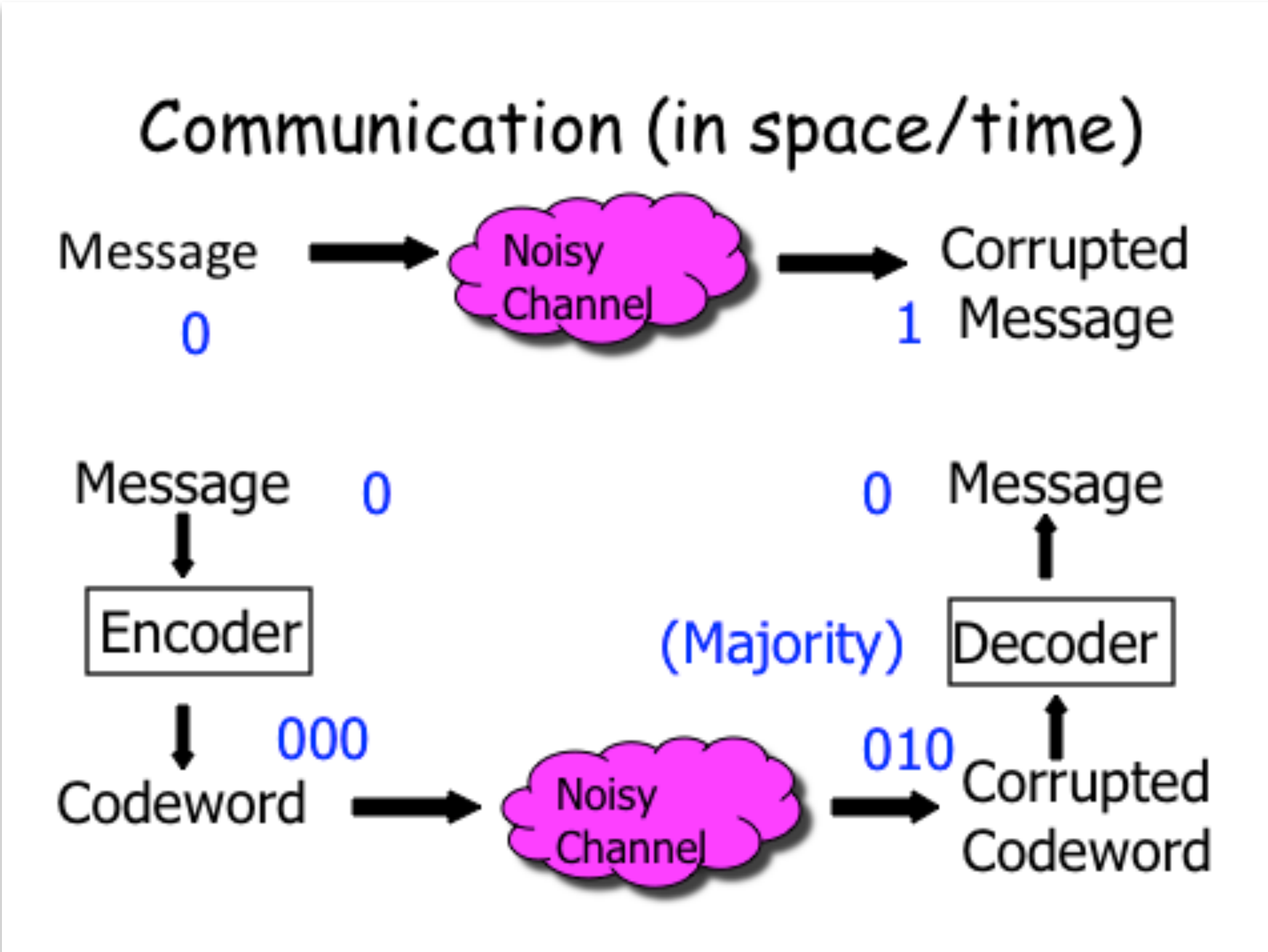
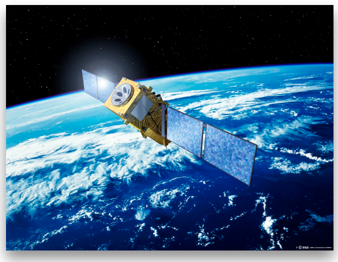
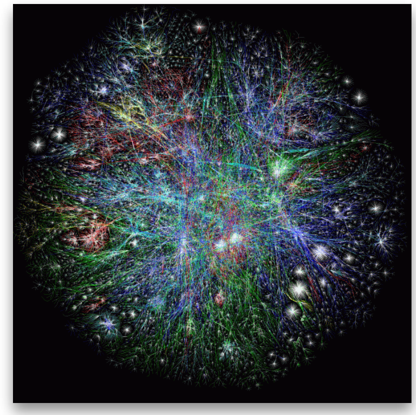
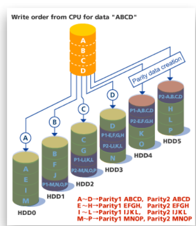
application: error correcting codes

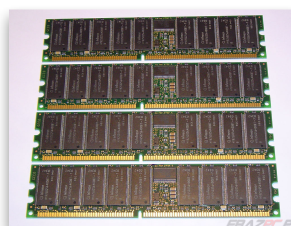


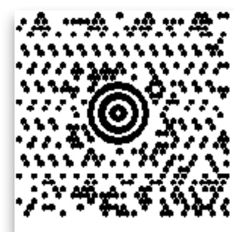
Codes are all around us











Goal: send a 4-bit message over a noisy communication channel.

Say, 1 bit in 10 is flipped in transit, independently.

What is the probability that the message arrives correctly?

Let X = # of errors; X ~ Bin(4, 0.1)

P(correct message received) = P(X=0)

$$P(X=0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561$$

Can we do better? Yes: error correction via redundancy.

E.g., send every bit in triplicate; use majority vote.

Let Y = # of errors in one trio; Y ~ Bin(3, 0.1); P(a trio is OK) =

$$P(Y \leq 1) = \binom{3}{0}(0.1)^0(0.9)^3 + \binom{3}{1}(0.1)^1(0.9)^2 = 0.972$$

If X' = # errors in triplicate msg, X' ~ Bin(4, 0.028), and

$$P(X'=0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168$$

The Hamming(7,4) code:

Have a 4-bit string to send over the network (or to disk)

Add 3 "parity" bits, and send 7 bits total

If bits are $b_1 b_2 b_3 b_4$ then the three parity bits are

$parity(b_1 b_2 b_3)$, $parity(b_1 b_3 b_4)$, $parity(b_2 b_3 b_4)$

Each bit is independently corrupted (flipped) in transit with probability 0.1

Z = number of bits corrupted ~ Bin(7, 0.1)

The Hamming code allow us to *correct* all 1 bit errors.

(E.g., if b_1 flipped, 1st 2 parity bits, but not 3rd, will look wrong; the only single bit error causing this symptom is b_1. Similarly for any other single bit being flipped. Some, but not all, multi-bit errors can be detected, but not corrected.)

P(correctable message received) = P(Z ≤ 1)

Using Hamming error-correcting codes: Z ~ Bin(7, 0.1)

$$P(Z \le 1) = \binom{7}{0}(0.1)^0(0.9)^7 + \binom{7}{1}(0.1)^1(0.9)^6 \approx 0.8503$$

Recall, uncorrected success rate is

$$P(X = 0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561$$

And triplicate code error rate is:

$$P(X' = 0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168$$

Hamming code is nearly as reliable as the triplicate code, with 5/12 ≈ 42% fewer bits. (& better with longer codes.)

Sending a bit string over the network

- $n = 4$ bits sent, each corrupted with probability 0.1
- X = # of corrupted bits, $X \sim \mathrm{Bin}(4, 0.1)$
- In real networks, large bit strings (length $n \approx 10^4$)
- Corruption probability is very small: $p \approx 10^{-6}$

Extreme n and p values arise in many cases

- # bit errors in file written to disk
- # of typos in a book
- # of elements in particular bucket of large hash table
- # of server crashes per day in giant data center
- # facebook login requests sent to a particular server


Siméon Poisson, 1781-1840

Suppose "events" happen, independently, at an *average* rate of λ per unit time. Let X be the *actual* number of events happening in a given time unit. Then X is a *Poisson* r.v. *with parameter* λ (denoted X ~ Poi(λ)) and has distribution (PMF):

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$$

Examples:

# of alpha particles emitted by a lump of radium in 1 sec.

# of traffic accidents in Seattle in one year

# of babies born in a day at UW Med center

# of visitors to my web page today

See B&T Section 6.2 for more on theoretical basis for Poisson.

X is a Poisson r.v. with parameter λ if it has PMF:

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}$$

Is it a valid distribution? Recall Taylor series:

$$e^{\lambda} = \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \cdots = \sum_{0 \le i} \frac{\lambda^i}{i!}$$

So

$$\sum_{0 \le i} P(X = i) = \sum_{0 \le i} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{0 \le i} \frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$$

$$
\begin{aligned}
E[X] &= \sum_{0 \le i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \\
&= \sum_{1 \le i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{1 \le i} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{0 \le j} \frac{\lambda^j}{j!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}
$$

i = 0 term is zero

j = i-1

As expected, given definition in terms of "average rate λ"

(Var[X] = λ, too; proof similar, see B&T example 6.20)

binomial random variable is Poisson in the limit

Poisson approximates binomial when n is large, p is small, and $\lambda = np$ is "moderate"

Formally, Binomial is Poisson in the limit as $n \to \infty$ (equivalently, $p \to 0$) while holding $np = \lambda$

X ~ Binomial(n,p)

$$
\begin{aligned}
P(X=i) &= \binom{n}{i} p^i (1-p)^{n-i} \\
&= \frac{n!}{i!(n-i)!}\left(\frac{\lambda}{n}\right)^i \left(1-\frac{\lambda}{n}\right)^{n-i}, \text{ where } \lambda = pn \\
&= \frac{n(n-1)\cdots(n-i+1)}{n^i}\ \frac{\lambda^i}{i!}\ \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \\
&= \underbrace{\frac{n(n-1)\cdots(n-i+1)}{(n-\lambda)^i}}_{\approx\, 1}\ \frac{\lambda^i}{i!}\ \underbrace{(1-\lambda/n)^n}_{\approx\, e^{-\lambda}} \\
&\approx 1 \cdot \frac{\lambda^i}{i!} \cdot e^{-\lambda}
\end{aligned}
$$

I.e., Binomial ≈ Poisson for large n, small p, moderate i, λ.

Recall example of sending bit string over a network

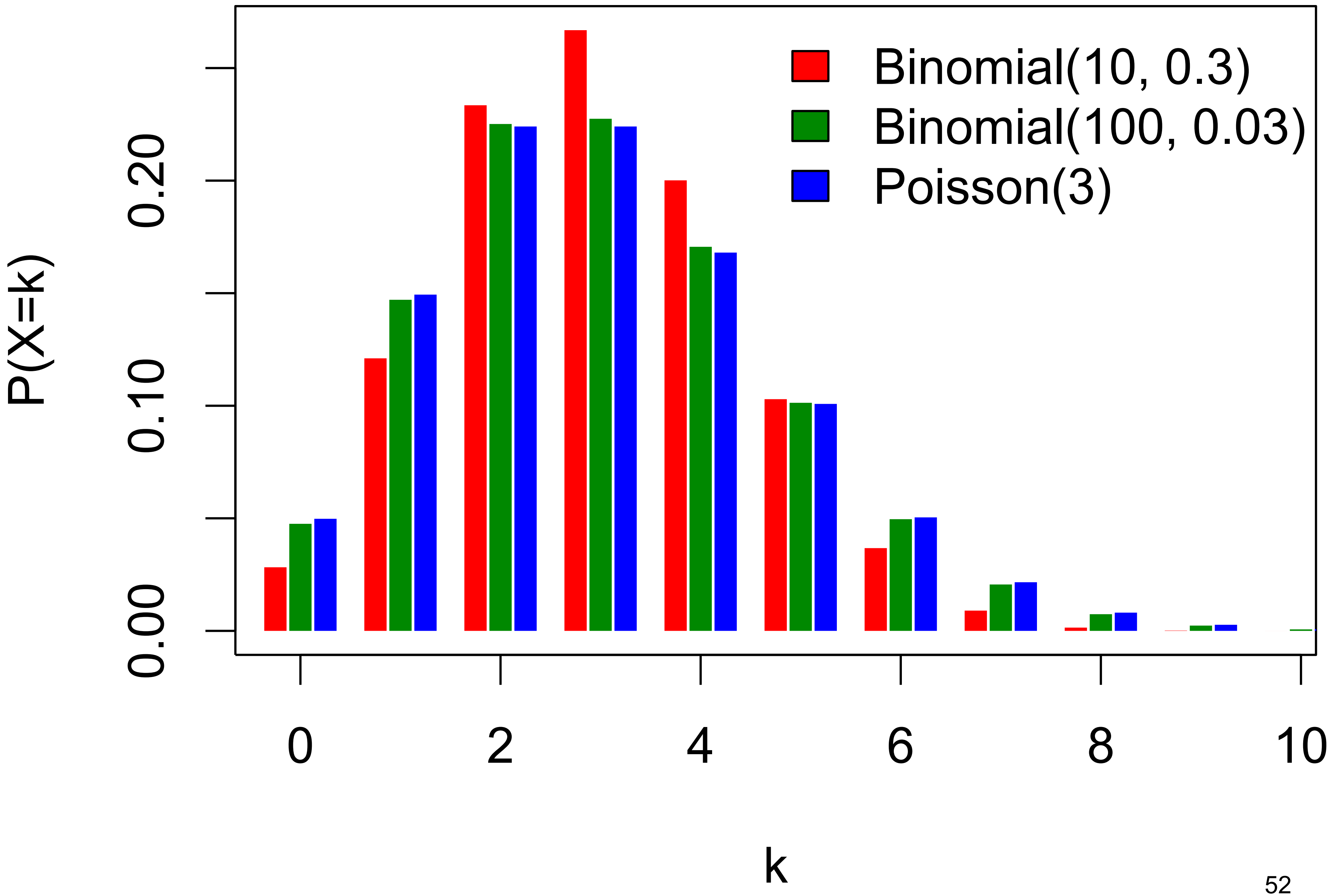
- Send bit string of length $n = 10^4$
- Probability of (independent) bit corruption is $p = 10^{-6}$
- $X \sim Poi(\lambda = 10^4 \cdot 10^{-6} = 0.01)$
- What is probability that message arrives uncorrupted?

$$P(X = 0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-0.01} \frac{0.01^0}{0!} \approx 0.9900498\underline{34}$$

Using $Y \sim Bin(10^4, 10^{-6})$:

- $P(Y=0) \approx 0.9900498\underline{29}$

binomial vs Poisson



Recall: if $Y \sim \text{Bin}(n,p)$, then:

$E[Y] = pn$

$Var[Y] = np(1-p)$

And if $X \sim \text{Poi}(\lambda)$ where $\lambda = np$ ($n \to \infty$, $p \to 0$) then

$E[X] = \lambda = np = E[Y]$

$Var[X] = \lambda \approx \lambda(1-\lambda/n) = np(1-p) = Var[Y]$

Expectation and variance of Poisson are the same (λ)

Expectation is the same as corresponding binomial

Variance almost the same as corresponding binomial

Note: when two different distributions share the same mean & variance, it suggests (but doesn't prove) that one may be a good approximation for the other.

In a series $X_1, X_2, \ldots$ of Bernoulli trials with success probability p, let Y be the index of the first success, i.e.,

$$X_1 = X_2 = \ldots = X_{Y-1} = 0 \;\&\; X_Y = 1$$

Then Y is a *geometric* random variable with parameter p.

Examples:

Number of coin flips until first head

Number of blind guesses on LSAT until I get one right

Number of darts thrown until you hit a bullseye

Number of random probes into hash table until empty slot

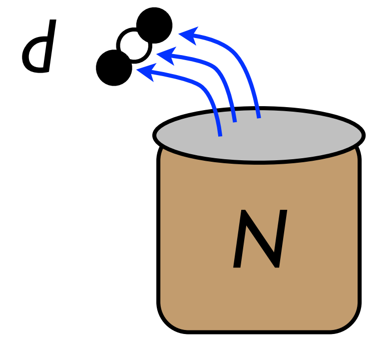
Number of wild guesses at a password until you hit it

$P(Y=k) = (1-p)^{k-1}p$; Mean $1/p$; Variance $(1-p)/p^2$

balls in urns – the hypergeometric distribution

B&T, exercise 1.61

Draw *d* balls (without replacement) from an urn containing *N*, of which *w* are white, the rest black.



Let X = number of white balls drawn

$$P(X = i) = \frac{\binom{w}{i}\binom{N-w}{d-i}}{\binom{N}{d}}, \quad i = 0, 1, \ldots, d$$

(note: n choose k = 0 if k < 0 or k > n)

E[X] = dp,   where p = w/N (the fraction of white balls)

proof: Let X_j be 0/1 indicator for j-th ball is white, $X = \Sigma X_j$

The X_j are *dependent*, but $E[X] = E[\Sigma X_j] = \Sigma E[X_j] = dp$

Var[X] = dp(1-p)(1-(d-1)/(N-1))

$N \approx 22500$ human genes, many of unknown function

Suppose in some experiment, $d = 1588$ of them were observed (say, they were all switched on in response to some drug)

A big question: What are they doing?

One idea: The Gene Ontology Consortium (www.geneontology.org) has grouped genes with known functions into categories such as "muscle development" or "immune system." Suppose 26 of your d genes fall in the "muscle development" category.

Just chance?

Or call Coach & see if he wants to dope some athletes?

Hypergeometric: GO has 116 genes in the muscle development category. If those are the white balls among 22500 in an urn, what is the probability that you would see 26 of them in 1588 draws?

Cao, et al., Developmental Cell 18, 662–674, April 20, 2010

data mining

Table 2. Gene Ontology Analysis on Differentially Bound Peaks in Myoblasts versus Myotubes

GO Categories Enriched in Genes Associated with Myotube-Increased Peaks

GOID	Term	P Value	OR[a]	Count[b]	Size[c]	Ont[d]
GO:0005856	cytoskeleton	2.05E-11	2.40	94	490	CC
GO:0043292	contractile fiber	6.98E-09	5.85	22	58	CC
GO:0030016	myofibril	1.96E-08	5.74	21	56	CC
GO:0044449	contractile fiber part	2.58E-08	5.97	20	52	CC
GO:0030017	sarcomere	4.95E-08	6.04	19	49	CC
GO:0008092	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	MF
GO:0007519	skeletal muscle development	2.50E-06	4.13	20	65	BP
GO:0015629	actin cytoskeleton	4.73E-06	3.08	27	111	CC
GO:0003779	actin binding	[illegible]	[illegible]	[illegible]	159	MF
GO:0006936	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	BP
GO:0044430	cytoskeletal part	[illegible]	[illegible]	[illegible]	294	CC
GO:0031674	I band	2.27E-05	5.67	12	32	CC
GO:0003012	muscle system process	2.54E-05	4.11	16	52	BP
GO:0030029	actin filament-based process	2.89E-05	2.73	27	119	BP
GO:0007517	muscle development	5.06E-05	2.69	26	116	BP

probability of seeing this many genes from a set of this size by chance according to the hypergeometric distribution.

E.g., if you draw 1588 balls from an urn containing 490 white balls and ≈22000 black balls, P(94 white) ≈2.05×10^{-11}

A differentially bound peak was associated to the closest gene (unique Entrez ID) measured by distance to TSS within CTCF flanking domains. OR: ratio of predicted to observed number of genes within a given GO category. Count: number of genes with differentially bound peaks. Size: total number of genes for a given functional group. Ont: the Geneontology. BP = biological process, MF = molecular function, CC = cellular component.

Supreme Court case: Berghuis v. Smith

If a group is underrepresented in a jury pool, how do you tell?

Justice Breyer [Stanford Alum] opened the questioning by invoking the binomial theorem. He hypothesized a scenario involving "**an urn with a thousand balls, and sixty are red, and nine hundred forty are black, and then you select them at random… twelve at a time.**" According to Justice Breyer and the binomial theorem, if the red balls were black jurors then "**you would expect… something like a third to a half of juries would have at least one black person**" on them.

- Justice Scalia's rejoinder: "We don't have any urns here."

Justice Breyer meets CSE 312

- Should model this combinatorially
  - Ball draws not independent trials (balls not replaced)
- Exact solution:

  $$P(\text{draw 12 black balls}) = \binom{940}{12} \Big/ \binom{1000}{12} \approx 0.4739$$

  $$P(\text{draw} \geq 1 \text{ red ball}) = 1 - P(\text{draw 12 black balls}) \approx 0.5261$$

- Approximation using Binomial distribution
  - Assume P(red ball) constant for every draw = 60/1000
  - X = # red balls drawn. X ~ Bin(12, 60/1000 = 0.06)
  - $P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0.4759 = 0.5240$

In Breyer's description, should actually expect just <u>over half</u> of juries to have at least one black person on them

Often care about 2 (or more) random variables *simultaneously*

measured X = height and Y = weight

X = cholesterol and Y = blood pressure

X_1, X_2, X_3 = work loads on servers A, B, C

Joint probability mass function:

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

Joint cumulative distribution function:

$F_{XY}(x, y) = P(X \leq x \,\&\, Y \leq y)$

Two joint PMFs

W \ Z	1	2	3
1	2/24	2/24	2/24
2	2/24	2/24	2/24
3	2/24	2/24	2/24
4	2/24	2/24	2/24

X \ Y	1	2	3
1	4/24	1/24	1/24
2	0	3/24	3/24
3	0	4/24	2/24
4	4/24	0	2/24

P(W = Z) = 3 * 2/24 = 6/24

P(X = Y) = (4 + 3 + 2)/24 = 9/24

Can look at arbitrary relationships between variables this way

Two joint PMFs

W \ Z	1	2	3	$f_W(w)$
1	2/24	2/24	2/24	6/24
2	2/24	2/24	2/24	6/24
3	2/24	2/24	2/24	6/24
4	2/24	2/24	2/24	6/24
$f_Z(z)$	8/24	8/24	8/24	

X \ Y	1	2	3	$f_X(x)$
1	4/24	1/24	1/24	6/24
2	0	3/24	3/24	6/24
3	0	4/24	2/24	6/24
4	4/24	0	2/24	6/24
$f_Y(y)$	8/24	8/24	8/24	

Marginal distribution of one r.v.: sum over the other:

$f_Y(y) = \Sigma_x f_{XY}(x,y)$ $\quad$ $f_X(x) = \Sigma_y f_{XY}(x,y)$

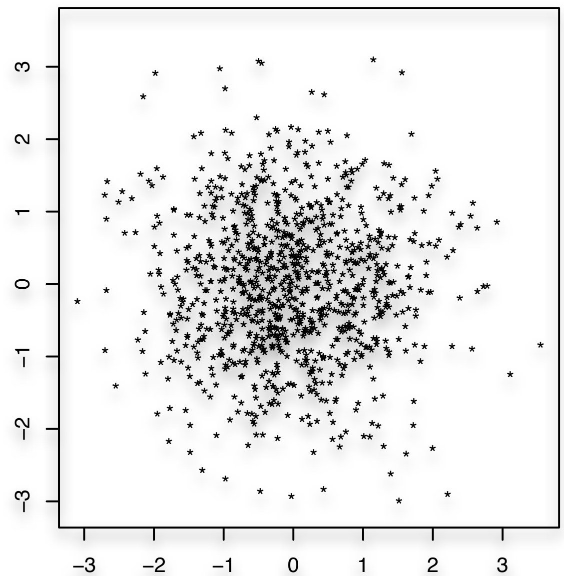
Question: Are W & Z independent? Are X & Y independent?

sampling from a (continuous) joint distribution

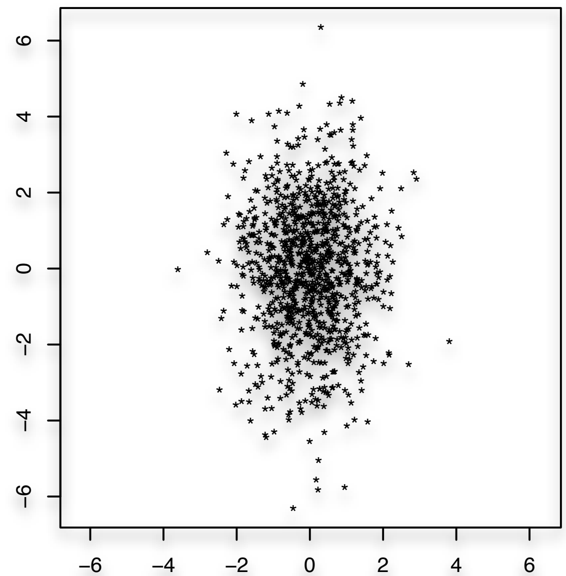
Top row; independent variables

bottom row: dependent variables

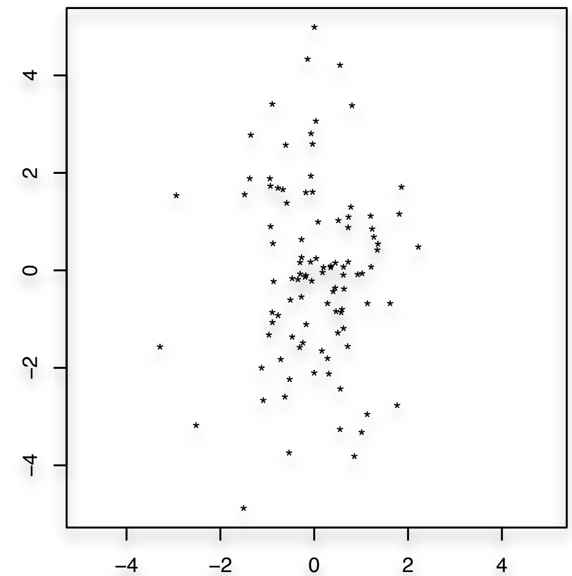
var(x)=1, var(y)=1, cov=0, n=1000



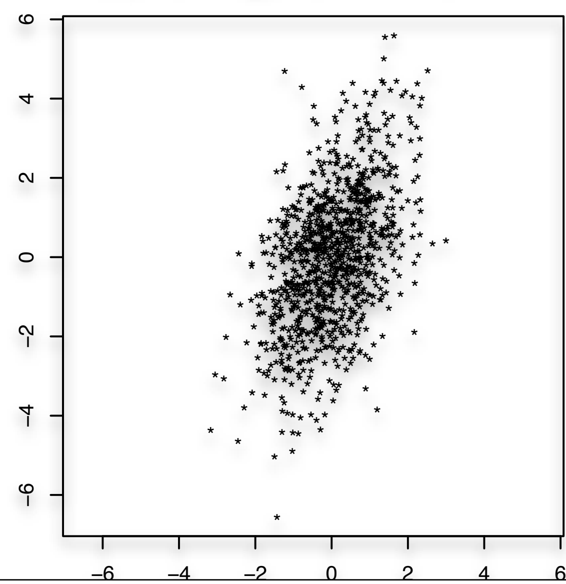
var(x)=1, var(y)=3, cov=0, n=1000



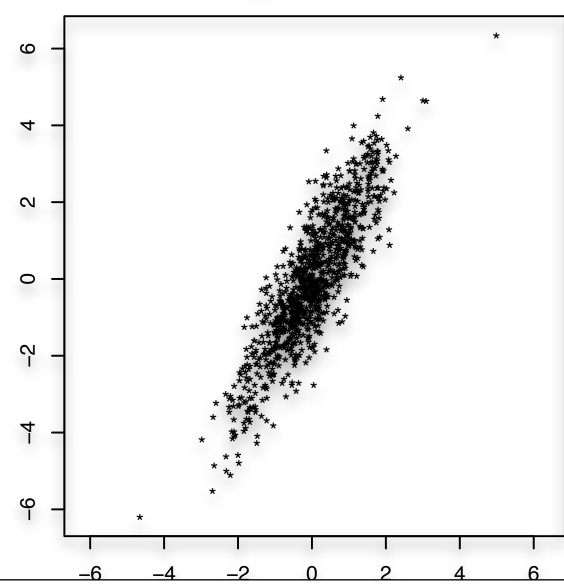
var(x)=1, var(y)=3, cov=0, n=100



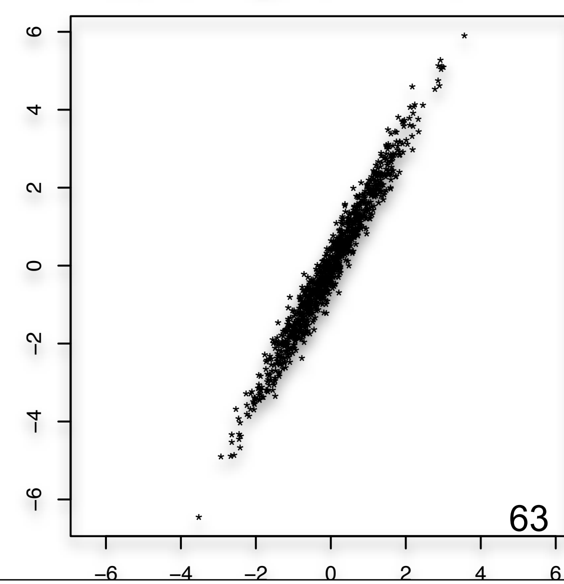
var(x)=1, var(y)=3, cov=0.8, n=1000



var(x)=1, var(y)=3, cov=1.5, n=1000



var(x)=1, var(y)=3, cov=1.7, n=1000



A function g(X,Y) defines a new random variable.

Its expectation is:

$$E[g(X, Y)] = \Sigma_x \Sigma_y\, g(x, y) f_{XY}(x,y)$$

Expectation is linear. I.e., if g is linear:

$$E[g(X, Y)] = E[a X + b Y + c] = a E[X] + b E[Y] + c$$

Example:

g(X,Y) = 2X-Y

E[g(X,Y)] = 72/24 = 3

E[g(X,Y)] = 2•2.5 - 2 = 3

X \ Y	1	2	3
1	1 • 4/24	0 • 1/24	-1 • 1/24
2	3 • 0/24	2 • 3/24	1 • 3/24
3	5 • 0/24	4 • 4/24	3 • 2/24
4	7 • 4/24	6 • 0/24	5 • 2/24

RV: a numeric function of the outcome of an experiment

Probability Mass Function p(x): prob that RV = x; $\Sigma p(x)=1$

Cumulative Distribution Function F(x): *probability that RV ≤ x*

Concepts generalize to *joint* distributions

Expectation:

of a random variable: $E[X] = \Sigma_x xp(x)$

of a function: if Y = g(X), then $E[Y] = \Sigma_x g(x)p(x)$

linearity:

$E[aX + b] = aE[X] + b$

$E[X+Y] = E[X] + E[Y]$; even if *dependent*

this interchange of "order of operations" is quite special to linear combinations. E.g. $E[XY] \neq E[X]*E[Y]$, *in general (but see below)*

Variance:

$Var[X] = E[(X-E[X])^2] = E[X^2] - (E[X])^2]$

Standard deviation: $\sigma = \sqrt{Var[X]}$

$Var[aX+b] = a^2 Var[X]$

If X & Y are *independent*, then

$E[X \cdot Y] = E[X] \cdot E[Y]$;

$Var[X+Y] = Var[X]+Var[Y]$

(These two equalities hold for *indp* rv's; but not in general.)