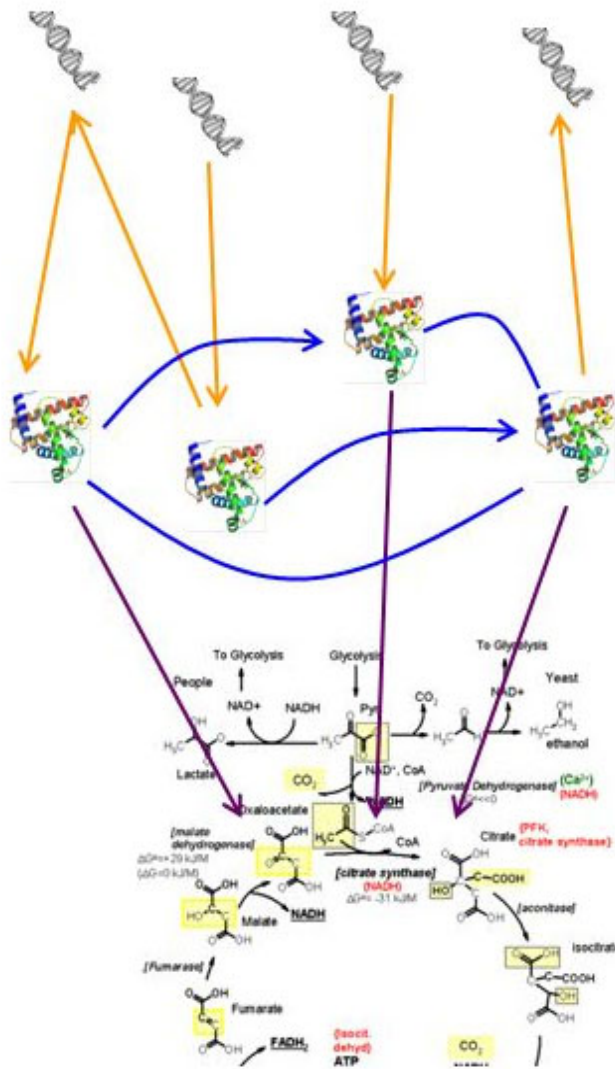


# Fitting a Gaussian distribution: a biological example

# Molecular binding is used at multiple levels

Each level has its own molecular interaction network

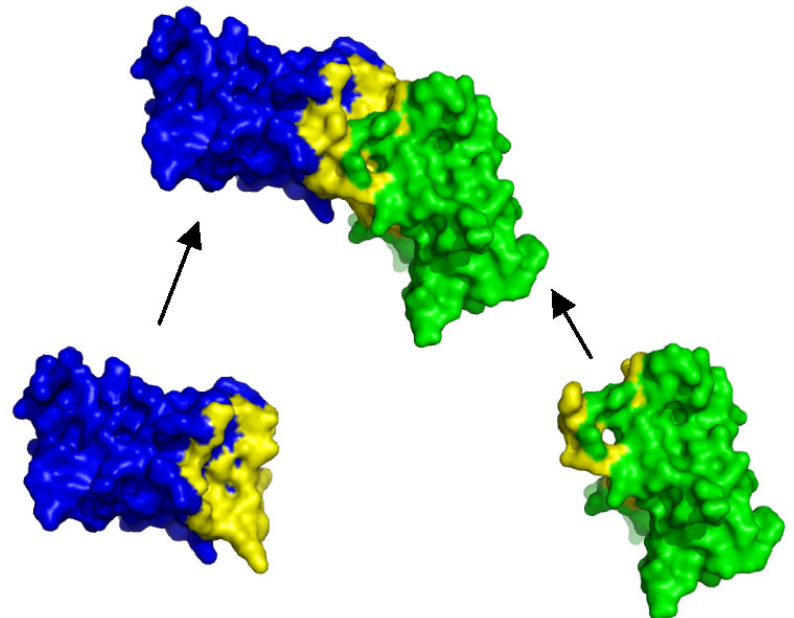
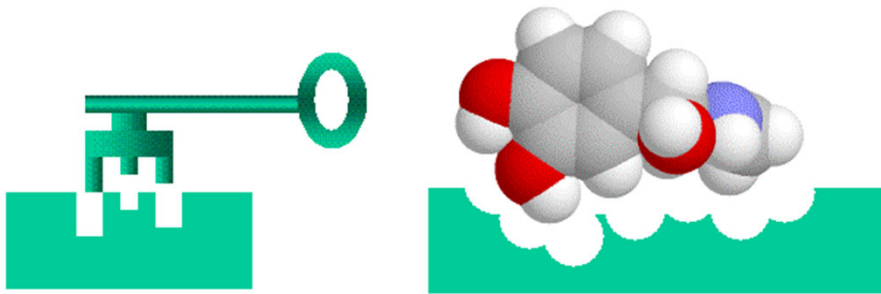


Regulatory network:  
RNA-level regulation  
By DNA-binding  
Proteins  
Protein-Protein (binding) Interaction Network

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



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

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

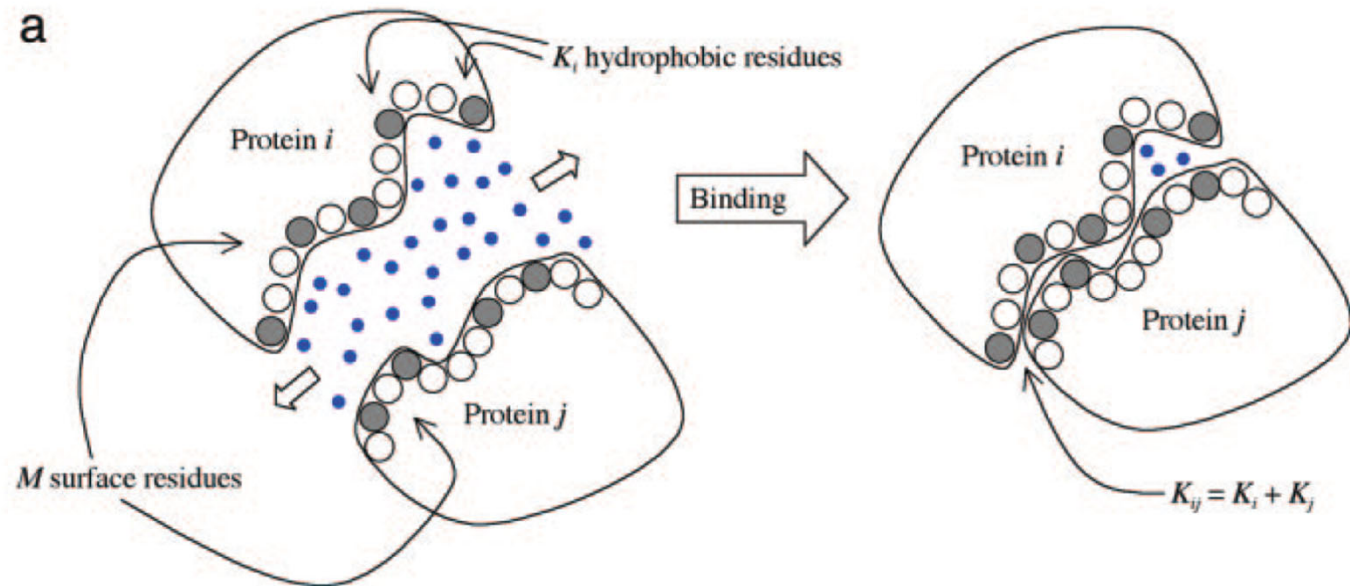
\*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; †Harvard College, 12 Oxford Street, Cambridge, MA 02138; and ‡Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

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* PPI experiments are compared with another, one finds that only  $\approx 150$  of the thousands of interactions identified in each experiment are recovered in the

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

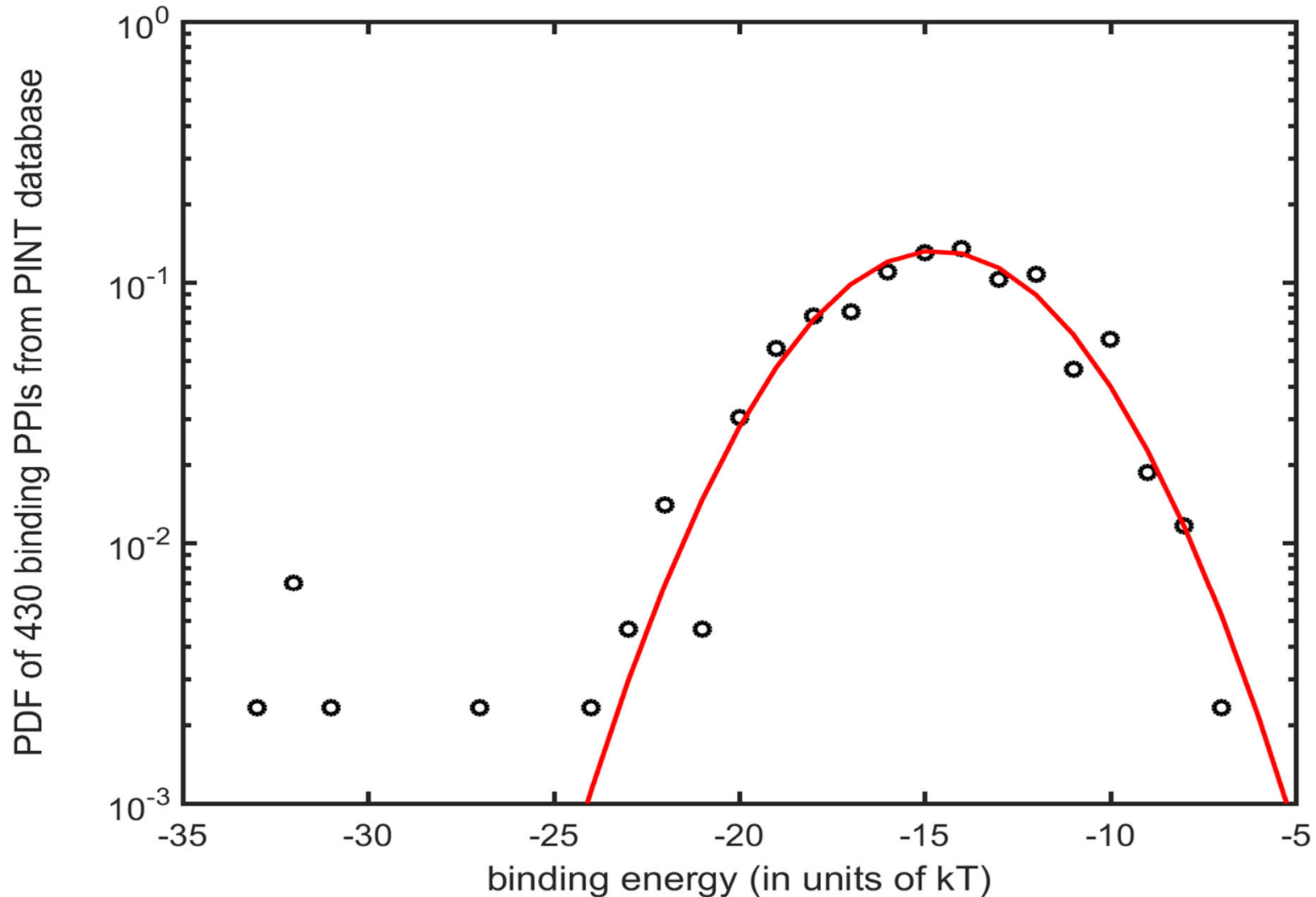


Predicted **Gaussian distribution**:  $\text{PDF}(E_{ij}=E)$ — because  $E_{ij}$  — **sum of hydrophobicities of many independent residues**

# Matlab exercise

- In Matlab load `PINT_binding_energy.mat` with binding energy  $E_{ij}$  (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)
- Fit Gaussian to the distribution of  $E_{ij}$  using `dfittool`
- Use “Exclude” button to generate the new exclusion rule to drop all points with  $X < -23$  from the fit
- Use “New Fit” button to generate the new “Normal” fit with the exclusion rule you just created
- Find mean ( $\mu$ ) and standard deviation ( $\sigma$ )
- 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 ?



J. Zhang, S. Maslov, E. Shakhnovich,  
Nature/EMBO Molecular Systems Biology (2008)

Data on binding interactions  
from PINT database

# 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  $C_j$  :  $D_{ij} = C_i C_j / K_{ij}$
- What is the distribution of  $K_{ij}$ ?
- Answer: it is called log-normal since the **logarithm of  $K_{ij}$**  is the **binding energy  $-E_{ij}/kT$**  which is normally distributed

# Lognormal Distribution

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

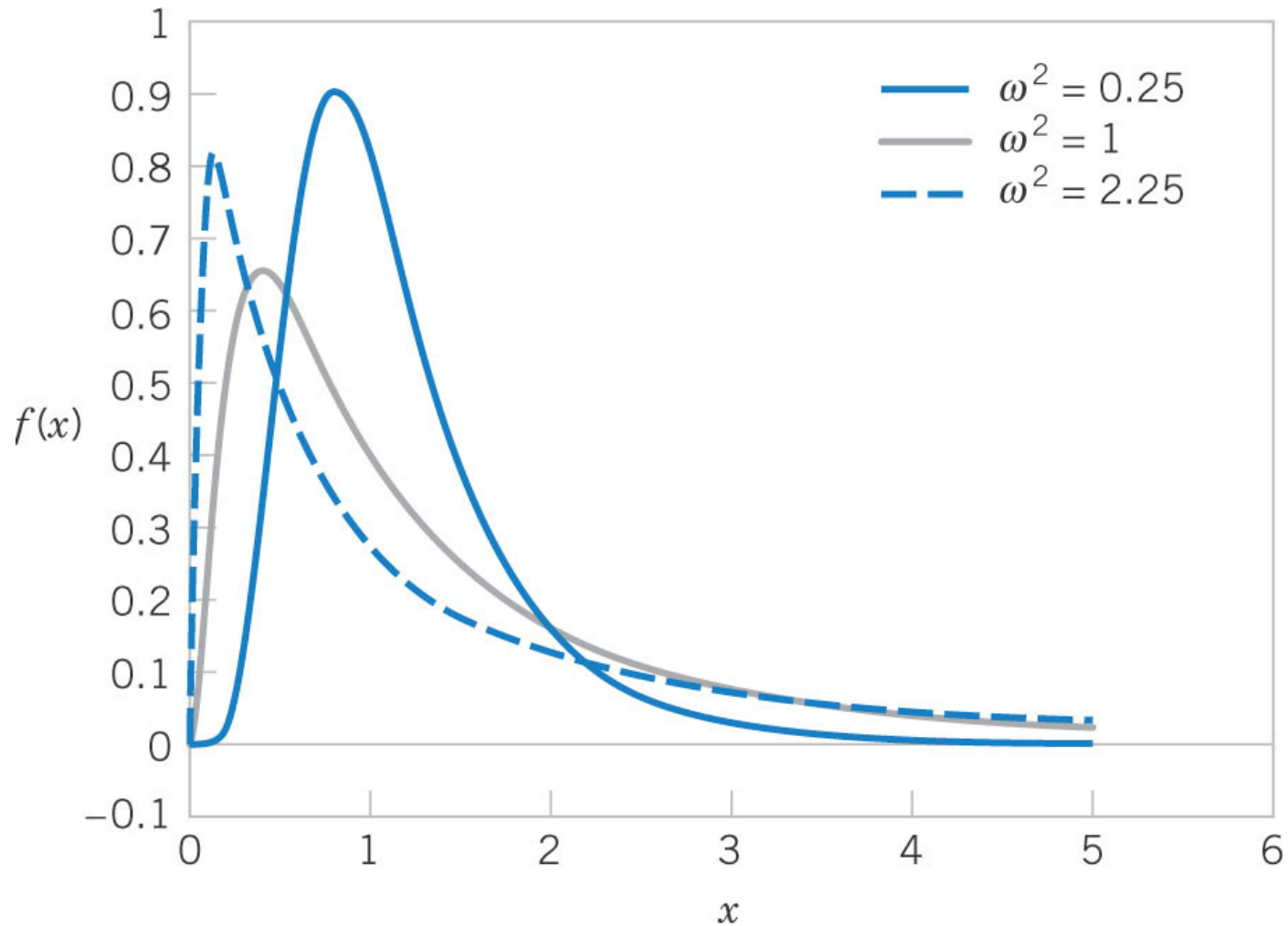
$$\begin{aligned}
 F(x) &= P[X \leq x] = P[\exp(W) \leq x] = P[W \leq \ln(x)] \\
 &= P\left[Z \leq \frac{\ln(x) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(x) - \theta}{\omega}\right] = \quad \text{for } x > 0 \\
 &= 0 \quad \text{for } x \leq 0
 \end{aligned}$$

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

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



# Lognormal Graphs



**Figure 4-27** Lognormal probability density functions with  $\theta = 0$  for selected values of  $\omega^2$ .

Credit: XKCD  
comics

# WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS  
WHY ARE AMERICANS AFRAID OF DRAGONS  
WHY IS HTTPS CROSSED OUT IN RED  
WHY IS THERE A LINE THROUGH HTTPS  
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK  
WHY IS HTTPS IMPORTANT

# QUESTIONS FOUND IN GOOGLE AUTOCOMLETE



WHY ARE THERE WEEKS  
WHY DO I FEEL DIZZY  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN

WHY AREN'T ECONOMISTS RICH

WHY ARE THERE SWARMS OF GNATS  
WHY IS THERE PHLEGM  
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN

WHY DO AMERICANS CALL IT SOCCER

WHY IS PSYCHIC WEAK TO BUG

WHY ARE MY EARS RINGING

WHY DO CHILDREN GET CANCER

WHY ARE THERE SO MANY AVENGERS

WHY IS POSEIDON ANGRY WITH ODYSSEUS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS THERE ICE IN SPACE

WHY IS WOLVERINE NOT IN THE AVENGERS

# WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD

WHY IS SPACE BLACK

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS OUTER SPACE SO COLD

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY ARE THERE PYRAMIDS ON THE MOON

WHY ARE THERE GHOSTS

WHY DO OWLS ATTACK PEOPLE

WHY IS NASA SHUTTING DOWN

WHY ARE THERE GHOSTS

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE GHOSTS

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE THERE GODS

WHY DO SPIDERS COME INSIDE

WHY ARE THERE GHOSTS

WHY ARE THERE TWO SPOCKS

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY IS LIFE SO BORING

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE CIGARETTES LEGAL

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY ARE THERE DUCKS IN MY POOL

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY IS JESUS WHITE

WHY DO SPIDER BITES ITCH

WHY ARE THERE GHOSTS

WHY IS THERE LIQUID IN MY EAR

WHY IS DYING SO SCARY

WHY ARE THERE GHOSTS

WHY DO Q TIPS FEEL GOOD

WHY DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS  
WHY DO I SAY UH  
WHY IS SEA SALT BETTER  
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS  
WHY IS THERE NOT A POKEMON MMO  
WHY IS THERE LAUGHING IN TV SHOWS  
WHY ARE THERE DOORS ON THE FREEWAY  
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING  
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA  
WHY ARE THERE SCARY SOUNDS IN MINECRAFT  
WHY IS THERE KICKING IN MY STOMACH  
WHY ARE THERE TWO SLASHES AFTER HTTP  
WHY ARE THERE CELEBRITIES  
WHY DO SNAKES EXIST  
WHY DO OYSTERS HAVE PEARLS  
WHY ARE DUCKS CALLED DUCKS  
WHY DO THEY CALL IT THE CLAP  
WHY ARE KYLE AND CARTMAN FRIENDS  
WHY IS THERE AN ARROW ON AANG'S HEAD  
WHY ARE TEXT MESSAGES BLUE  
WHY ARE THERE MUSTACHES ON CLOTHES  
WHY ARE THERE MUSTACHES ON CARS  
WHY ARE THERE MUSTACHES EVERYWHERE  
WHY ARE THERE SO MANY BIRDS IN OHIO  
WHY IS THERE SO MUCH RAIN IN OHIO  
WHY IS OHIO WEATHER SO WEIRD

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE THERE BRIDESMAIDS  
WHY DO DYING PEOPLE REACH UP  
WHY AREN'T THERE VARICOSE ARTERIES  
WHY ARE OLD KUNGONS DIFFERENT

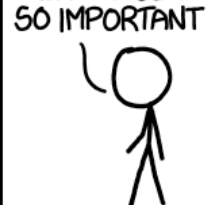
WHY ARE THERE SQUIRRELS



WHY ARE THERE TINY SPIDERS IN MY HOUSE  
WHY DO SPIDERS COME INSIDE  
WHY ARE THERE HUGE SPIDERS IN MY HOUSE  
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE  
WHY ARE THERE SPIDERS IN MY ROOM  
WHY ARE THERE SO MANY SPIDERS IN MY ROOM  
WHY DO SPIDER BITES ITCH  
WHY IS DYING SO SCARY  
WHY IS THERE NO GPS IN LAPTOPS  
WHY DO KNEES CLICK  
WHY AREN'T THERE E GRADES  
WHY IS ISOLATION BAD  
WHY DO BOYS LIKE ME  
WHY DON'T BOYS LIKE ME  
WHY IS THERE ALWAYS A JAVA UPDATE  
WHY ARE THERE RED DOTS ON MY THIGHS  
WHY IS LYING GOOD

WHY ARE THERE FEMALE MR NIMES

WHY IS SEX SO IMPORTANT



WHY ARE THERE FEMALE MR NIMES



WHY IS MT VESUVIUS THERE  
WHY DO THEY SAY T MINUS  
WHY ARE THERE OBELISKS  
WHY ARE WRESTLERS ALWAYS WET  
WHY ARE OCEANS BECOMING MORE ACIDIC  
WHY IS ARWEN DYING  
WHY AREN'T MY QUAIL LAYING EGGS  
WHY AREN'T MY QUAIL EGGS HATCHING  
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY IS LIFE SO BORING

WHY ARE CIGARETTES LEGAL  
WHY ARE THERE DUCKS IN MY POOL  
WHY IS JESUS WHITE  
WHY IS THERE LIQUID IN MY EAR  
WHY DO Q TIPS FEEL GOOD  
WHY DO GOOD PEOPLE DIE



WHY ARE DOGS AFRAID OF FIREWORKS  
WHY IS THERE NO KING IN ENGLAND

WHY IS PROGRAMMING SO HARD  
WHY IS THERE A 0 OHM RESISTOR  
WHY DO AMERICANS HATE SOCCER  
WHY DO RHYMES SOUND GOOD  
WHY DO TREES DIE  
WHY IS THERE NO SOUND ON CNN  
WHY AREN'T POKEMON REAL  
WHY AREN'T BULLETS SHARP  
WHY DO DREAMS SEEM SO REAL

# Multiple random variables, Correlations

# 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:  $10 < \text{Gene expression} < 100$
    - 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, Power law);  
**PMF**:  $f(x) = \text{Prob}(X=x)$ ; **CDF**:  $F(x) = \text{Prob}(X \leq x)$ ;
  - **Continuous** (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal);  
**PDF**:  $f(x)$  such that  $\text{Prob}(X \text{ inside } A) = \int_A f(x) dx$ ; **CDF**:  $F(x) = \text{Prob}(X \leq x)$
- **Next step**: work with **multiple random variables** measured together in the same series of random experiments



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

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

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

Montgomery Runger 5th edition Equation (5-1)

# Example 5-1: # Repeats vs. Signal Bars

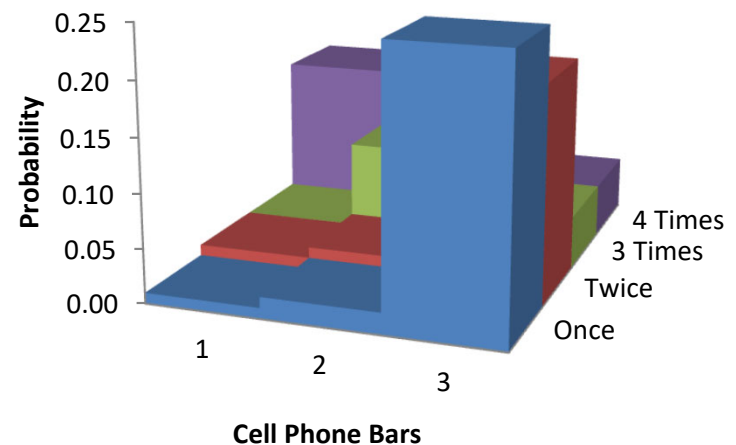
You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

- 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

**Figure 5-1** Joint probability distribution of  $X$  and  $Y$ . The table cells are the probabilities. Observe that more bars relate to less repeating.

**Bar Chart of Number of Repeats vs. Cell Phone Bars**



# 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.

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

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

y = number of times city name is stated	x = number of bars of signal strength			$f_Y(y) =$
	1	2	3	
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

Called **marginal** because they are **written in the margins**

**Figure 5-6** From the prior example, the joint PMF is shown in green while the two marginal PMFs are shown in purple.



# 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		

$$\mu_X = E(X) = 2.35; \quad \sigma_X^2 = V(X) = 6.15 - 2.35^2 = 6.15 - 5.52 = 0.6275$$

$$\mu_Y = E(Y) = 2.49; \quad \sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 6.20 = 1.4099$$

# Conditional Probability Distributions

Recall that  $P(B|A) = \frac{P(A \cap B)}{P(A)}$

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

From Example 5-1

$$P(Y=1 | X=3) = 0.25/0.55 = 0.455$$

$$P(Y=2 | X=3) = 0.20/0.55 = 0.364$$

$$P(Y=3 | X=3) = 0.05/0.55 = 0.091$$

$$P(Y=4 | X=3) = 0.05/0.55 = 0.091$$

$$\text{Sum} = 1.00$$

y = number of times city name is stated	x = number of bars of signal strength			$f_Y(y) =$
	1	2	3	
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

Note that there are 12 probabilities conditional on  $X$ , and 12 more probabilities conditional upon  $Y$ .

Reminder

# Statistically independent events

Always true:  $P(A \cap B) = P(A | B) \cdot P(B) = P(B | A) \cdot P(A)$

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

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

# Joint Random Variable Independence

- Random variable independence means that knowledge of the value of  $X$  does not change any of the probabilities associated with the values of  $Y$ .
- Opposite: Dependence implies that the values of  $X$  are influenced by the values of  $Y$

# Independence for Discrete Random Variables

- Remember independence of events (slide 13 lecture 4) : Events are independent if **any one** of the three conditions are met:
  - 1)  $P(A|B) = P(A \cap B) / P(B) = P(A)$  or
  - 2)  $P(B|A) = P(A \cap B) / P(A) = P(B)$  or
  - 3)  $P(A \cap B) = P(A) \cdot P(B)$
- Random variables independent if **all events**  $A$  that  $Y=y$  and  $B$  that  $X=x$  are independent if any one of these conditions is met:
  - 1)  $P(Y=y | X=x) = P(Y=y)$  for any  $x$  or
  - 2)  $P(X=x | Y=y) = P(X=x)$  for any  $y$  or
  - 3)  $P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$**for every pair  $x$  and  $y$**

# X and Y are Bernoulli variables

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

Are they independent?

A. yes

B. no

C. I don't know

Get your i-clickers

# 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

C. I don't know

Get your i-clickers