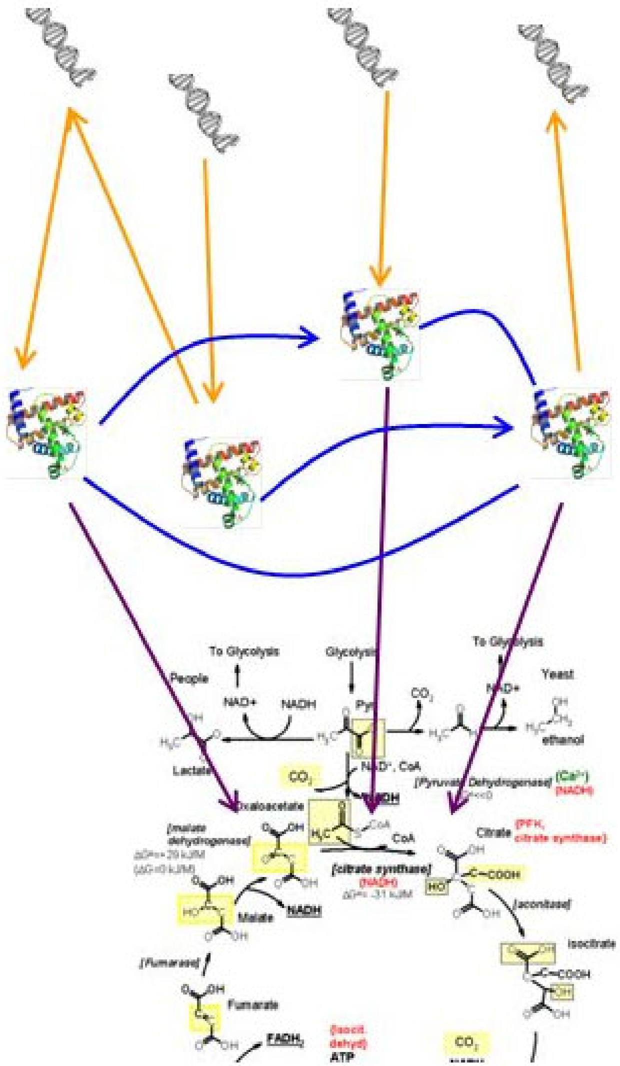
Fitting a Gaussian distribution: a biological example

Molecular binding is used at multiple levels

Each level has its own molecular interaction network



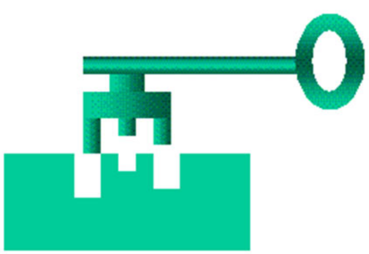
Regulatory network:
RNA-level regulation
By DNA-binding Proteins

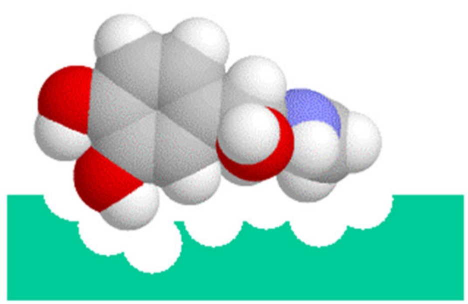
Protein-Protein (binding) Interaction Network

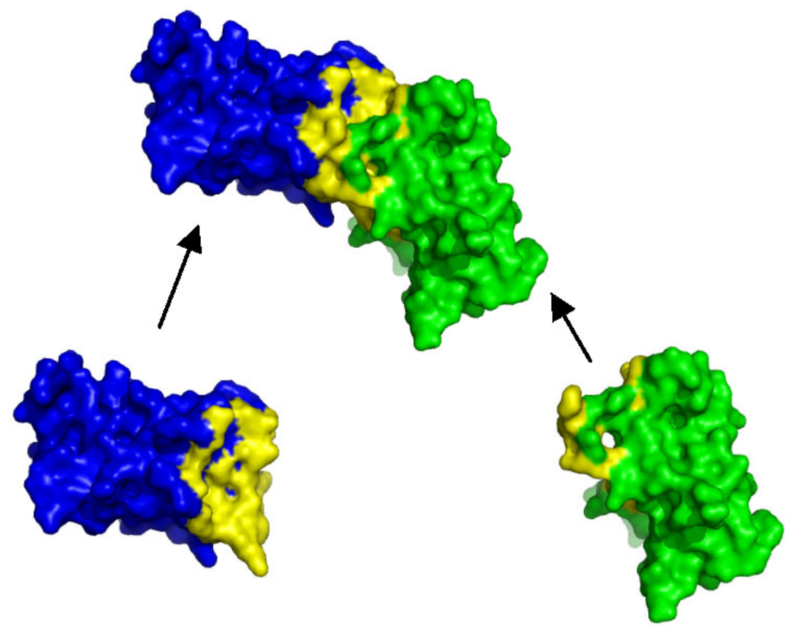
Protein-Metabolite Interactions:
Metabolic network

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







PNAS

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

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

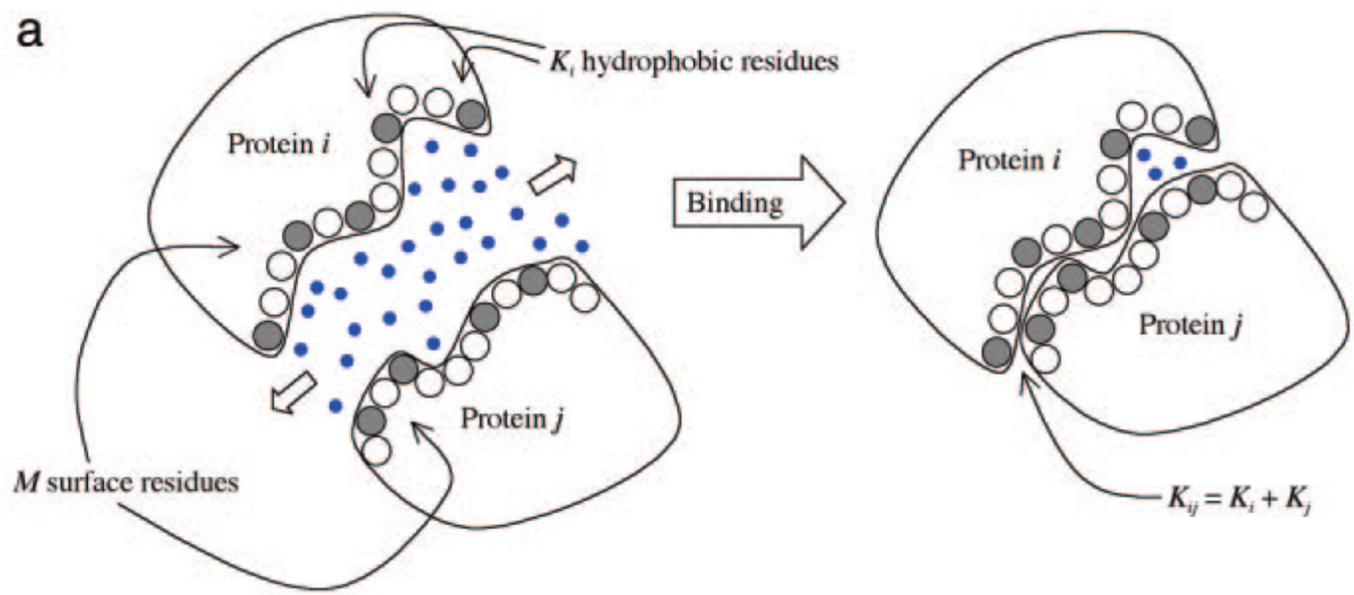
*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; †Harvard College, 12 Oxford Street, Cambridge, MA 02138; and ‡Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

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

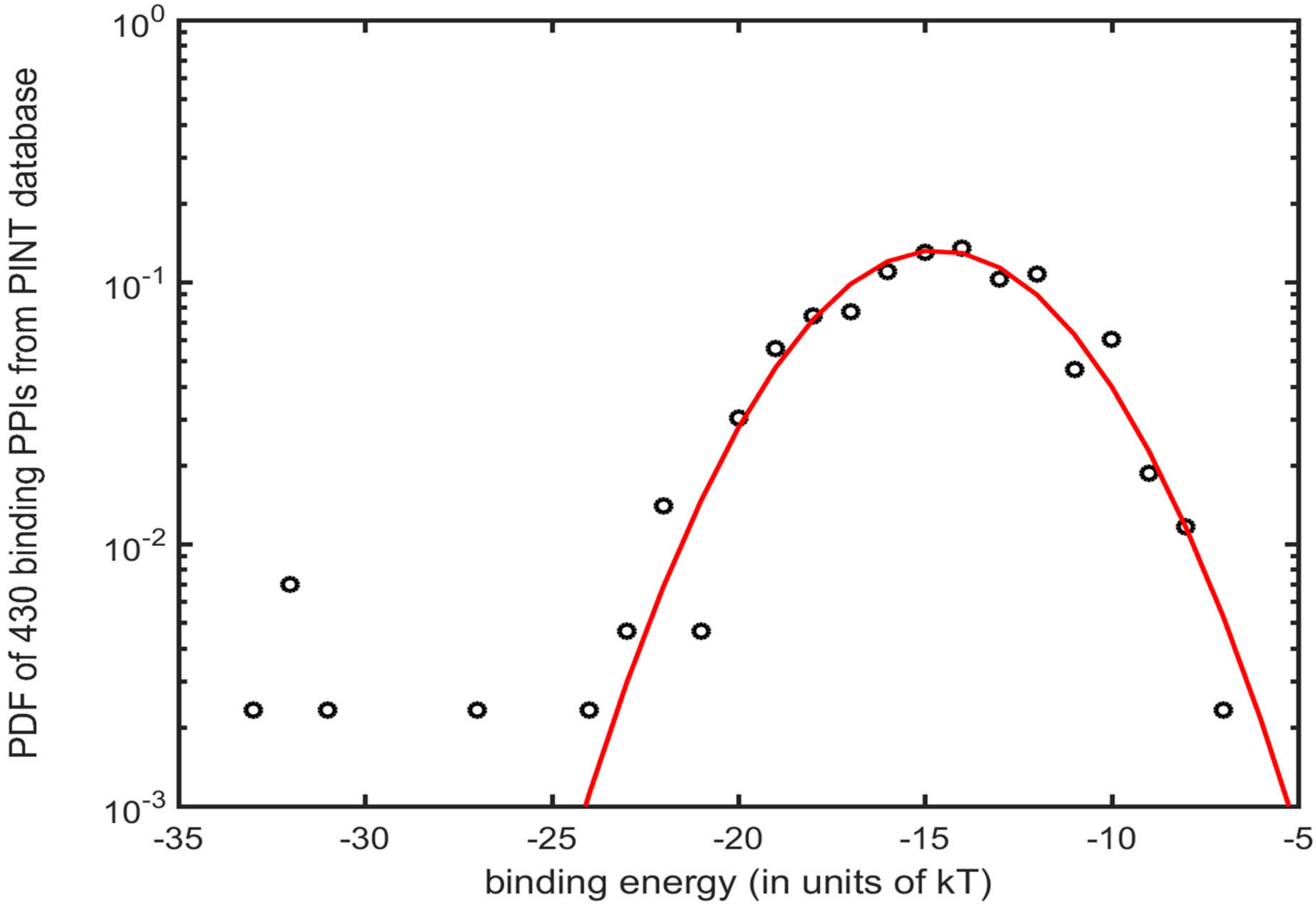


Predicted Gaussian distribution: PDF(E_{ij}=E)– because E_{ij} – sum of hydrophobicities of many independent residues

Matlab exercise

- 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 : $D_{ij}=C_iC_j/K_{ij}$
- 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 θ and variance of ω^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.

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

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

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

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

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

Lognormal Graphs

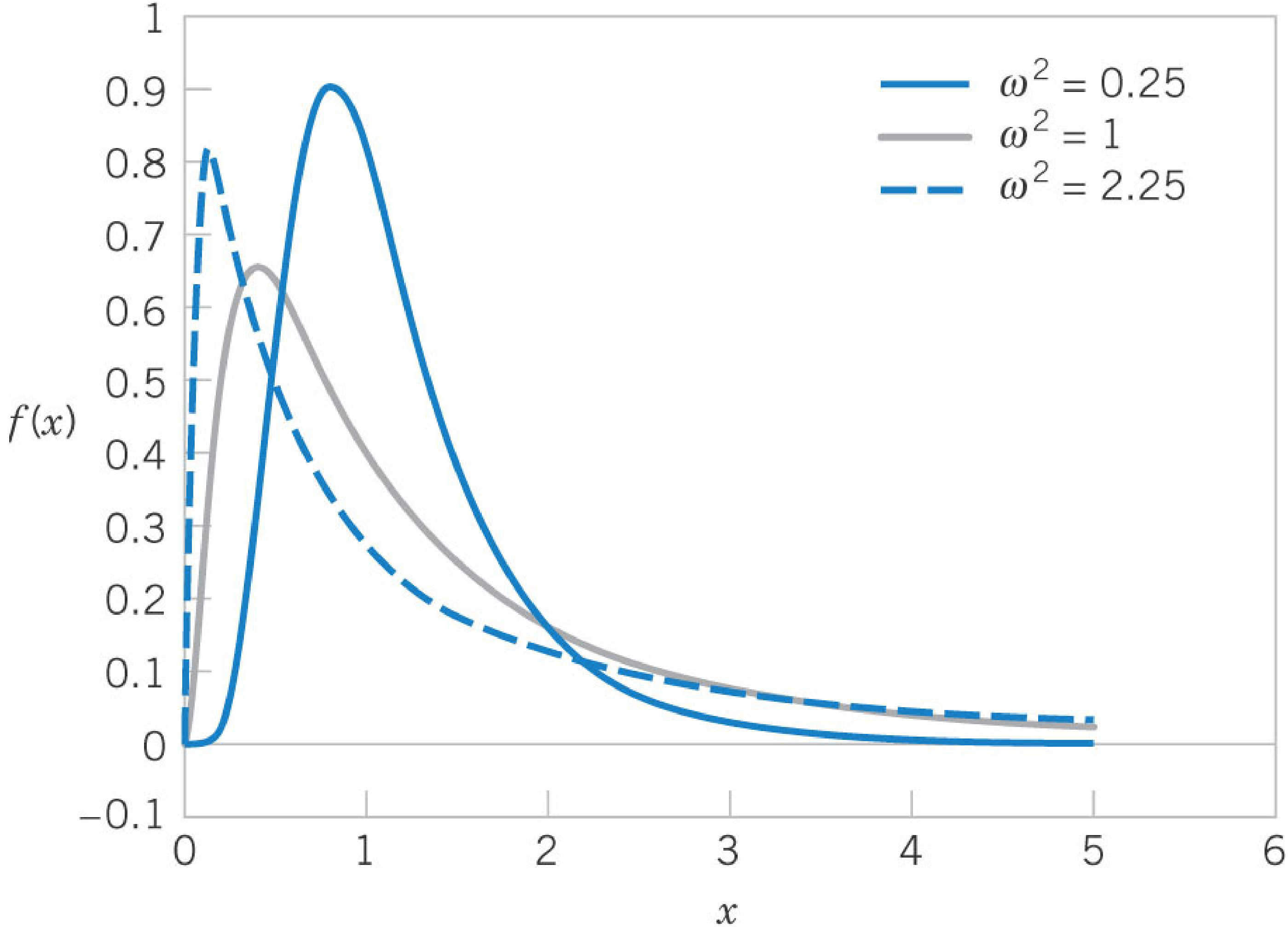
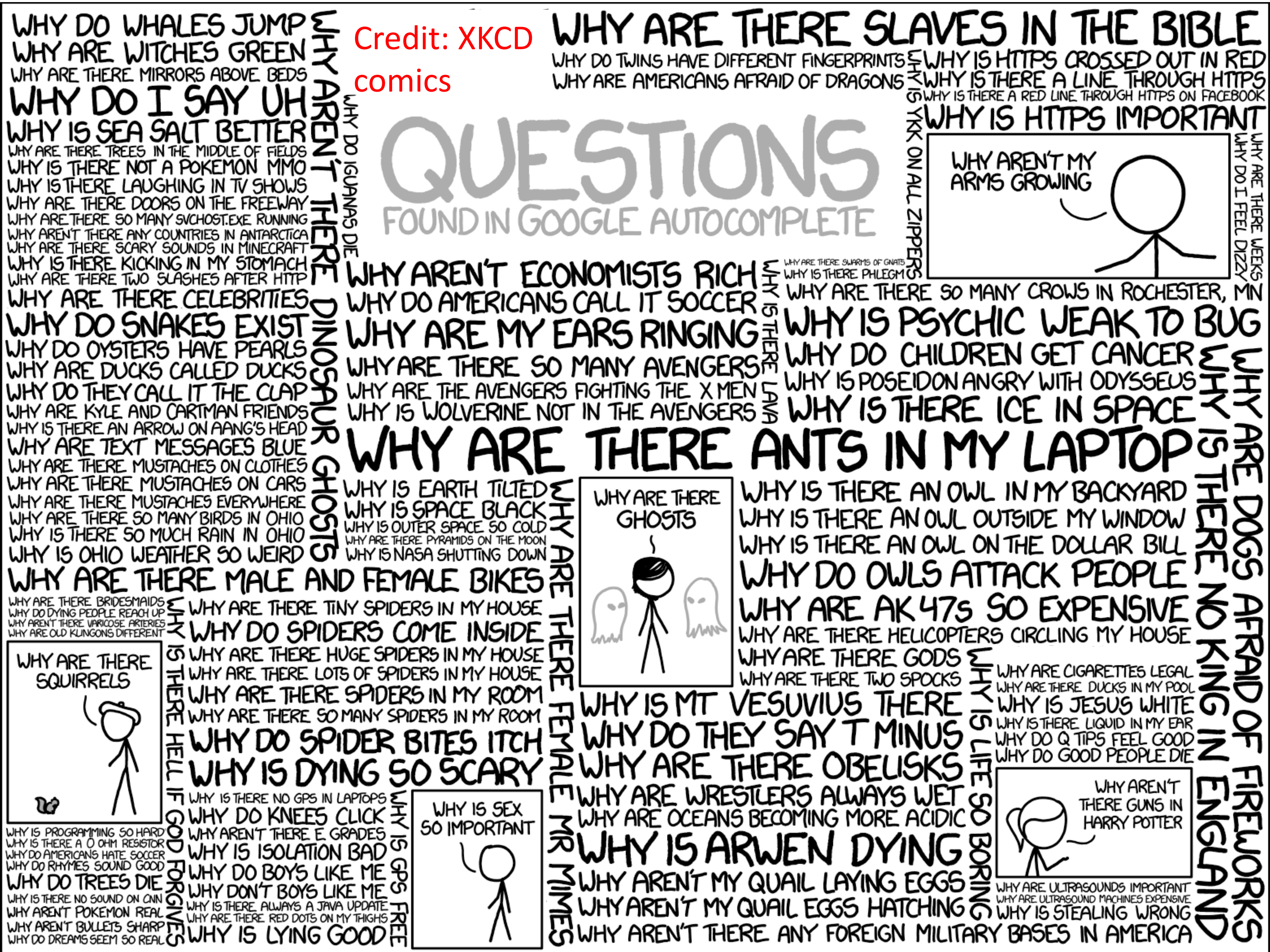


Figure 4-27 Lognormal probability density functions with $\theta = 0$ for selected values of ω^2.

Credit: XKCD comics



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<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)=Prob(X=x); CDF: F(x)=Prob(X≤x);
  - Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal);
    PDF: f(x) such that Prob(X inside A)= $\int_A f(x)dx$; CDF: F(x)=Prob(X≤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 $f_{XY}(x, y)$, satifies:

(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

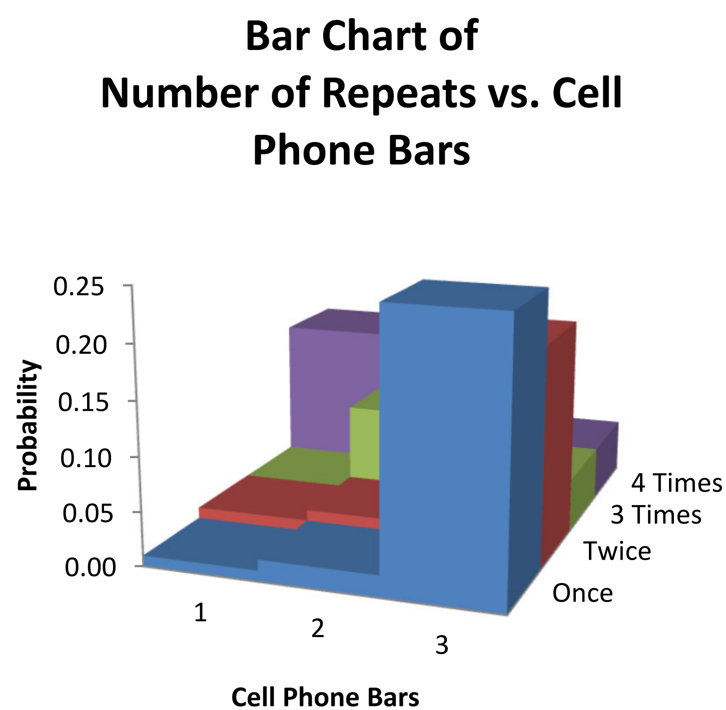


Figure 5-1 Joint probability distribution of X and Y. The table cells are the probabilities. Observe that more bars relate to less repeating.

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

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

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		

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

$\mu_Y = E(Y) = 2.49$; $\sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 16.20 = 1.4099$

Conditional Probability Distributions

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

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

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

From Example 5-1

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

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

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

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

Sum = *1.00*

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

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 $E_1, E_2, \ldots, E_n$ 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

- 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