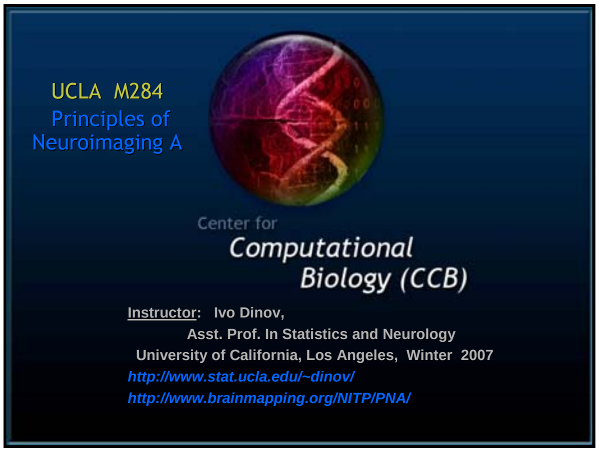
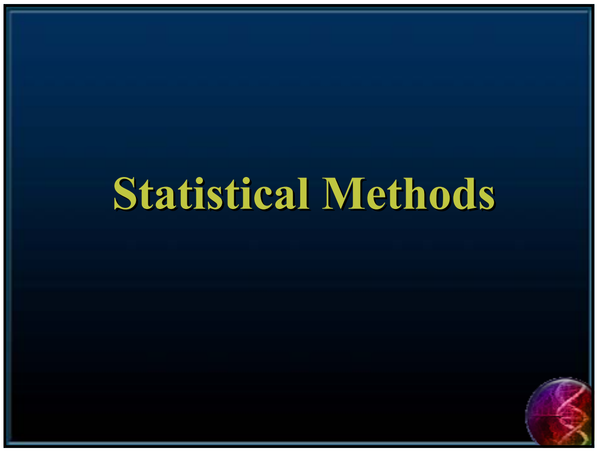


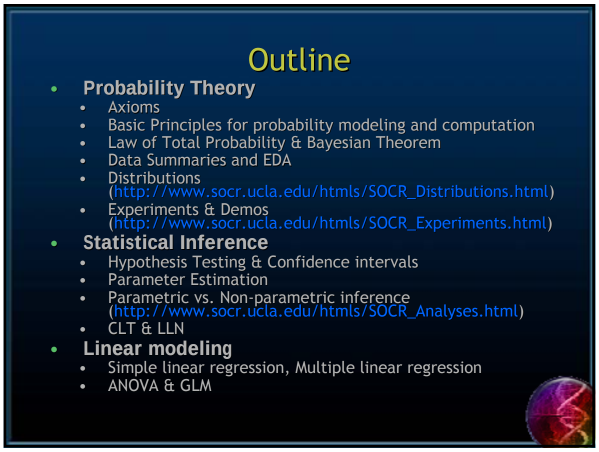
UCLA M284
Principles of Neuroimaging A

Center for
Computational
Biology (CCB)

Instructor: **Ivo Dinov,**
Asst. Prof. In Statistics and Neurology
University of California, Los Angeles, Winter 2007
http://www.stat.ucla.edu/~dinov/
http://www.brainmapping.org/NITP/PNA/

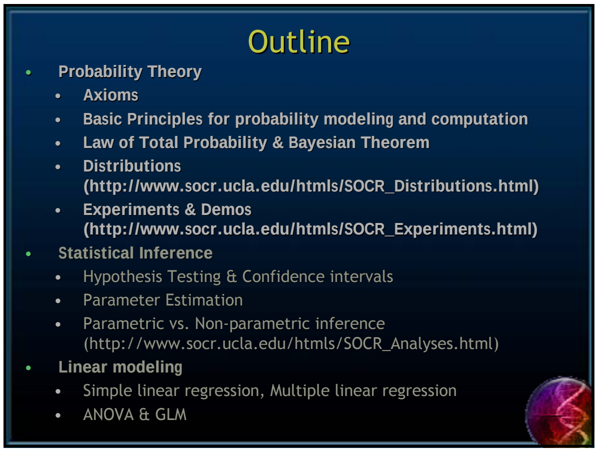


Statistical Methods



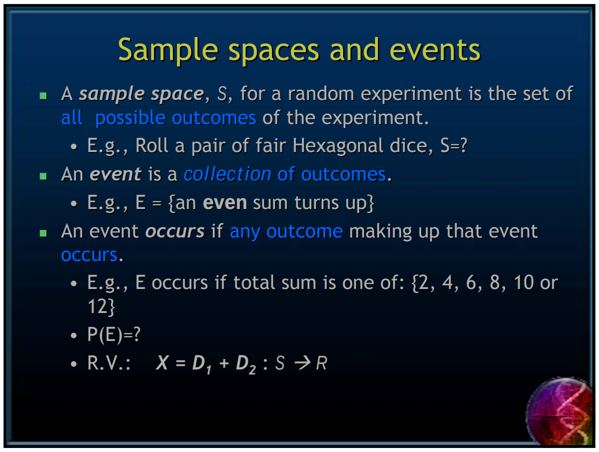
Outline

- **Probability Theory**
  - Axioms
  - Basic Principles for probability modeling and computation
  - Law of Total Probability & Bayesian Theorem
  - Data Summaries and EDA
  - Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
  - Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)
- **Statistical Inference**
  - Hypothesis Testing & Confidence intervals
  - Parameter Estimation
  - Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)
  - CLT & LLN
- **Linear modeling**
  - Simple linear regression, Multiple linear regression
  - ANOVA & GLM



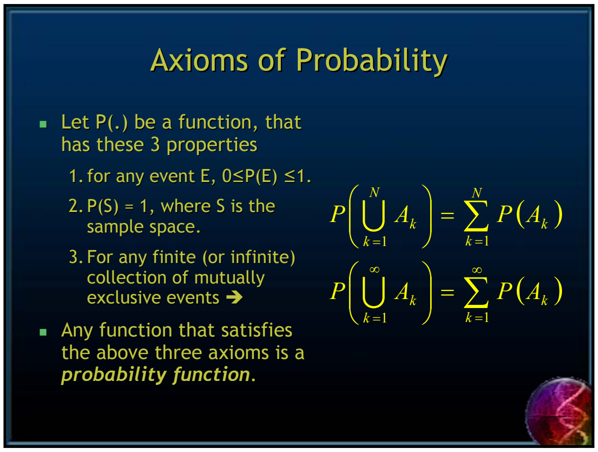
Outline

- **Probability Theory**
  - **Axioms**
  - **Basic Principles for probability modeling and computation**
  - **Law of Total Probability & Bayesian Theorem**
  - **Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)**
  - **Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)**
- Statistical Inference
  - Hypothesis Testing & Confidence intervals
  - Parameter Estimation
  - Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)
- Linear modeling
  - Simple linear regression, Multiple linear regression
  - ANOVA & GLM



Sample spaces and events

- A ***sample space***, S, for a random experiment is the set of all possible outcomes of the experiment.
  - E.g., Roll a pair of fair Hexagonal dice, S=?
- An ***event*** is a *collection* of outcomes.
  - E.g., E = {an **even** sum turns up}
- An event ***occurs*** if any outcome making up that event occurs.
  - E.g., E occurs if total sum is one of: {2, 4, 6, 8, 10 or 12}
  - P(E)=?
  - R.V.: $X = D_1 + D_2 : S \rightarrow R$



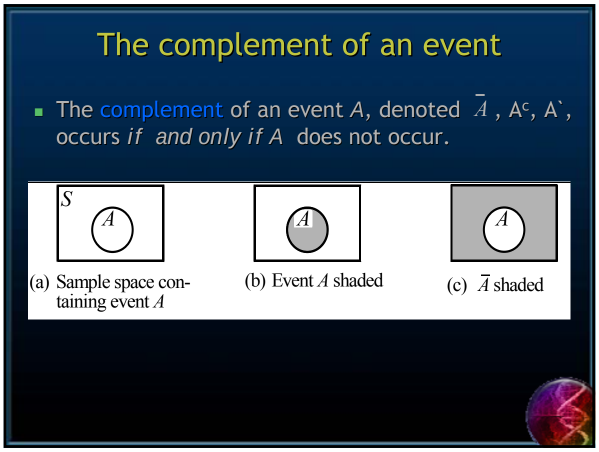
Axioms of Probability

- Let P(.) be a function, that has these 3 properties
  1. for any event E, 0≤P(E) ≤1.
  2. P(S) = 1, where S is the sample space.
  3. For any finite (or infinite) collection of mutually exclusive events →

$$P\left(\bigcup_{k=1}^{N} A_k\right) = \sum_{k=1}^{N} P(A_k)$$

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

- Any function that satisfies the above three axioms is a ***probability function***.



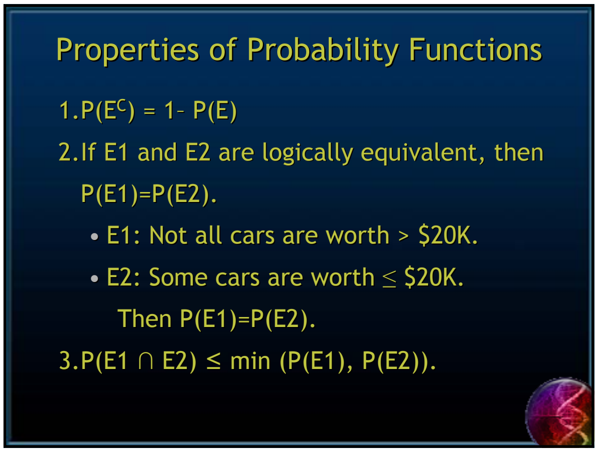
The complement of an event

- The complement of an event *A*, denoted $\bar{A}$, A^c, A`, occurs *if and only if A* does not occur.

(a) Sample space containing event *A*

(b) Event *A* shaded

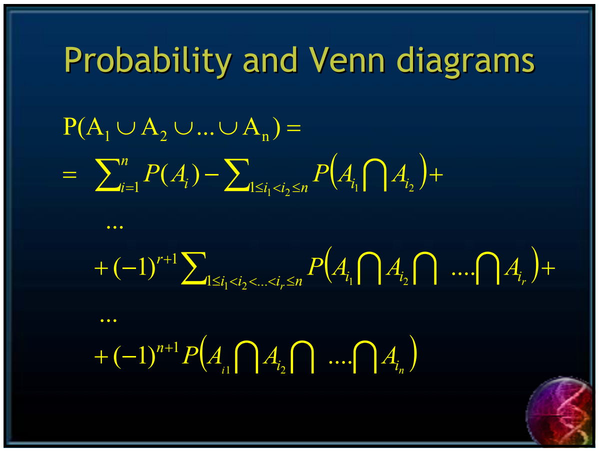
(c) $\bar{A}$ shaded



Properties of Probability Functions

1. P(E^C) = 1- P(E)
2. If E1 and E2 are logically equivalent, then P(E1)=P(E2).
   - E1: Not all cars are worth > $20K.
   - E2: Some cars are worth ≤ $20K.

   Then P(E1)=P(E2).
3. P(E1 ∩ E2) ≤ min (P(E1), P(E2)).



Probability and Venn diagrams

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) =$$

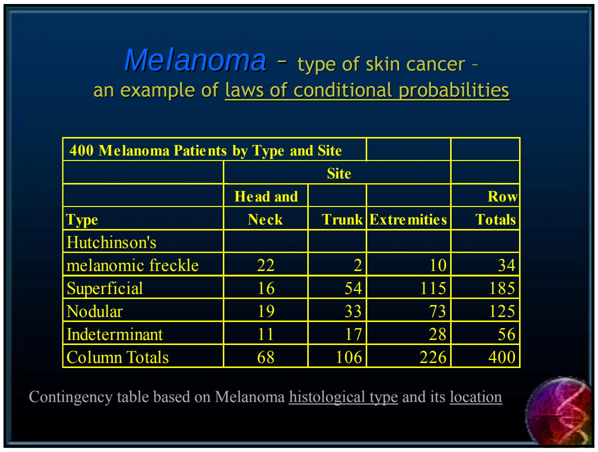
$$= \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i_1 < i_2 \le n} P\left(A_{i_1} \bigcap A_{i_2}\right) +$$

$$\ldots$$

$$+ (-1)^{r+1} \sum_{1 \le i_1 < i_2 < \ldots < i_r \le n} P\left(A_{i_1} \bigcap A_{i_2} \bigcap \ \ldots . \bigcap A_{i_r}\right) +$$

$$\ldots$$

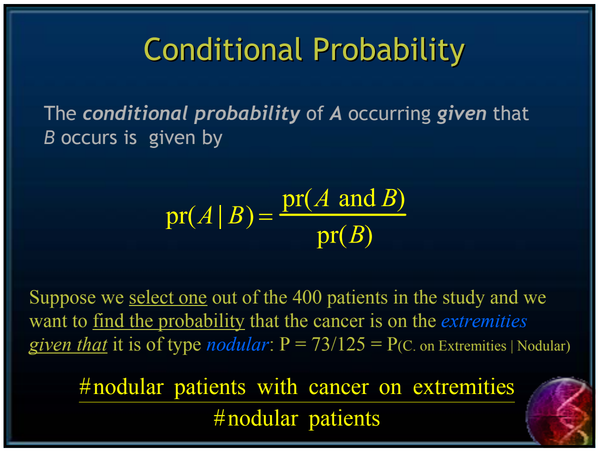
$$+ (-1)^{n+1} P\left(A_{i1} \bigcap A_{i_2} \bigcap \ \ldots . \bigcap A_{i_n}\right)$$



Melanoma – type of skin cancer – an example of laws of conditional probabilities

400 Melanoma Patients by Type and Site				
	Site			
	Head and			Row
Type	Neck	Trunk	Extremities	Totals
Hutchinson's				
melanomic freckle	22	2	10	34
Superficial	16	54	115	185
Nodular	19	33	73	125
Indeterminant	11	17	28	56
Column Totals	68	106	226	400

Contingency table based on Melanoma histological type and its location



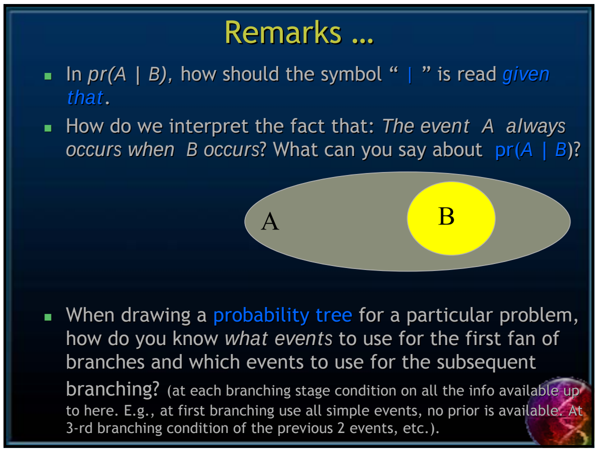
Conditional Probability

The ***conditional probability*** of ***A*** occurring ***given*** that *B* occurs is given by

$$\mathrm{pr}(A \mid B) = \frac{\mathrm{pr}(A \text{ and } B)}{\mathrm{pr}(B)}$$

Suppose we select one out of the 400 patients in the study and we want to find the probability that the cancer is on the *extremities* *given that* it is of type *nodular*: P = 73/125 = P(C. on Extremities | Nodular)

$$\frac{\#\text{nodular patients with cancer on extremities}}{\#\text{nodular patients}}$$



Remarks …

- In *pr(A | B)*, how should the symbol " | " is read *given that*.
- How do we interpret the fact that: *The event A always occurs when B occurs*? What can you say about pr(*A* | *B*)?

- When drawing a probability tree for a particular problem, how do you know *what events* to use for the first fan of branches and which events to use for the subsequent branching? (at each branching stage condition on all the info available up to here. E.g., at first branching use all simple events, no prior is available. At 3-rd branching condition of the previous 2 events, etc.).

Statistical independence

- Events *A* and *B* are *statistically independent* if knowing whether *B* has occurred gives <u>no new information</u> about the chances of *A* occurring,

  i.e. if **pr(*A* | *B*) = pr(*A*)**
- Similarly, P(*B* | *A*) = P(*B*), since

  $P(B|A)=P(B \& A)/P(A) = P(\cancel{A|B})P(B)/\cancel{P(A)} = P(B)$
- If *A* and *B* are *statistically independent*, then

$$\mathrm{pr}(A \text{ and } B) = \mathrm{pr}(A) \times \mathrm{pr}(B)$$

Inverting Conditional Probabilities

$P(A \cap B) = P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$

Formula summary cont.

Multiplication Rule under independence:

- If *A* and *B* are independent events, then

  $P(A \cap B) = P(A)\,P(B)$
- If $A_1, A_2, \ldots, A_n$ are mutually independent,

$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2) \ldots P(A_n)$

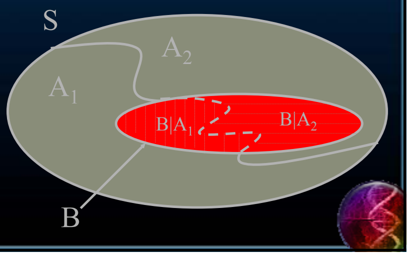
Law of Total Probability

- If $\{A_1, A_2, \ldots, A_n\}$ are a partition of the sample space (mutually exclusive and $\cup A_i = S$) *then for any event B*

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n)$$

Ex:

$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$



Bayesian Rule

- If $\{A_1, A_2, \ldots, A_n\}$ are a non-trivial partition of the sample space (mutually exclusive and $\cup A_i = S$, $P(A_i)>0$) *then for any non-trivial event B (P(B)>0)*

$P(A_i \mid B) = P(A_i \cap B) / P(B) = [P(B \mid A_i) \times P(A_i)] / P(B)$ =

$$= \frac{P(B \mid A_i) \times P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k) P(A_k)}$$

Bayesian Rule

$$P(A_i) = \frac{P(A_i \mid B) \times P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k) P(A_k)}$$

D = the test person has the disease.

T = the test result is positive.

Ex: (Laboratory blood test)

<u>Assume:</u>

P(positive Test| Disease) = 0.95

P(positive Test| no Disease)=0.01

P(Disease) = 0.005

<u>Find:</u>

P(Disease|positive Test)=?

P(D | T) = ?

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T \mid D) \times P(D)}{P(T \mid D) \times P(D) + P(T \mid D^c) \times P(D^c)}$$

$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = \frac{0.00475}{0.02465} = 0.193$$

Classes vs. Evidence Conditioning

- **Classes**: healthy(NC), cancer
- **Evidence**: positive mammogram (pos), negative mammogram (neg)
- If a woman has a positive mammogram result, what is the probability that she has breast cancer?

$$P(class \mid evidence) = \frac{P(evidence \mid class) \times P(class)}{\sum_{classes} P(evidence \mid class) \times P(class)}$$

$$P(cancer\ ) = 0.01$$

$$P(pos \mid cancer\ ) = 0.8$$

$$P(pos \mid healthy) = 0.1$$

$$P(cancer\ \mid pos\ ) = ?$$

P(C|P)=P(P|C)xP(C)/[P(P|C)xP(C)+ P(P|H)xP(H)]

P(C|P)=0.8x0.01 / [0.8x0.01 + 0.1x0.99] = ?

Bayesian Rule (different data/example!)

True Disease State

Test Results

	No Disease	Disease	Total
Negative	OK (0.98505)	False Negative II (0.00025)	0.9853
Positive	False Positive I (0.00995)	OK (0.00475)	0.0147
Total	0.995	0.005	1.0

$$P\left(T \bigcap D^C\right) = P\left(T \mid D^C\right) \times P\left(D^C\right) = 0.01 \times 0.995 = 0.00995$$

Power of Test = $1 - P(T^C \mid D)$ = 0.00025/0.005 = 0.95

Sensitivity: TP/(TP+FN) = 0.00475/(0.00475+ 0.00025)= 0.95

Specificity: TN/(TN+FP) = 0.98505/(0.98505+ 0.00995) = 0.99

4 Factors affecting the power

- **Larger:** ➔ **Causes:**
- Sample size (positive)
- Sample variance (negative)
- Effect size (positive)
- The chosen level for α (positive)

EDA

Probability Theory

- Axioms
- Basic Principles for probability modeling and computation
- Law of Total Probability & Bayesian Theorem

Data Summaries and EDA

- Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
- Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)

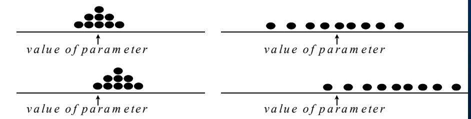
Statistical Inference

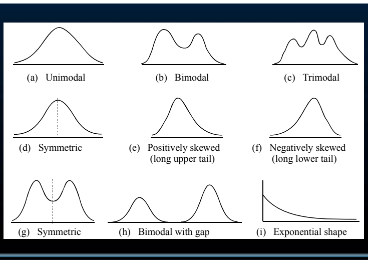
- Hypothesis Testing & Confidence intervals
- Parameter Estimation
- Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)

Linear modeling

- Simple linear regression, Multiple linear regression
- ANOVA & GLM

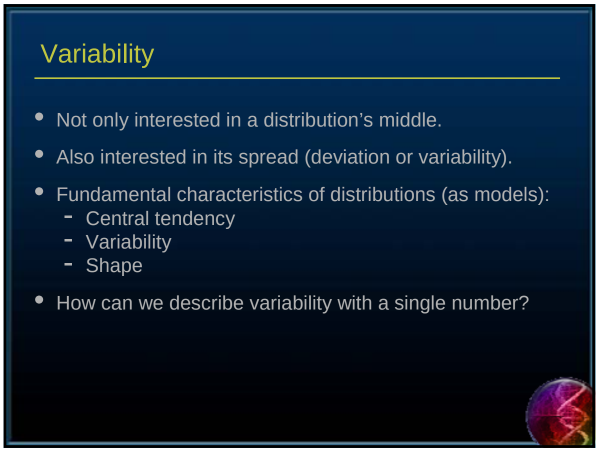
The Big Three: Center, Spread and Shape!





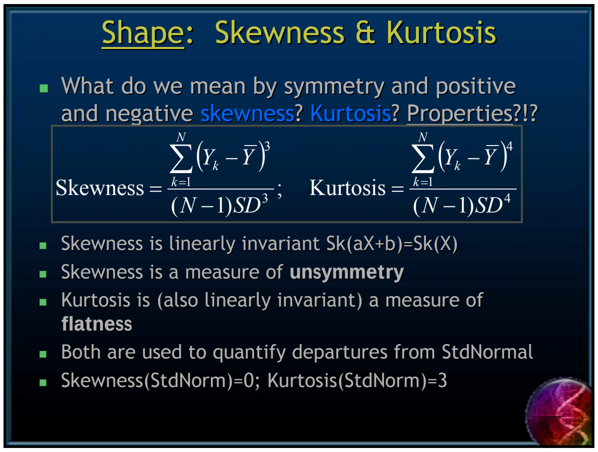
Center: Central tendency, the *middle,* as a single number

- **Mode**: The most frequent score in the distribution.
- **Median**: The centermost score if there are an odd number of scores or the average of the two centermost scores if there are an even number of scores.
- **Mean**: The sum of all (numeric) observations divided by the number of scores (arithmetic average).



Variability

- Not only interested in a distribution's middle.
- Also interested in its spread (deviation or variability).
- Fundamental characteristics of distributions (as models):
  - Central tendency
  - Variability
  - Shape
- How can we describe variability with a single number?

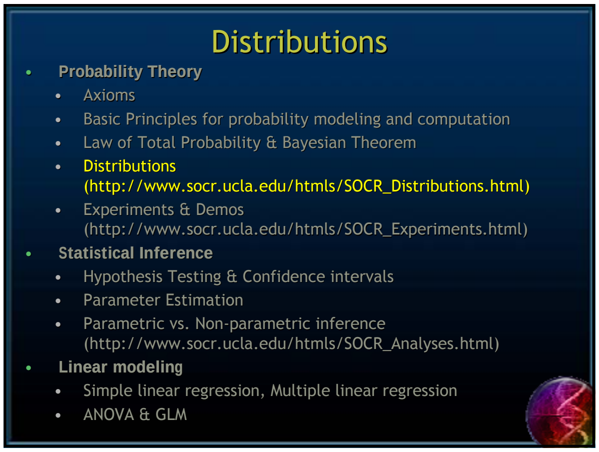


Shape: Skewness & Kurtosis

- What do we mean by symmetry and positive and negative skewness? Kurtosis? Properties?!?

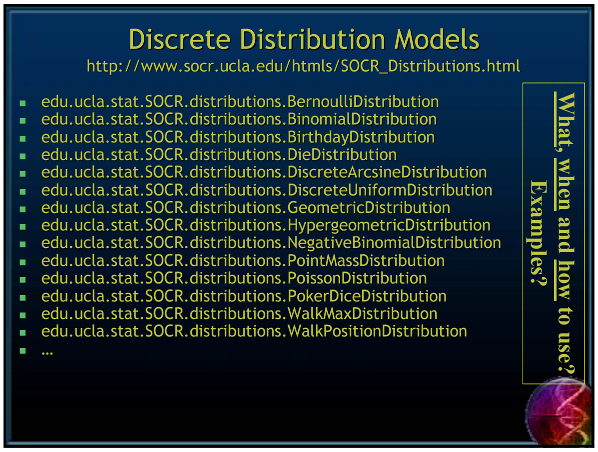
$$\text{Skewness} = \frac{\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^3}{(N-1)SD^3}; \quad \text{Kurtosis} = \frac{\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^4}{(N-1)SD^4}$$

- Skewness is linearly invariant Sk(aX+b)=Sk(X)
- Skewness is a measure of **unsymmetry**
- Kurtosis is (also linearly invariant) a measure of **flatness**
- Both are used to quantify departures from StdNormal
- Skewness(StdNorm)=0; Kurtosis(StdNorm)=3



Distributions

- **Probability Theory**
  - Axioms
  - Basic Principles for probability modeling and computation
  - Law of Total Probability & Bayesian Theorem
  - Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
  - Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)
- **Statistical Inference**
  - Hypothesis Testing & Confidence intervals
  - Parameter Estimation
  - Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)
- **Linear modeling**
  - Simple linear regression, Multiple linear regression
  - ANOVA & GLM

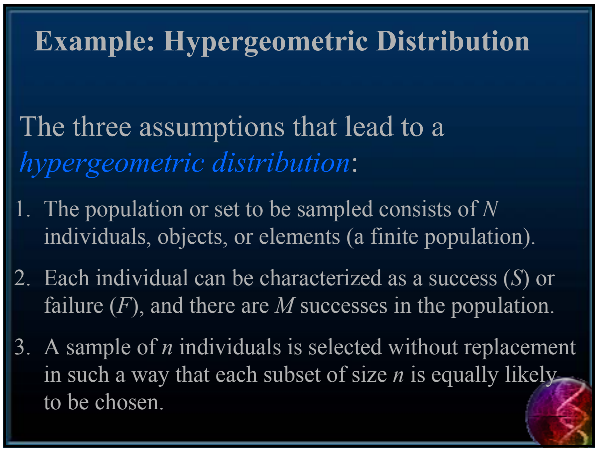


Discrete Distribution Models

http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

- edu.ucla.stat.SOCR.distributions.BernoulliDistribution
- edu.ucla.stat.SOCR.distributions.BinomialDistribution
- edu.ucla.stat.SOCR.distributions.BirthdayDistribution
- edu.ucla.stat.SOCR.distributions.DieDistribution
- edu.ucla.stat.SOCR.distributions.DiscreteArcsineDistribution
- edu.ucla.stat.SOCR.distributions.DiscreteUniformDistribution
- edu.ucla.stat.SOCR.distributions.GeometricDistribution
- edu.ucla.stat.SOCR.distributions.HypergeometricDistribution
- edu.ucla.stat.SOCR.distributions.NegativeBinomialDistribution
- edu.ucla.stat.SOCR.distributions.PointMassDistribution
- edu.ucla.stat.SOCR.distributions.PoissonDistribution
- edu.ucla.stat.SOCR.distributions.PokerDiceDistribution
- edu.ucla.stat.SOCR.distributions.WalkMaxDistribution
- edu.ucla.stat.SOCR.distributions.WalkPositionDistribution
- …

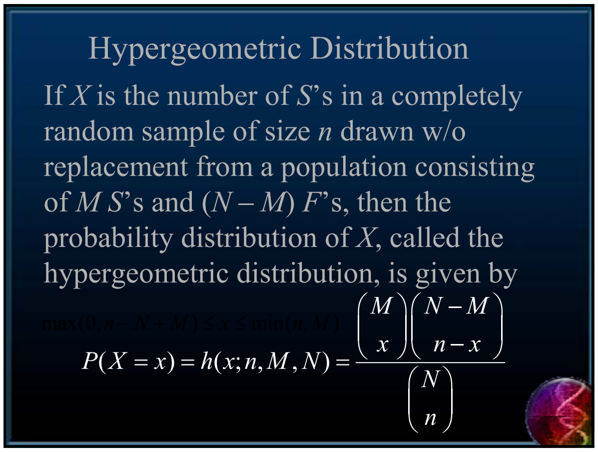
What, when and how to use? Examples?



Example: Hypergeometric Distribution

The three assumptions that lead to a *hypergeometric distribution*:

1. The population or set to be sampled consists of N individuals, objects, or elements (a finite population).
2. Each individual can be characterized as a success (S) or failure (F), and there are M successes in the population.
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.



Hypergeometric Distribution

If X is the number of S's in a completely random sample of size n drawn w/o replacement from a population consisting of M S's and $(N-M)$ F's, then the probability distribution of X, called the hypergeometric distribution, is given by

$$\max(0, n-N+M) \le x \le \min(n, M)$$

$$P(X=x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Hypergeometric Mean and Variance

$$E(X) = n \cdot \frac{M}{N} \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)$$

Ball_and_Urn_Experiment – HyperGeometric Distribution & Binomial Approximation to HyperGeometric
http://www.socr.ucla.edu/htmls/SOCR_Experiments.html

Computation of a Binomial pmf

$$b(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$0 \le x \le n$$

Hypergeometric Distribution & Binomial

- Binomial approximation to Hypergeometric

$$\frac{n}{N} \text{ is small (usually } < 0.1\text{), then } \frac{M}{N} \approx p$$

$$HyperGeom(x;N,n,M) \underset{M/N=p}{\overset{approaches}{\Longrightarrow}} Bin(x;n,p)$$

Ex: 4,000 out of 10,000 residents are against a new tax. 15 residents are selected at random and surveyed.

P(at most 7 favor the new tax) = ?

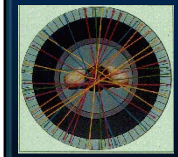
http://socr.stat.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm

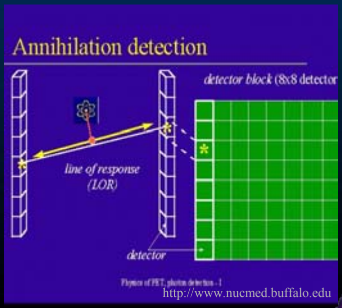
HyperGeometric and Binomial Experiment/Distributions

Poisson Distribution - Definition

- Used to model counts - number of arrivals (k) on a given interval …
- The Poisson distribution is also sometimes referred to as the **distribution of rare events**. Examples of Poisson distributed variables are number of accidents per person, number of sweepstakes won per person, or the number of catastrophic defects found in a production process.

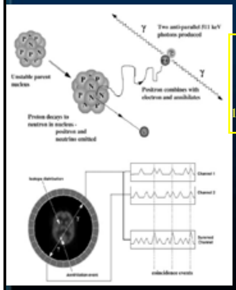
Functional Brain Imaging - Positron Emission Tomography (PET)





http://www.nucmed.buffalo.edu

Functional Brain Imaging - Positron Emission Tomography (PET)



Isotope	Energy (MeV)	Range(mm)	1/2-life	Appl.
^{11}C	0.96	1.1	20 min	receptors
^{15}O	1.7	1.5	2 min	stroke/activation
^{18}F	0.6	1.0	110 min	neurology
^{124}I	~2.0	1.6	4.5 days	oncology

Poisson Distribution - Mean

- Used to model counts - number of arrivals (k) on a given interval …
- Y~Poisson(λ), then P(Y=k) = $\frac{\lambda^k e^{-\lambda}}{k!}$, k = 0, 1, 2, …
- Mean of Y, $\mu_Y = \lambda$, since

$$E(Y)=\sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} =$$

$$=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

Poisson Distribution - Variance

- Y~Poisson(λ), then P(Y=k) = $\frac{\lambda^k e^{-\lambda}}{k!}$, k = 0, 1, 2, …
- Variance of Y, $\sigma_Y = \lambda^{1/2}$, since

$$\sigma_Y^2 = Var(Y)=\sum_{k=0}^{\infty}(k-\lambda)^2\frac{\lambda^k e^{-\lambda}}{k!} = \ldots = \lambda$$

- For example, suppose that Y denotes the number of blocked shots (arrivals) in a randomly sampled game for the UCLA Bruins men's basketball team. Then a Poisson distribution with mean=4 *may be* used to model Y .

Poisson as an ***approximation*** to Binomial

- Suppose we have a sequence of Binomial(n, p_n) models, with lim(n p_n) → λ, as n→infinity.
- For each 0<=y<=n, if Y_n~ Binomial(n, p_n), then
  - P(Y_n=y)= $\binom{n}{y}p_n^{\,y}(1-p_n)^{n-y}$
  - But this converges to:

$$\binom{n}{y}p_n^{\,y}(1-p_n)^{n-y}\xrightarrow[\substack{n\longrightarrow\infty\\ n\times p_n\longrightarrow\lambda}]{\text{WHY?}}\frac{\lambda^y e^{-\lambda}}{y!}$$

- Thus, Binomial(n, p_n) → Poisson(λ)

Poisson as an ***approximation*** to Binomial

- **Rule of thumb** is that approximation is good if:
  - **n>=100**
  - **p<=0.01**
  - **λ =n p <=20**
- Then, Binomial(n, p_n) → Poisson(λ)
- Validate using:
- http://www.socr.ucla.edu/htmls/SOCR_Experiments.html
- Binomial, HyperGeometric and Poisson Experiments

Example using Poisson approx to Binomial

- Suppose P(Disease) = 0.0001=10^{-4}. Find the probability that a village of 25,000 people has > 2 people having the disease!
- Y~ Binomial(25,000, 0.0001) , find P(Y>2). Note that Z~Poisson(λ =n p =25,000 x 0.0001=2.5)

$$P(Z>2) = 1-P(Z\le 2) = 1-\sum_{z=0}^{2}\frac{2.5^z}{z!}e^{-2.5} =$$

$$1-\left(\frac{2.5^0}{0!}e^{-2.5}+\frac{2.5^1}{1!}e^{-2.5}+\frac{2.5^2}{2!}e^{-2.5}\right) = 0.456$$

Why bother discussing distributions?

- Provide a rich source of (analytical) Models.
- General properties of processes may be studies without regard to the underlying molecular, physiological, genotypic or phenotypic properties or characteristics of the phenomenon.
- Easy to fit models to data and make inference using the model instead of limited data.
- Low computational costs (efficiency)
- What else?
- **Example:** http://www.socr.ucla.edu/htmls/SOCR_Modeler.html

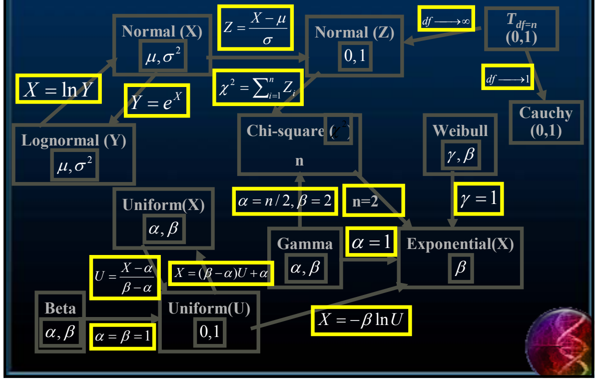
Continuous Distribution Models

http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

- Beta Distribution, edu.ucla.stat.SOCR.distributions.BetaDistribution
- Beta (Generalized) Distribution, edu.ucla.stat.SOCR.distributions.BetaGeneralDistribution
- CauchyDistribution, edu.ucla.stat.SOCR.distributions.CauchyDistribution
- Chi-Square Distribution, edu.ucla.stat.SOCR.distributions.ChiSquareDistribution
- Circle Distribution, edu.ucla.stat.SOCR.distributions.CircleDistribution
- Continuous Uniform Distribution, edu.ucla.stat.SOCR.distributions.ContinuousUniformDistribution
- Exponential Distribution, edu.ucla.stat.SOCR.distributions.ExponentialDistribution
- Fisher's F Distribution, edu.ucla.stat.SOCR.distributions.FisherDistribution
- Gamma Distribution, edu.ucla.stat.SOCR.distributions.GammaDistribution
- GeneralCauchyDistribution, edu.ucla.stat.SOCR.distributions.GeneralCauchyDistribution
- Gilbrats Distribution, edu.ucla.stat.SOCR.distributions.GilbratsDistribution
- GumbelDistribution, edu.ucla.stat.SOCR.distributions.GumbelDistribution
- Half-Normal Distribution, edu.ucla.stat.SOCR.distributions.HalfNormalDistribution
- Laplace Distribution, edu.ucla.stat.SOCR.distributions.LaplaceDistribution
- Logistic Distribution, edu.ucla.stat.SOCR.distributions.LogisticDistribution
- Log-Normal Distribution, edu.ucla.stat.SOCR.distributions.LogNormalDistribution
- Maxwell Distribution, edu.ucla.stat.SOCR.distributions.MaxwellDistribution
- MixtureDistribution, edu.ucla.stat.SOCR.distributions.MixtureDistribution
- Normal Distribution, edu.ucla.stat.SOCR.distributions.NormalDistribution
- Pareto Distribution, edu.ucla.stat.SOCR.distributions.ParetoDistribution
- Rayleigh Distribution, edu.ucla.stat.SOCR.distributions.RayleighDistribution
- Student's T Distribution, edu.ucla.stat.SOCR.distributions.StudentDistribution
- Triangle Distribution, edu.ucla.stat.SOCR.distributions.TriangleDistribution
- Weibull Distribution, edu.ucla.stat.SOCR.distributions.WeibullDistribution

What, when and how to use? Examples?

Relation among Distributions



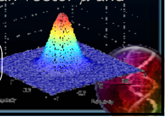
The Normal Distribution

- A normal density curve can be summarized with the following formula:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

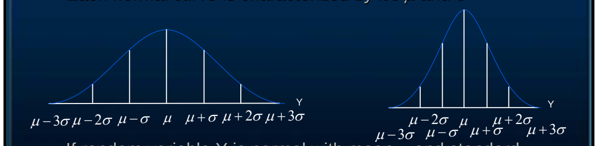
- Every normal curve uses this formula, what makes them different is what gets plugged in for μ and σ
- Each normal curve is centered at μ and the width depends on σ
- (small = tall, large = short/wide).
- *d*-dimensional Gaussian distributions with mean vector μ and covariance matrix Σ:

$$p(\overline{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} Exp\left(-\frac{1}{2}(\overline{x}-\overline{\mu})^T \Sigma^{-1}(\overline{x}-\overline{\mu})\right)$$



The Normal Distribution

- Each normal curve is characterized by it's μ and σ



- If random variable Y is normal with mean μ and standard deviation σ, we write
  - Y ~ N(μ , σ^2)
  - http://www.SOCR.ucla.edu/htmls/SOCR_Distributions.html

Definition of the expected value, in general

- The expected value:

$$E(X) = \sum_{\text{all } x} x\, P(x) \left(= \int_{\text{all } X} x\, P(x) dx \right)$$

- = Sum of (value times probability of value)

Example

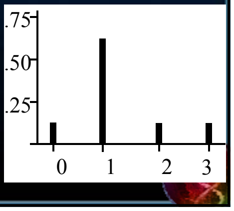
In the at least one of each or at most 3 children example, where X ={number of Girls} we have:

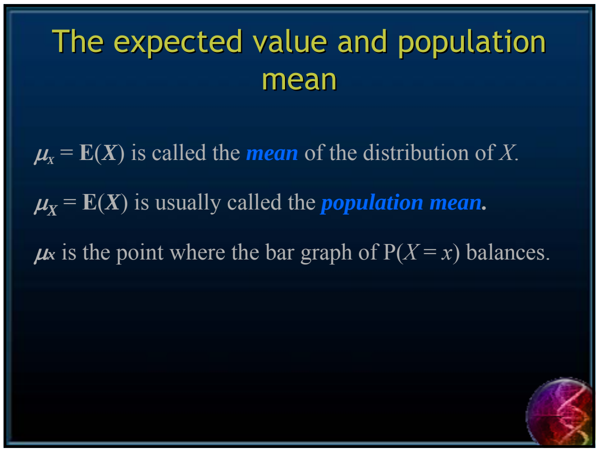
X	0	1	2	3
pr(*x*)	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

$$E(X) = \sum_x x\,P(x)$$

$$= 0 \times \frac{1}{8} + 1 \times \frac{5}{8} + 2 \times \frac{1}{8} + 3 \times \frac{1}{8}$$

$$= 1.25$$



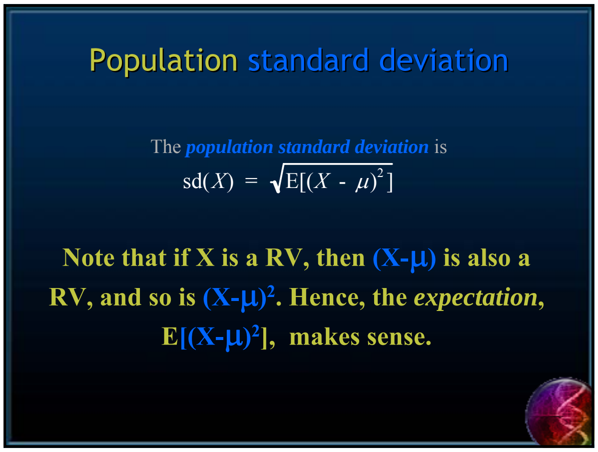


The expected value and population mean

μ_x = **E**(*X*) is called the ***mean*** of the distribution of X.

μ_X = **E**(*X*) is usually called the ***population mean*****.**

μ_x is the point where the bar graph of P($X = x$) balances.

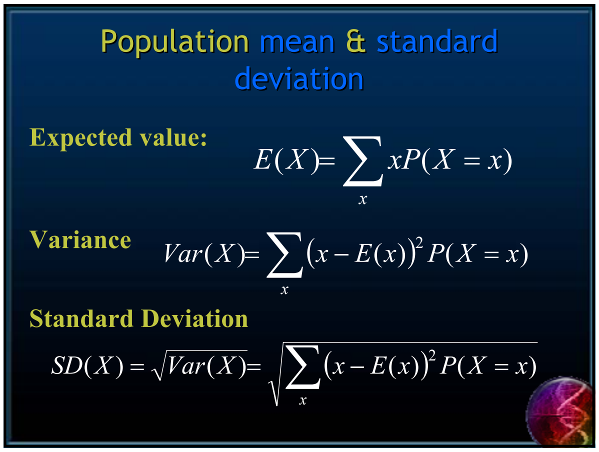


Population standard deviation

The ***population standard deviation*** is

$$\text{sd}(X) = \sqrt{\text{E}[(X - \mu)^2]}$$

Note that if X is a RV, then (X-μ) is also a RV, and so is (X-μ)^2. Hence, the *expectation*, E[(X-μ)^2], makes sense.



Population mean & standard deviation

Expected value:

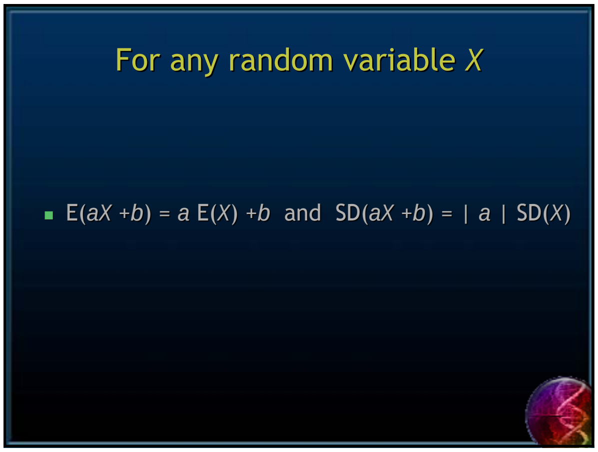
$$E(X) = \sum_x xP(X = x)$$

Variance

$$Var(X) = \sum_x (x - E(x))^2 P(X = x)$$

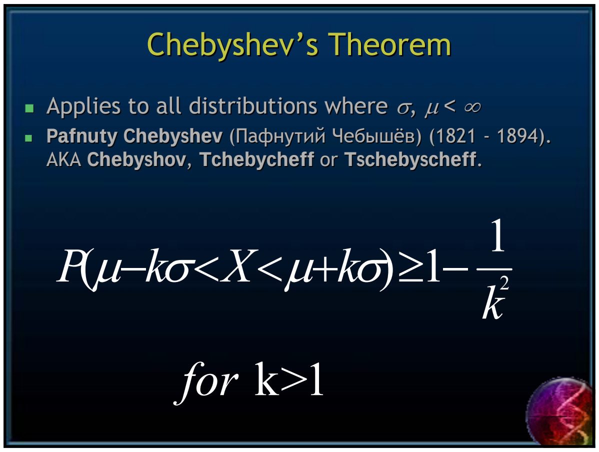
Standard Deviation

$$SD(X) = \sqrt{Var(X)} = \sqrt{\sum_x (x - E(x))^2 P(X = x)}$$



For any random variable *X*

- E($aX + b$) = a E(X) + b and SD($aX + b$) = | a | SD(X)

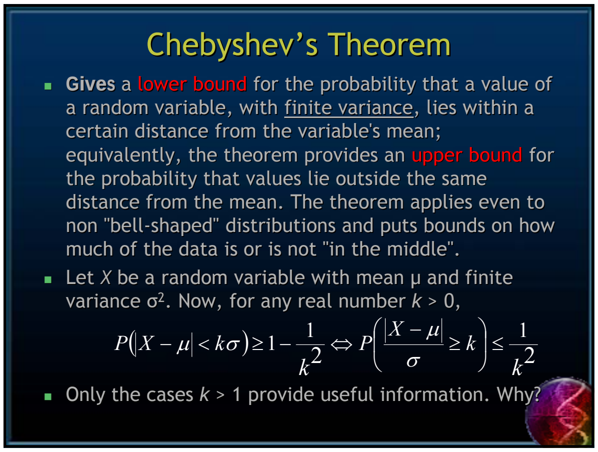


Chebyshev’s Theorem

- Applies to all distributions where $\sigma, \mu < \infty$
- **Pafnuty Chebyshev** (Пафнутий Чебышёв) (1821 - 1894). AKA **Chebyshov**, **Tchebycheff** or **Tschebyscheff**.

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

for k>1

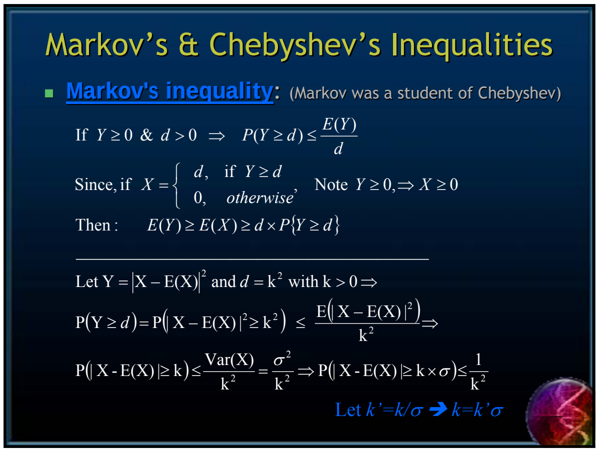


Chebyshev’s Theorem

- **Gives** a lower bound for the probability that a value of a random variable, with finite variance, lies within a certain distance from the variable's mean; equivalently, the theorem provides an upper bound for the probability that values lie outside the same distance from the mean. The theorem applies even to non "bell-shaped" distributions and puts bounds on how much of the data is or is not "in the middle".
- Let *X* be a random variable with mean μ and finite variance σ^2. Now, for any real number $k > 0$,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \Leftrightarrow P\left(\frac{|X - \mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$

- Only the cases $k > 1$ provide useful information. Why?



Markov’s & Chebyshev’s Inequalities

- **Markov's inequality**: (Markov was a student of Chebyshev)

$$\text{If } Y \geq 0 \ \& \ d > 0 \ \Rightarrow \ P(Y \geq d) \leq \frac{E(Y)}{d}$$

$$\text{Since, if } X = \begin{cases} d, & \text{if } Y \geq d \\ 0, & \textit{otherwise} \end{cases}, \quad \text{Note } Y \geq 0, \Rightarrow X \geq 0$$

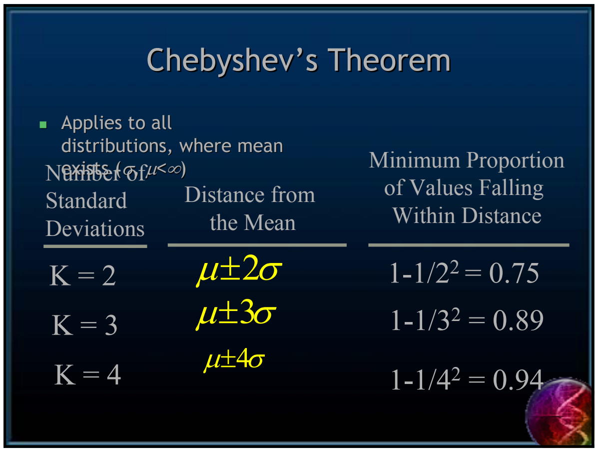
$$\text{Then}: \quad E(Y) \geq E(X) \geq d \times P\{Y \geq d\}$$

$$\text{Let Y} = |\text{X} - \text{E(X)}|^2 \text{ and } d = \text{k}^2 \text{ with k} > 0 \Rightarrow$$

$$\text{P}(\text{Y} \geq d) = \text{P}(|\text{X} - \text{E(X)}|^2 \geq \text{k}^2) \ \leq \ \frac{\text{E}(|\text{X} - \text{E(X)}|^2)}{\text{k}^2} \Rightarrow$$

$$\text{P}(|\text{X - E(X)}| \geq \text{k}) \leq \frac{\text{Var(X)}}{\text{k}^2} = \frac{\sigma^2}{\text{k}^2} \Rightarrow \text{P}(|\text{X - E(X)}| \geq \text{k} \times \sigma) \leq \frac{1}{\text{k}^2}$$

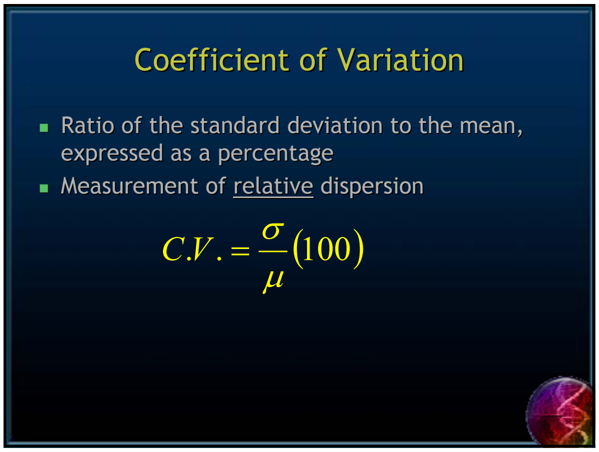
Let $k' = k/\sigma$ ➔ $k = k'\sigma$



Chebyshev’s Theorem

- Applies to all distributions, where mean exists ($\sigma, \mu < \infty$)

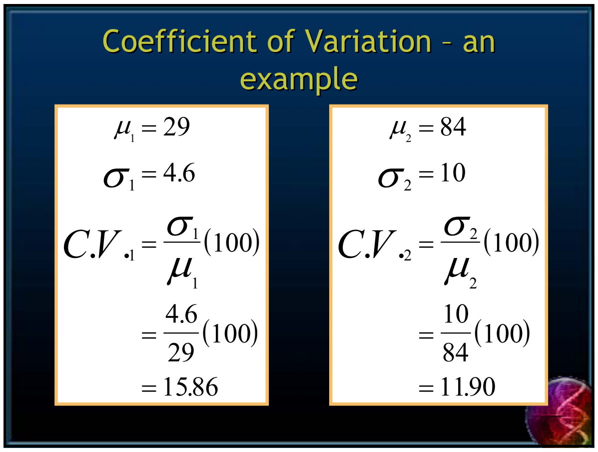
Number of Standard Deviations	Distance from the Mean	Minimum Proportion of Values Falling Within Distance
K = 2	$\mu \pm 2\sigma$	$1\text{-}1/2^2 = 0.75$
K = 3	$\mu \pm 3\sigma$	$1\text{-}1/3^2 = 0.89$
K = 4	$\mu \pm 4\sigma$	$1\text{-}1/4^2 = 0.94$



Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

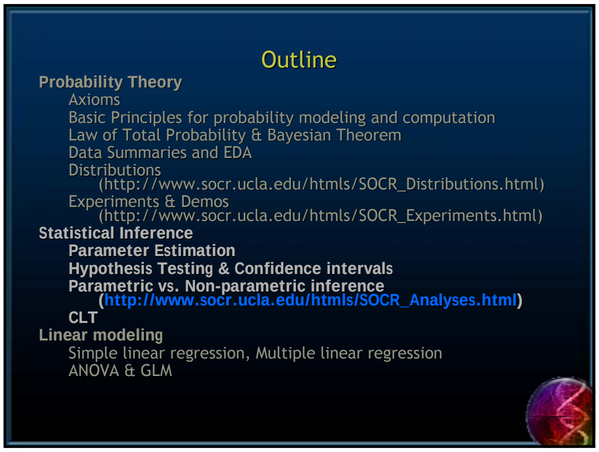
$$C.V. = \frac{\sigma}{\mu}(100)$$



Coefficient of Variation - an example

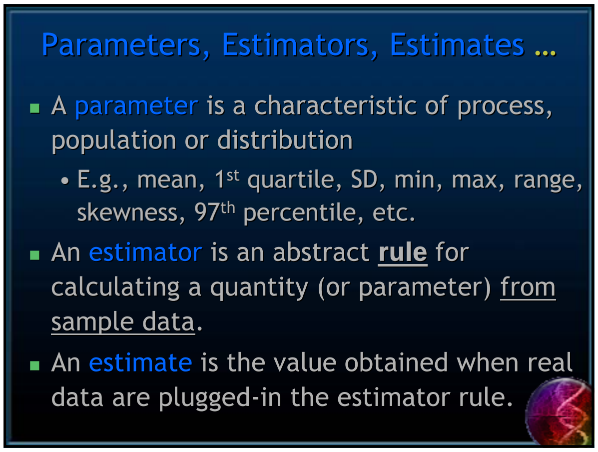
$$\mu_1 = 29$$
$$\sigma_1 = 4.6$$
$$C.V._1 = \frac{\sigma_1}{\mu_1}(100) = \frac{4.6}{29}(100) = 15.86$$

$$\mu_2 = 84$$
$$\sigma_2 = 10$$
$$C.V._2 = \frac{\sigma_2}{\mu_2}(100) = \frac{10}{84}(100) = 11.90$$



Outline

- **Probability Theory**
  - Axioms
  - Basic Principles for probability modeling and computation
  - Law of Total Probability & Bayesian Theorem
  - Data Summaries and EDA
  - Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
  - Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)
- **Statistical Inference**
  - **Parameter Estimation**
  - **Hypothesis Testing & Confidence intervals**
  - **Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)**
  - **CLT**
- **Linear modeling**
  - Simple linear regression, Multiple linear regression
  - ANOVA & GLM



Parameters, Estimators, Estimates …

- A parameter is a characteristic of process, population or distribution
  - E.g., mean, 1st quartile, SD, min, max, range, skewness, 97th percentile, etc.
- An estimator is an abstract **rule** for calculating a quantity (or parameter) from sample data.
- An estimate is the value obtained when real data are plugged-in the estimator rule.