Recap: Normal Distribution

The heights of adult men in the United States are normally distributed with a mean of 175 cm and a standard deviation of 8 cm.

Suppose a car is built so that anyone between the height of 150 cm and 185 cm can drive it.

What is the probability that a randomly selected man will not be able to drive this car?

Recap: Normal Distribution

The heights of adult men in the United States are normally distributed with a mean of 175 cm and a standard deviation of 8 cm.

Suppose a car is built so that anyone between the height of 150 cm and 185 cm can drive it.

What is the probability that a randomly selected man will not be able to drive this car?

$X \sim \text{Norm}(\mu=175, \sigma=8)$

$P(X<150) + P(X>185)$

Recap: Normal Distribution

$X \sim \text{Norm}(\mu=175, \sigma=8)$

$P(X<150) + P(X>185)$

$= P((X-175)/8<(150-175)/8) + P((X-175)/8>(185-175)/8)$

$= P(Z < -3.125) + P(Z > 1.25)$

$= (1 - \Phi(3.125)) + (1 - \Phi(1.25))$

$= (1-0.999126) + (1-0.894350)$

$= 0.000874 + 0.10565$

$= 0.106524$

$P(X < \mu - \sigma) = P(X > \mu + \sigma) = (1-0.68)/2=0.16=16\%$

$P(X < \mu - 2\sigma) = P(X > \mu + 2\sigma) = (1- 0.95)/2=0.023=2.3\%$

$P(X < \mu - 3\sigma) = P(X > \mu + 3\sigma) = (1- 0.997)/2=0.0013=0.13\%$

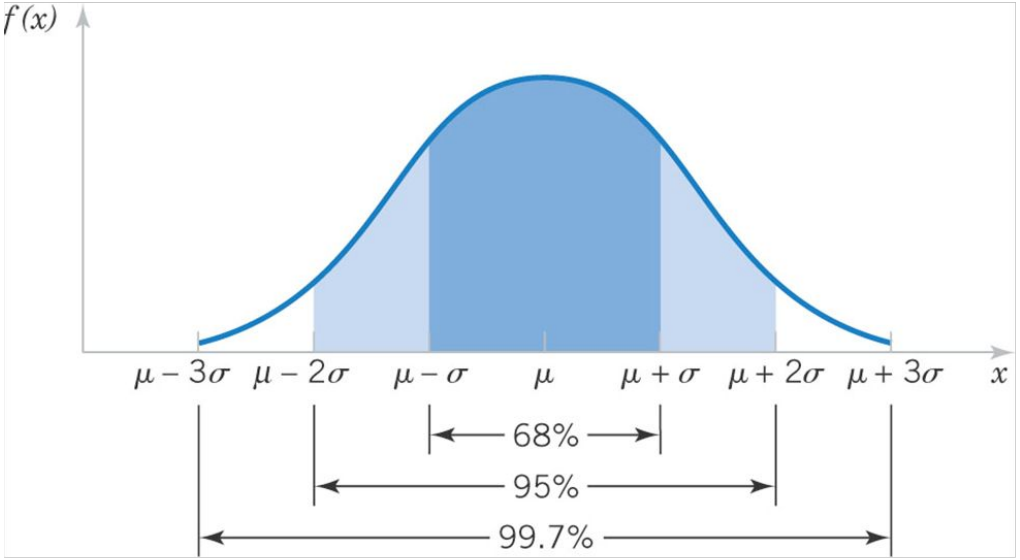
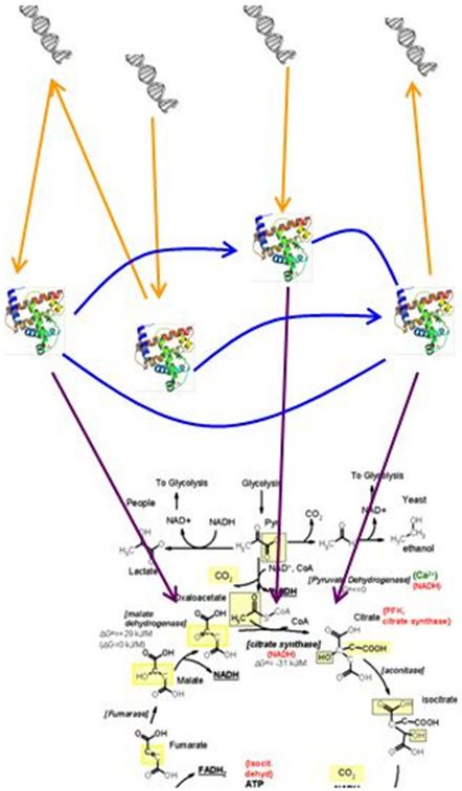


Figure 4-12 Probabilities associated with a normal distribution – well worth remembering to quickly estimate probabilities.

Molecular binding is used at multiple levels

Each level has its own molecular interaction network



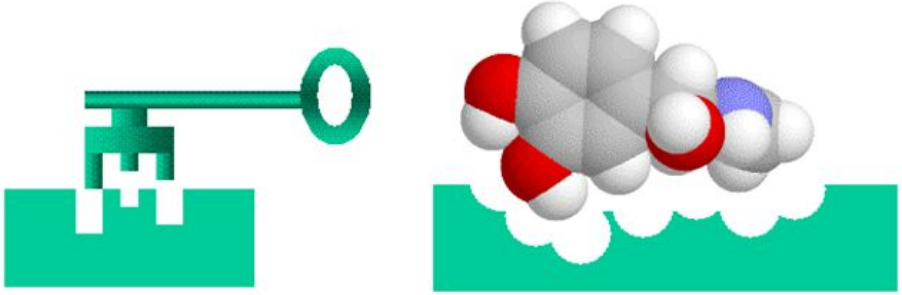
Regulatory network:
RNA-level regulation
by DNA-binding proteins

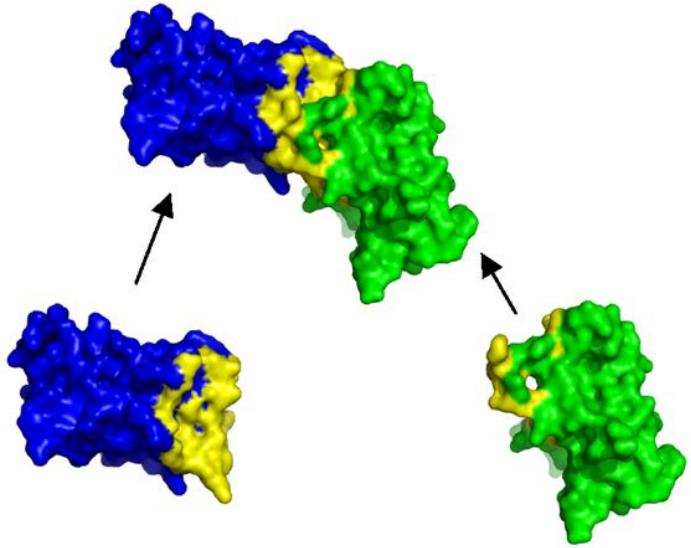
Protein-Protein Interaction Network

Protein-Metabolite Interactions:
Metabolic network

Biological example of a Gaussian: Energy of Protein-Protein Binding Interactions

- Proteins and other biomolecules (metabolites, drugs, DNA) specifically (and non-specifically) bind each other
- For specific bindings: Lock-and-Key theory
- For non-specific bindings: random contacts





PNAS

A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg†, and Eugene I. Shakhnovich‡§

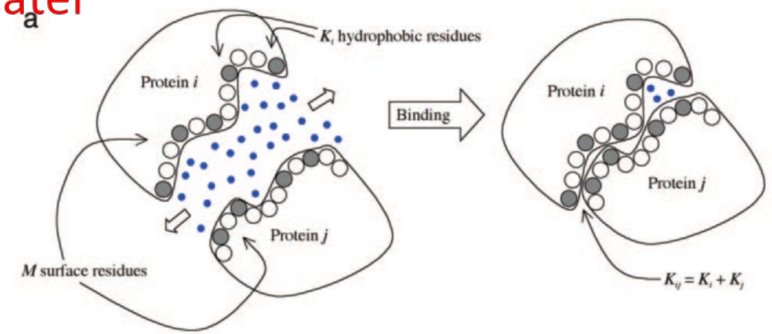
*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; †Harvard College, 12 Oxford Street, Cambridge, MA 02138; and ‡Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Communicated by David Chandler, University of California, Berkeley, CA, November 10, 2005 (received for review September 23, 2005)

It has recently been demonstrated that many biological networks exhibit a "scale-free" topology, for which the probability of observing a node with a certain number of edges (k) follows a power law: i.e., $p(k) \sim k^{-\gamma}$. This observation has been reproduced by

(19–22). Indeed, when the two major *S. cerevisiae* p
protein interaction (PPI) experiments are compared w
another, one finds that only ≈150 of the thousands of
tions identified in each experiment are recovered in th

Most Binding energy is due to hydrophobic amino-acid residues being screened from water



Predicted Gaussian distribution: PDF(E_{ij}=E)– because E_{ij} – sum of hydrophobicities of many independent residues

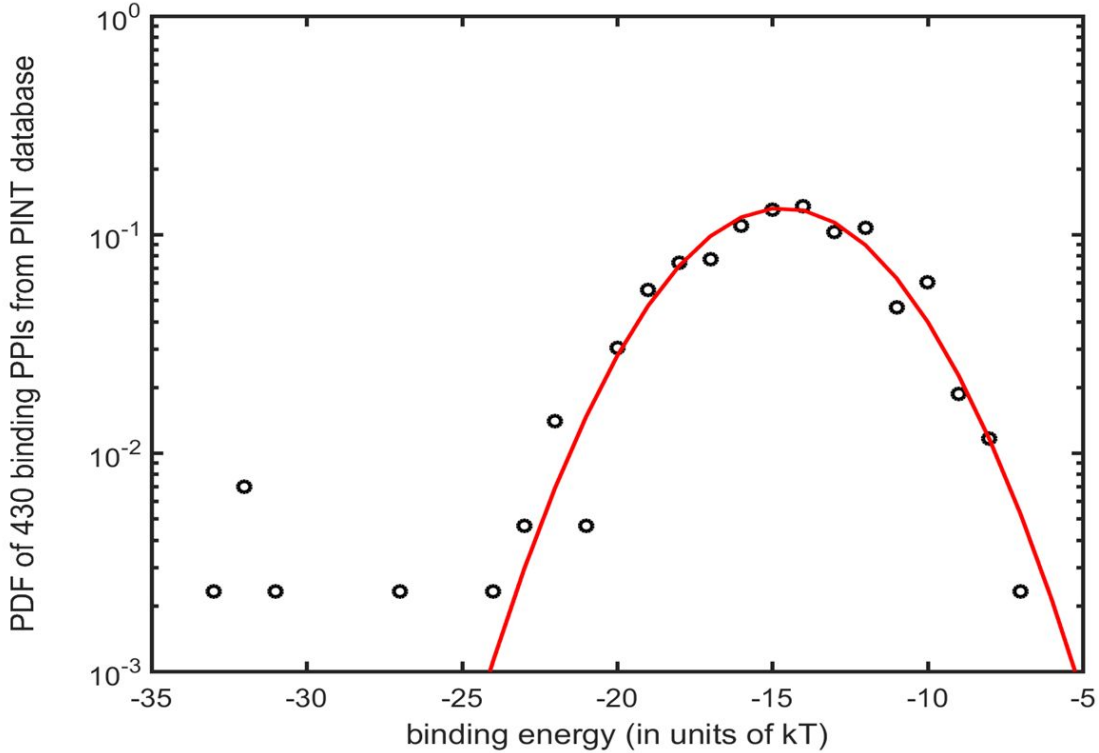
Matlab exercise

1. In Matlab load PINT_binding_energy.mat with binding energy Eij (in units of kT at room temperature) for 430 pairs of interacting proteins from human, yeast, etc.

   Data collected in 2007 from the PINT database
   http://www.bioinfodatabase.com/pint/
   and analyzed in J. Zhang, S. Maslov, E. Shakhnovich, Molecular Systems Biology (2008)

2. Fit Gaussian to the distribution of Eij using dfittool
3. Use “Exclude” button to generate the new exclusion rule to drop all points with X<-23 from the fit
4. Use "New Fit" button to generate the new “Normal” fit with the exclusion rule you just created
5. Find mean (mu) and standard deviation (sigma)
6. Select “probability plot” from “Display type” dropdown menu to evaluate the quality of the plot. Where does the probability plot deviate from a straight line?

How does it compare with the experimental data?



Dissociation constant

- Interaction between two molecules (say, proteins) is usually described in terms of dissociation constant
  $K_{ij}=1M \exp(-E_{ij}/kT)$
- Law of Mass Action: the concentration D_{ij} of a heterodimer formed out of two proteins with free (monomer) concentrations C_i and C_j : $D_{ij}=C_iC_j/K_{ij}$

- What is the distribution of K_{ij}?
  - it is called log-normal since the logarithm of K_{ij} is the binding energy $-E_{ij}/kT$ which is normally distributed

Lognormal Distribution

Lognormal distribution

- Let W denote a normal random variable with mean of θ and variance of ω^2, i.e., E(W) = θ and V(W) = ω^2
- As a change of variable, let X = e^W = exp(W) and W = ln(X)
- Now X is a lognormal random variable.

$$F(x) = P[X \le x] = P[\exp(W) \le x] = P[W \le \ln(x)]$$

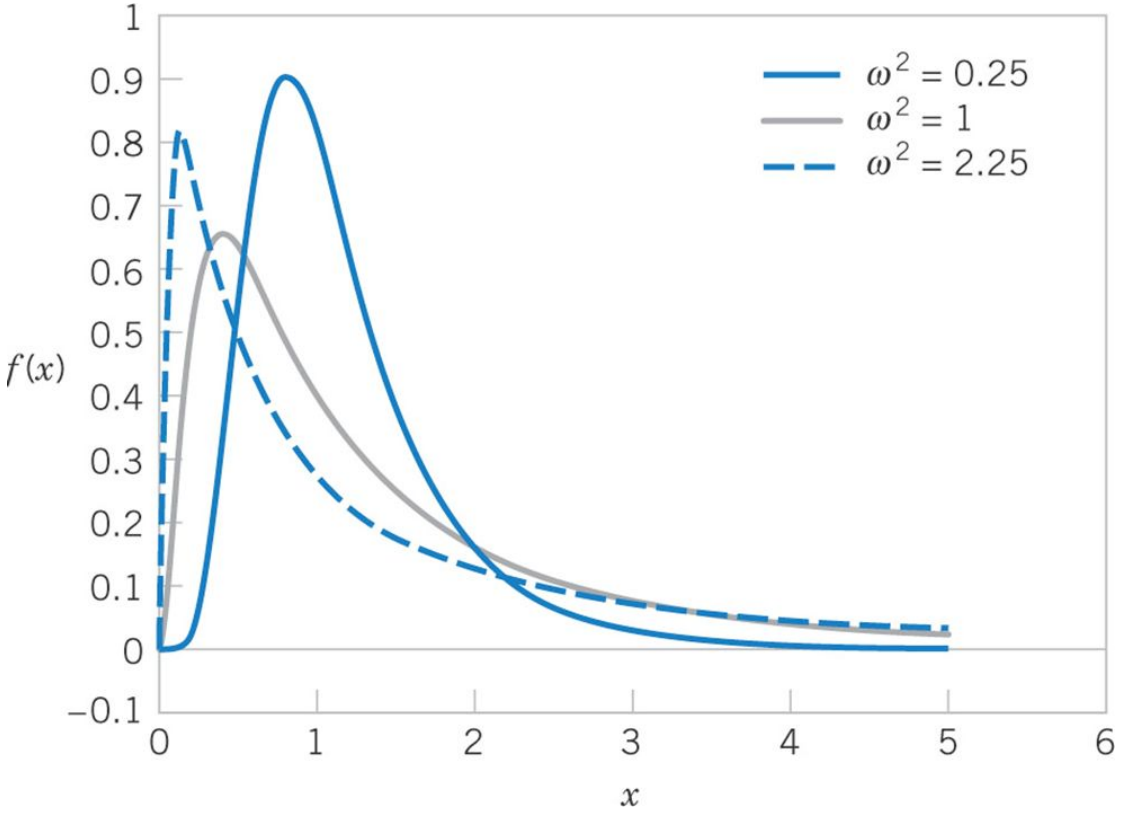
$$= P\left[Z \le \frac{\ln(x)-\theta}{\omega}\right] = \Phi\left[\frac{\ln(x)-\theta}{\omega}\right] = \quad \text{for} \quad x > 0$$

$$= 0 \quad \text{for} \quad x \le 0$$

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{x\omega\sqrt{2\pi}} e^{-\left[\frac{\ln(x)-\theta}{2\omega}\right]^2} \qquad \text{for } 0 < x < \infty$$

$$E(X) = e^{\theta+\omega^2/2} \qquad \text{and} \qquad V(X) = e^{2\theta+\omega^2}\left(e^{\omega^2}-1\right) \qquad (4\text{-}22)$$

Lognormal distribution



What we learned so far…

- Random Events:
  - Working with events as sets: union, intersection, etc.
    - Some events are simple: Head vs Tails, Cancer vs Healthy
    - Some are more complex: 10<Gene expression<100
    - Some are even more complex: Series of dice rolls: 1,3,5,3,2
  - Conditional probability: $P(A|B)=P(A \cap B)/P(B)$
  - Independent events: $P(A|B)=P(A)$ or $P(A \cap B)= P(A)*P(B)$
  - Bayes theorem: relates $P(A|B)$ to $P(B|A)$
- Random variables:
  - Mean, Variance, Standard deviation. How to work with $E(g(X))$
  - Discrete (Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative binomial, Hypergeometric, Power law); PMF: $f(x)=Prob(X=x)$; CDF: $F(x)=Prob(X\leq x)$;
  - Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal); PDF: $f(x)$ such that $Prob(X \text{ inside } A)= \int_A f(x)dx$; CDF: $F(x)=Prob(X\leq x)$

Next step: work with **multiple random variables** measured together in the same series of random experiments

Joint Probability Distributions

Concept of Joint Probabilities

Biological systems are usually described not by a single random variable but by many random variables

Example: The expression state of a human cell:
20,000 random variables X_i for each of its genes

A joint probability distribution describes the behavior of several random variables

We will start with just two random variables X and Y and generalize when necessary

Joint Probability Mass Function Defined

The joint probability mass function of the discrete random variables X and Y, denoted as $f_{XY}(x,y)$, satisfies:

(1) $f_{XY}(x,y) \geq 0$ All probabilities are non−negative

(2) $\sum_x \sum_y f_{XY}(x,y) = 1$ The sum of all probabilities is 1

(3) $f_{XY}(x,y) = P(X = x, Y = y)$

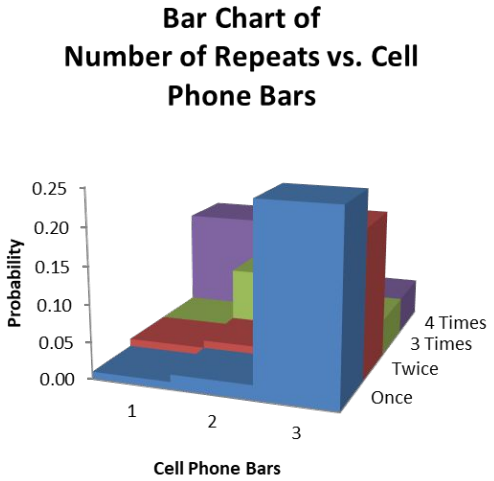
Example: # Repeats vs. Signal Bars

You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

Let Y denote the number of times you have to state your departure city.

Let X denote the number of bars of signal strength on you cell phone.

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05



Marginal Probability Distributions (discrete)

For a discrete joint PDF, there are marginal distributions for each random variable, formed by summing the joint PMF over the other variable.

$$f_X(x) = \sum_y f_{XY}(x, y)$$

$$f_Y(y) = \sum_x f_{XY}(x, y)$$

Called marginal because they are written in the margins

Example: # Repeats vs. Signal Bars

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_{Y(}y) =$
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x) =$	0.20	0.25	0.55	1.00

$$f_X(x) = \sum_y f_{XY}(x, y)$$

$$f_Y(y) = \sum_x f_{XY}(x, y)$$

Mean & Variance of X and Y are calculated using marginal distributions

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_Y(y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

$\mu_X = E(X)$
$\mu_Y = E(Y)$

$\sigma_X^2 = V(X) = E(X^2) - E(X)^2$
$\sigma_Y^2 = V(Y) = E(Y^2) - E(Y)^2$

Mean & Variance of X and Y are calculated using marginal distributions

y = number of times city name is stated	x = number of bars of signal strength					
	1	2	3	$f(y)$ =	$y*f(y)$ =	$y^2*f(y)$ =
1	0.01	0.02	0.25	0.28	0.28	0.28
2	0.02	0.03	0.20	0.25	0.50	1.00
3	0.02	0.10	0.05	0.17	0.51	1.53
4	0.15	0.10	0.05	0.30	1.20	4.80
$f(x)$ =	0.20	0.25	0.55	1.00	2.49	7.61
$x*f(x)$ =	0.20	0.50	1.65	2.35		
$x^2*f(x)$ =	0.20	1.00	4.95	6.15		

$\mu_X = E(X)$

$\mu_Y = E(Y)$

$\sigma_X^2 = V(X) = E(X^2) - E(X)^2$

$\sigma_Y^2 = V(Y) = E(Y^2) - E(Y)^2$

Mean & Variance of X and Y are calculated using marginal distributions

y = number of times city name is stated	x = number of bars of signal strength					
	1	2	3	$f(y)$ =	$y*f(y)$ =	$y^2*f(y)$ =
1	0.01	0.02	0.25	0.28	0.28	0.28
2	0.02	0.03	0.20	0.25	0.50	1.00
3	0.02	0.10	0.05	0.17	0.51	1.53
4	0.15	0.10	0.05	0.30	1.20	4.80
$f(x)$ =	0.20	0.25	0.55	1.00	2.49	7.61
$x*f(x)$ =	0.20	0.50	1.65	2.35		
$x^2*f(x)$ =	0.20	1.00	4.95	6.15		

$\mu_X = E(X) = 2.35$

$\mu_Y = E(Y) = 2.49$

$\sigma_X^2 = V(X) = E(X^2) - E(X)^2$
$= 6.15 - 2.35^2 = 0.6275$

$\sigma_Y^2 = V(Y) = E(Y^2) - E(Y)^2$
$= 7.61 - 2.49^2 = 1.4099$

Conditional Probability Distributions

$P(Y=y|X=x)=P(X=x,Y=y)/P(X=x)=$

$=f(x,y)/f_X(x)$

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_Y(y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

$P(Y=1|X=3) = 0.25/0.55 = 0.455$
$P(Y=2|X=3) = 0.20/0.55 = 0.364$
$P(Y=3|X=3) = 0.05/0.55 = 0.091$
$P(Y=4|X=3) = 0.05/0.55 = 0.091$

Statistically Independent Events

- **Two events**

Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A|B) = P(A)$

(2) $P(B|A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$

- **Multiple events**

The events $E_1, E_2, \ldots, E_n$ are independent if and only if for any subset of these events $E_{i_1}, E_{i_2}, \ldots, E_{i_k}$,

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \times P(E_{i_2}) \times \cdots \times P(E_{i_k})$$

Joint Random Variable Independence

- Joint random variables are independent if any of the following are met

1) $P(Y=y \mid X=x)=P(Y=y)$ for any x or
2) $P(X=x \mid Y=y)=P(X=x)$ for any y or
3) $P(X=x, Y=y)=P(X=x)\cdot P(Y=y)$

- If X and Y and independent, the knowledge of the value of X does not change the probabilities for the values of Y
- If X and Y are dependent, the values of Y and influenced by the values of X

X and Y are Bernoulli variables

	Y=0	Y=1
X=0	2/6	1/6
X=1	2/6	1/6

What is the marginal $P_Y(Y=0)$?

A. 1/6
B. 2/6
C. 3/6
D. 4/6

X and Y are Bernoulli variables

2/6+2/6 = 4/6

	Y=0	Y=1
X=0	2/6	1/6
X=1	2/6	1/6

What is the marginal $P_Y(Y=0)$?

A. 1/6
B. 2/6
C. 3/6
D. 4/6

X and Y are Bernoulli variables

2/6+2/6 = 4/6 2/6

	Y=0	Y=1	
X=0	2/6	1/6	2/6+1/6 = 3/6
X=1	2/6	1/6	2/6+1/6 = 3/6

What is the marginal $P_Y(Y=0)$?

A. 1/6
B. 2/6
C. 3/6
D. 4/6

X and Y are Bernoulli variables

2/6+2/6 = 4/6 2/6

	Y=0	Y=1	
X=0	2/6	1/6	2/6+1/6 = 3/6
X=1	2/6	1/6	2/6+1/6 = 3/6

What is the conditional $P(X=0|Y=1)$?

A. 2/6
B. 1/2
C. 1/6
D. 4/6

X and Y are Bernoulli variables

2/6+2/6 = 4/6 2/6

	Y=0	Y=1	
X=0	2/6	1/6	2/6+1/6 = 3/6
X=1	2/6	1/6	2/6+1/6 = 3/6

What is the conditional P(X=0|Y=1)?

A. 2/6
B. 1/2
C. 1/6
D. 4/6

P(X=0|Y=1) = P(X=0,Y=1)/P(Y=1) = (1/6)/(2/6) = 1/2

X and Y are Bernoulli variables

2/6+2/6 = 4/6 2/6

	Y=0	Y=1	
X=0	2/6	1/6	2/6+1/6 = 3/6
X=1	2/6	1/6	2/6+1/6 = 3/6

Are they independent?

A. yes

B. no

X and Y are Bernoulli variables

2/6+2/6 = 4/6 2/6

	Y=0	Y=1	
X=0	2/6	1/6	2/6+1/6 = 3/6
X=1	2/6	1/6	2/6+1/6 = 3/6

Are they independent?

A. yes

B. no

X and Y are Bernoulli variables

	Y=0	Y=1
X=0	1/2	0
X=1	0	1/2

Are they independent?

A. yes
B. no

X and Y are Bernoulli variables

	Y=0	Y=1
X=0	1/2	0
X=1	0	1/2

Are they independent?

A. yes

B. no