$.



$P[\text{intersect}] = E[X] = a$

$X = \begin{cases} 1 & \text{if intersects} \\ 0 & \text{o.w.} \end{cases}$

$X = \#\ \text{intersections}.$

$$\left.\begin{aligned} X &= X_1 + \cdots + X_{\frac{1}{\varepsilon}} \\ E[X_i] &= \varepsilon E[X] = \varepsilon \cdot a \end{aligned}\right\} \text{linearity of expectation}$$

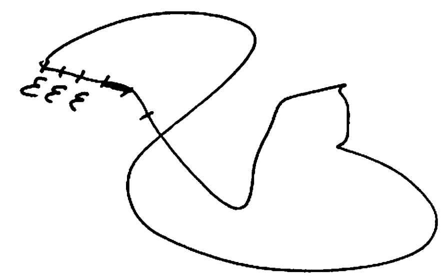
Needle of length ℓ: Y intersections

$E[Y] = \ell \cdot a$.

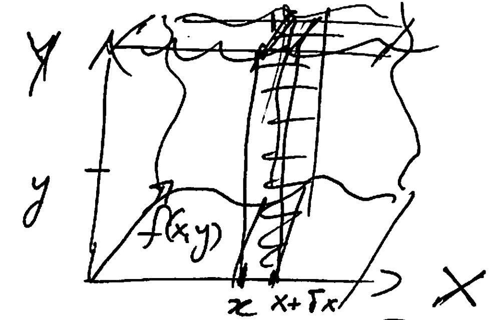
$\ell = \pi$

$2 = E[Y] = \pi \cdot a$

$a = 2/\pi$.



X & Y indep r.v.



$\boxed{f(x,y)}$ joint density funct.

$f_1(x)$ $\quad f_2(y)$

$$\longleftrightarrow \quad P[a \le X \le b \;\&\; c \le Y \le d] = P[a \le X \le b]\; P[c \le Y \le d].$$

$$f_1(x)\,\delta x \quad \cancel{f_1(x)} = \int_{x}^{x+\delta x}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \; = \text{marginal wrt } x.$$

$$\cancel{f_2(y)} = \cancel{\int^{y+\delta y}\int}$$

$$\boxed{f(x,y) = f_1(x)\, f_2(y)}$$

$$P[x \le X \le x+\delta x \;\&\; y \le Y \le y+\delta y] = P[x \le X \le x+\delta x] \cdot P[y \le Y \le y+\delta y].$$

$$= f(x,y)\,\cancel{\delta x}\,\delta y = f_1(x)\,\cancel{\delta x}\; f_2(y)\,\delta y$$

Normal Distribution / Gaussian.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$\mu = E[X]$.
$\sigma^2 = Var$
$\sigma = s.d.$

X mean $= \mu$
var $= \sigma^2$

$$Y = \frac{X-\mu}{\sigma}$$

Y mean $= 0$
var $= 1$.

$$Y = N[0, 1].$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

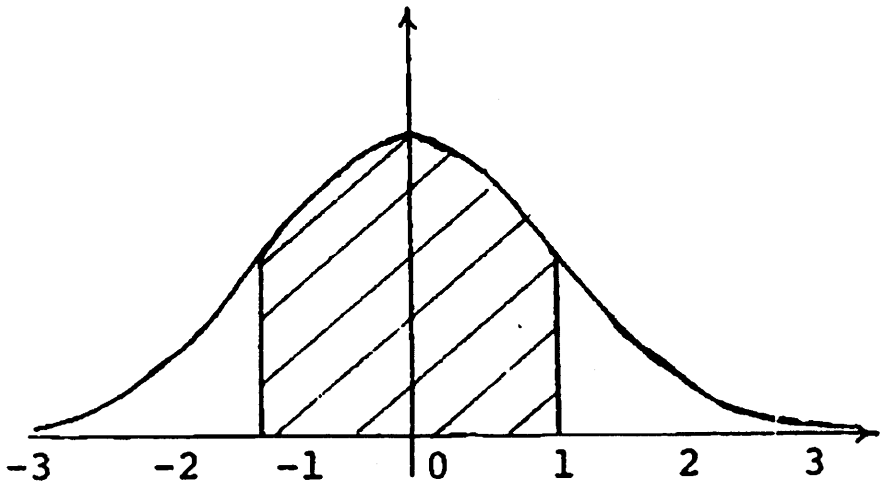
$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 1.$$

$E[Y] = 0$

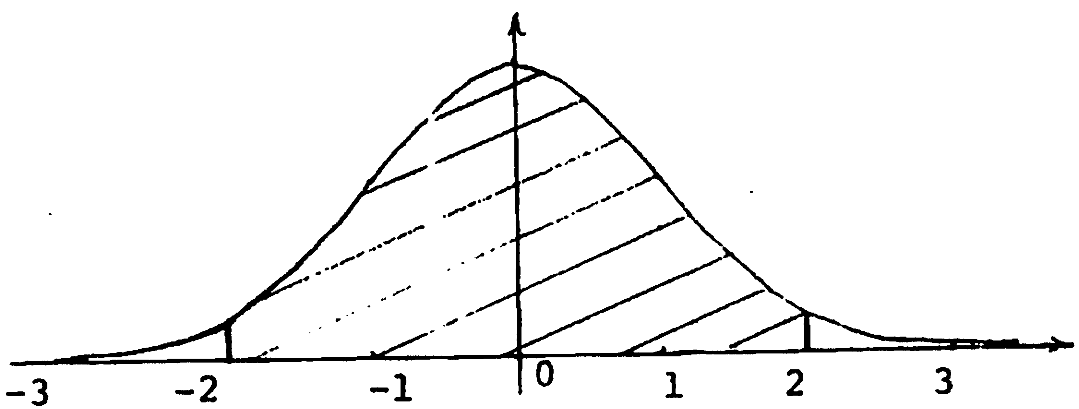
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y\, e^{-y^2/2} dy = 0 = \frac{1}{\sqrt{2\pi}}\underbrace{\int_{-\infty}^{0} \text{—}}_{=-A} + \underbrace{\int_{0}^{\infty} \text{—}}_{=A}$$

$= 0$

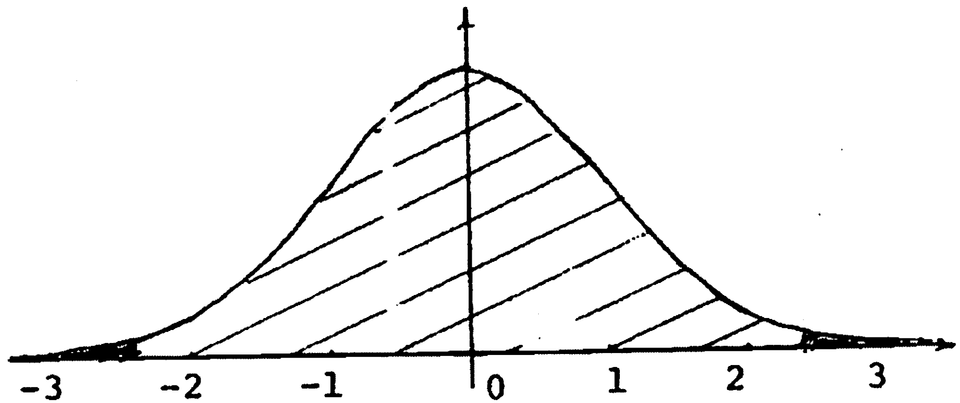
$$E[Y^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \underline{\underline{e^{-y^2/2}}} dy. = 1.$$



The area within one standard deviation of the mean is 68.27% of the total area, i.e. $P(-1 \leq X \leq 1) = .6827$.



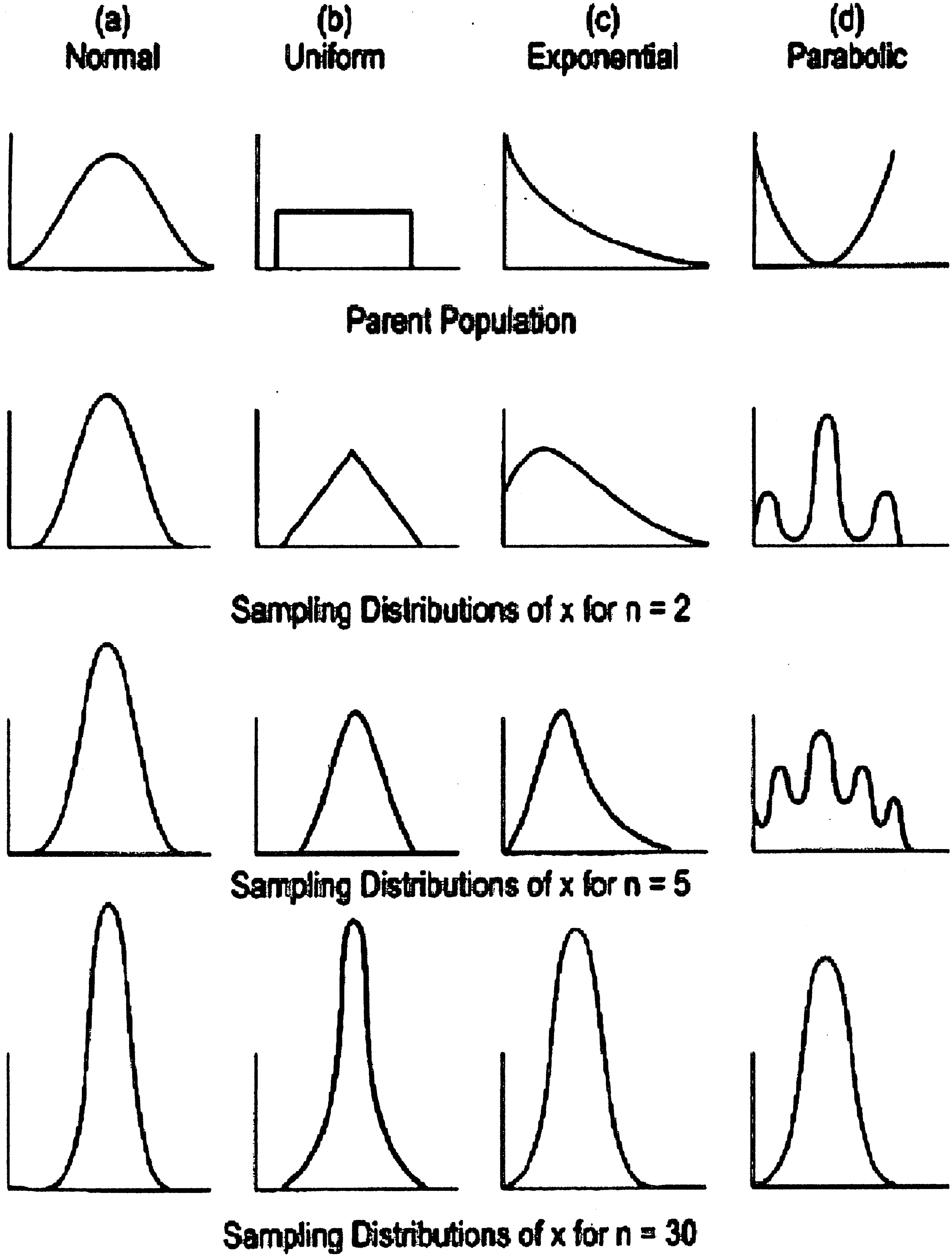
The area within two standard deviations of the mean is 95.45%, of the total area, i.e. $P(-2 \leq X \leq 2) = .9545$.

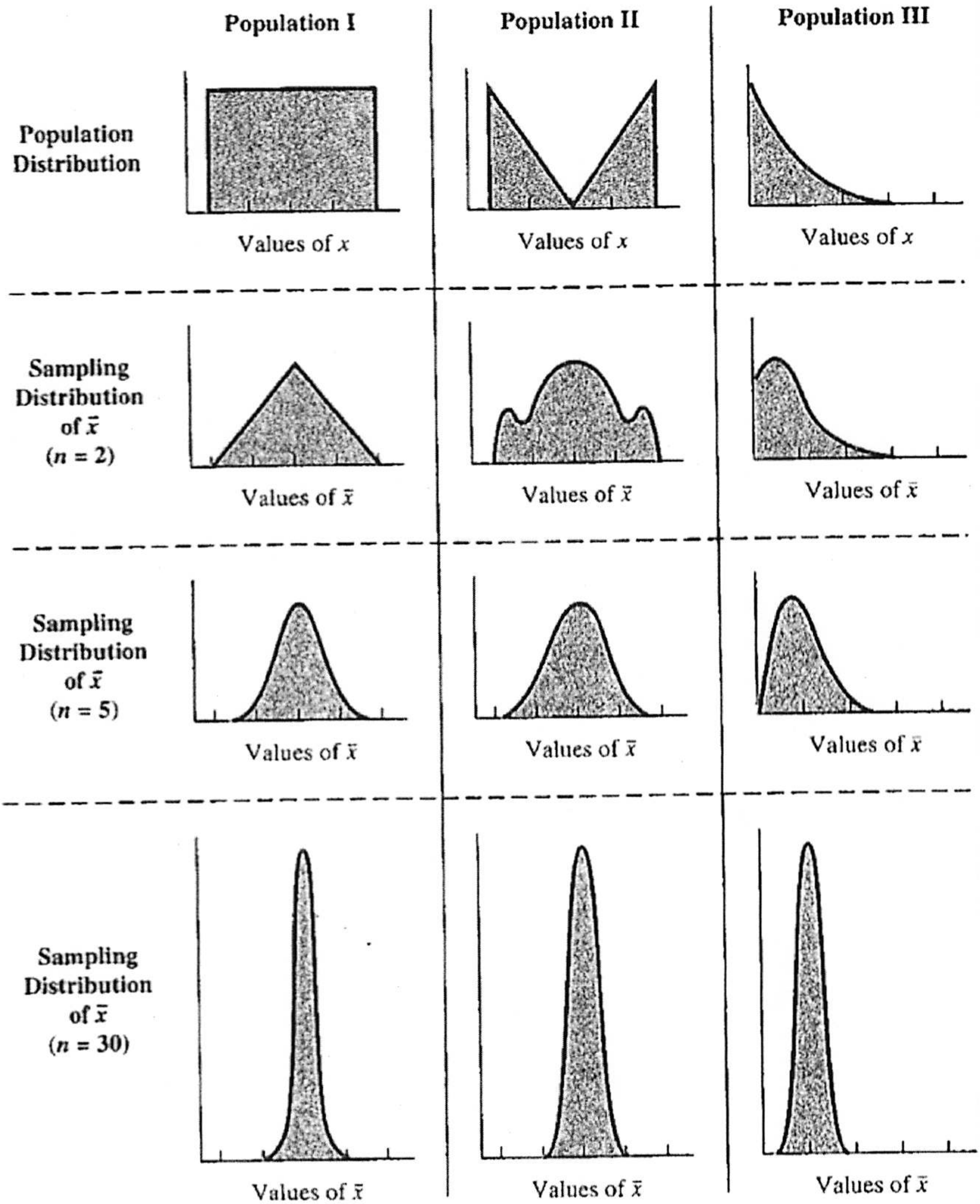


The area within three standard deviations of the mean is 99.73% of the total area, i.e. $P(-3 \leq X \leq 3) = .9973$.

The area within 1.96 standard deviations of the mean is 0.95

The area within 2.58 standard deviations of the mean is 0.99


(a)
Normal
(b)
Uniform
(c)
Exponential
(d)
Parabolic
Parent Population
Sampling Distributions of x for n = 2
Sampling Distributions of x for n = 5
Sampling Distributions of x for n = 30


Population I
Population II
Population III
Population Distribution
Values of x
Values of x
Values of x
Sampling Distribution of $\bar{x}$ ($n = 2$)
Values of $\bar{x}$
Values of $\bar{x}$
Values of $\bar{x}$
Sampling Distribution of $\bar{x}$ ($n = 5$)
Values of $\bar{x}$
Values of $\bar{x}$
Values of $\bar{x}$
Sampling Distribution of $\bar{x}$ ($n = 30$)
Values of $\bar{x}$
Values of $\bar{x}$
Values of $\bar{x}$

Central limit theorem:

$X_1, X_2, \ldots, X_n$ any distr. with mean μ and var σ^2.

$$Y = \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n \to \infty]{} N\left[\mu, \frac{\sigma^2}{n}\right]$$

Significance levels:

$\underbrace{\text{Told coin is fair}}_{S_{100}}$: Toss 100 times. # H's = $\underline{\underline{41}}$.

S_{100} $\quad E[S_{100}] = 50 \quad Var(S_n) = 100 \times \frac{1}{2} \times \frac{1}{2} = 25.$

$\sigma = 5.$

$$\frac{S_{100} - 50}{5} \sim N[0, 1]$$

$$P\left[-1.96 \le \frac{S_{100} - 50}{5} \le 1.96\right] = \underline{\underline{.95}}. \leftarrow \text{significant.}$$

$\underline{\underline{.99}} \leftarrow$ very significant.

$$P\left[40.2 \le S_{100} \le 59.8\right] = .95.$$

Outcome of experiment was not significant in ~~Exp~~ trying to disprove hypothesis.

* Correlation vs causation.

+ve correlation taxes & test scores.

X

* Simpson's paradox:

1973 U.C. Berkeley.

44% male

30% female.

Department	#male applicants	#female applicants	%male admit	%female admit
A	825	108	62	82
B	560	25	63	68
C	325	593	37	34
D	417	375	33	35