Covariance, Correlation Markov and Chebyshev

Discrete random variables

Probability associated with each combination of values

	X=0 0.4	X=1 0.3	X=2 0.3
Y=0 0.6	0.3	0.2	0.1
Y=1 0.4	0.1	0.1	0.2

Dependent Random Variables

Probability associated with each combination of values

marginal distributions

		X=0	X=1	X=2
		0.4	0.3	0.3
Y=0	0.6	0.3	0.2	0.1
Y=1	0.4	0.1	0.1	0.2

Independent random variables

$$\forall x, y \quad P(X = x \wedge Y = y) = P(X = x)P(Y = y)$$

Independent random variables

$$\forall x, y \quad P(X = x \wedge Y = y) = P(X = x)P(Y = y)$$

	X=0 0.2	X=1 0.1	X=2 0.7
Y=0 0.4	0.08	0.04	0.28
Y=1 0.6	0.12	0.06	0.42

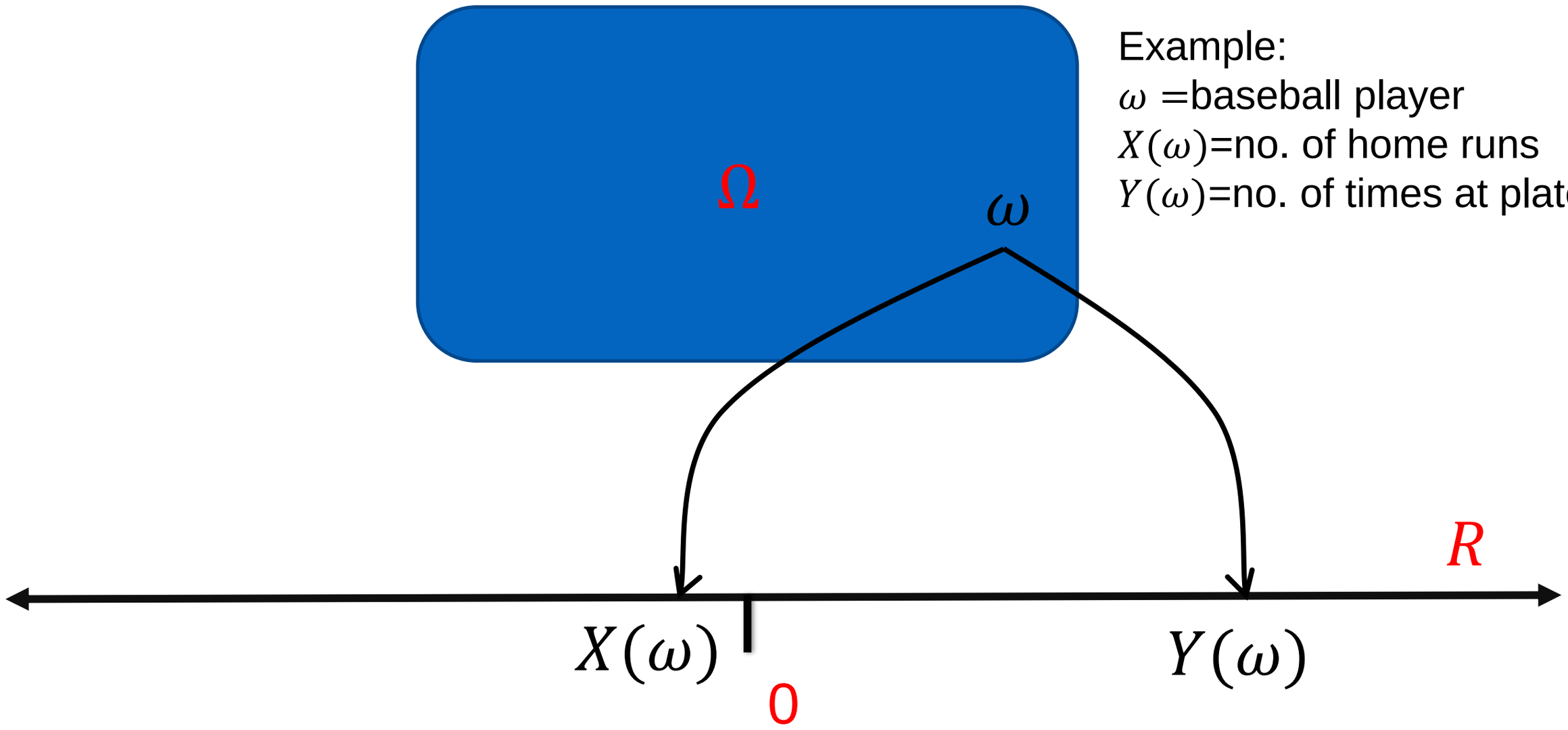
Independent random variables

$$\forall x, y \quad P(X = x \wedge Y = y) = P(X = x)P(Y = y)$$

marginal distributions

		X=0	X=1	X=2
		0.2	0.1	0.7
Y=0	0.4	0.08	0.04	0.28
Y=1	0.6	0.12	0.06	0.42

A random variable is a function from the sample space Ω to the real line R



Example:
ω =baseball player
$X(\omega)$=no. of home runs
$Y(\omega)$=no. of times at plate

Sum of two random variables

- The sum of two random variables is a random variable:

$$S(\omega) = X(\omega) + Y(\omega)$$

- The expected value is defined to be:

$$E(X) \doteq \sum_{i=1}^{n} X(\omega_i)P(\omega_i), \;\; E(Y) \doteq \sum_{i=1}^{n} Y(\omega_i)P(\omega_i), \qquad E(S) \doteq \sum_{i=1}^{n} S(\omega_i)P(\omega_i),$$

- We can prove the relation $E(X+Y) = E(X) + E(Y)$ in the following way:

- $E(X+Y) = E(S) =$

$$= \sum_{i=1}^{n} S(\omega_i)P(\omega_i) = \sum_{i=1}^{n} \big(X(\omega_i) + Y(\omega_i)\big)P(\omega_i) =$$

$$= \sum_{i=1}^{n} X(\omega_i)P(\omega_i) + \sum_{i=1}^{n} Y(\omega_i)P(\omega_i) = E(X) + E(Y)$$

Product of two random variables

- The product of two random variables is a random variable:

$$M(\omega) = X(\omega) \times Y(\omega)$$

- The expected value is defined to be:

$$E(X) \doteq \sum_{i=1}^{n} X(\ \omega_i) P(\omega_i),\ \ E(Y) \doteq \sum_{i=1}^{n} Y(\omega_i) P(\omega_i), \qquad E(M) \doteq \sum_{i=1}^{n} M(\omega_i) P(\omega_i),$$

- Lets analyze $E(XY)$:

$$E(XY) = E(M) = \sum_{i=1}^{n} M(\omega_i) P(\omega_i) = \sum_{i=1}^{n} \big(X(\omega_i) Y(\omega_i)\big) P(\omega_i) = *$$

- $P(\omega_i)$ can be replaced by $\mathrm{P}(X(\omega) = x\ and\ Y(\omega) = y)$. By summing over the possible values of x, y rather than over the outcomes ω_i . Using these3 observations we e can rewrite the last expression as

$$* = \sum_{x,y} xy\ \mathrm{P}(X(\omega) = x\ and\ Y(\omega) = y) = \#$$

- If $X(\omega),\ Y(\omega)$ are independent random variables then

$$P(X(\omega) = x\ and\ Y(\omega) = y) = P(X(\omega) = x) \times\ P(Y(\omega) = y)$$

- Which implies:

$$\# = \sum_{x,y} xy\ \mathrm{P}(X(\omega) = x\ and\ Y(\omega) = y) = \sum_{x,y} xy\ \mathrm{P}(X(\omega) = x)\ P(Y(\omega) = y) =$$

$$= \sum_{x} x\ \mathrm{P}(X(\omega) = x) \times \sum_{y} y\ \mathrm{P}(Y(\omega) = y) = E(X)E(Y)$$

Expected value for a product of independent RVs

$$E(XY) = \sum_x \sum_y xyP(X = x \wedge Y = y) =$$

Expected value for a product of independent RVs

$$E(XY) = \sum_x \sum_y xyP(X = x \wedge Y = y) =$$

$$= \sum_x \sum_y xyP(X = x)P(Y = y) = \sum_x xP(X = x) \sum_y yP(Y = y) =$$

Expected value for a product of independent RVs

$$E(XY) = \sum_x \sum_y xyP(X = x \wedge Y = y) =$$

$$= \sum_x \sum_y xyP(X = x)P(Y = y) = \sum_x xP(X = x) \sum_y yP(Y = y) =$$

$$= E(X)E(Y)$$

Covariance

Recall $\mu_X \doteq E(X),\ \mu_Y \doteq E(Y)$

$$\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) =$$

$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y = E(XY) - E(X)E(Y)$$

$$= E\big((X-\mu_X)(X-\mu_X)\big) =$$

Recall $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \mathrm{Cov}(X,X)$

$\mathrm{Cov}(X,Y) \neq 0$ implies that X and Y are not independent

but $\mathrm{Cov}(X,Y)=0$ **does not imply** that X and Y are independent

Why do we need Corr in addition to Cov ?

- Suppose Cov(X,Y)=3 and Cov(X,Z)=1
- Is X more correlated with Y than with Z?
- Not necessarily, it might be that Y=3Z
- We want to have a measure of correlation that is independent of scaling, i.e.

$$\forall a > 0, b > 0: \ \ Corr(aX, bY) = Corr(X, Y)$$

Correlation coefficient

$$\mathrm{Corr}(X,Y) \doteq \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Corr(aX+c,bY+d)=Corr(X,Y) if a,b>0
- Corr(X,Y) varies from -1 to +1
- Corr(X,Y)>0 ⇔ X and Y are "correlated"
- Corr(X,Y)<0 ⇔ X and Y are "anti-correlated"
- Corr(X,Y)=1 ⇔ X=aY, a>0
- Corr(X,Y)=-1 ⇔ X=aY a<0

Correlation coefficient

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

- The covariance depends on scaling and units, correlation coefficient does not

$$\forall a > 0, b > 0 \qquad Corr(aX,bY) = Corr(X,Y)$$

- The correlation coefficient varies between -1 and 1.

correlation vs. dependence

- If X, Y are independent then $Cov(X, Y) = 0$ and $Corr(X, Y) = 0$
- If $Corr(X, Y) \neq 0$ then X, Y are dependent.
- No implications in the opposite directions
- If X is a random variable that takes n discrete values, and Y is a random variable that takes m discrete values, Then checking for independence requires checking nm values.
- Checking for zero correlation requires calculation of just one value.
- Correlation is the quick and dirty way to detect strong dependencies, but it cannot find them all.

Examples

	X=1	X=2	X=3	X=4
Y=1	1/4	1/4	0	0
Y=2	0	0	0	0
Y=3	0	0	1/4	1/4

$$\mu(X)=2.5,\ \mu(Y)=2$$

$$\text{cov}(X,Y)=\frac{1}{4}(-1.5*-1)+\frac{1}{4}(-.5*-1)+\frac{1}{4}(.5*1)+\frac{1}{4}(1.5*1)=1$$

	X=1	X=2	X=3	X=4
Y=1	0	0	0	1/4
Y=2	0	1/4	1/4	0
Y=3	1/4	0	0	0

$$\mu(X)=2.5,\ \mu(Y)=2$$

$$\text{cov}(X,Y)=\frac{1}{4}(-1.5*1)+\frac{1}{4}(-.5*0)+\frac{1}{4}(.5*9)+\frac{1}{4}(1.5*-1)=-\frac{3}{4}$$

	X=1	X=2	X=3	X=4
Y=1	1/4	0	0	1/4
Y=2	0	0	0	0
Y=3	1/4	0	0	1/4

$$\mu(X)=2.5,\ \mu(Y)=2$$

$$\operatorname{cov}(X,Y)=\frac{1}{4}(-1.5*1)+\frac{1}{4}(-1.5*-1)+\frac{1}{4}(1.5*1)+\frac{1}{4}(1.5*-1)=0$$

P(X=1)=P(X=4)=1/2, P(Y=1)=P(Y=3)=1/2

X and Y are independent because all of the joint probabilities are either 0 or 1/4

	X=1	X=2	X=3	X=4
Y=1	1/8	0	0	1/8
Y=2	0	1/4	1/4	0
Y=3	1/8	0	0	1/8

1. Cov(X,Y)=0
2. X and Y are independent

A. 1 and 2 B. 1 and not 2 C. not 1 and 2 D. not 1 and not 2

Back to convergence of averages

Estimating click-through rates

- Suppose that we have two ads and we know that one has click-through-rate of 2%, and the other has click-through-rate of 1.75%, but we don't know which is which.

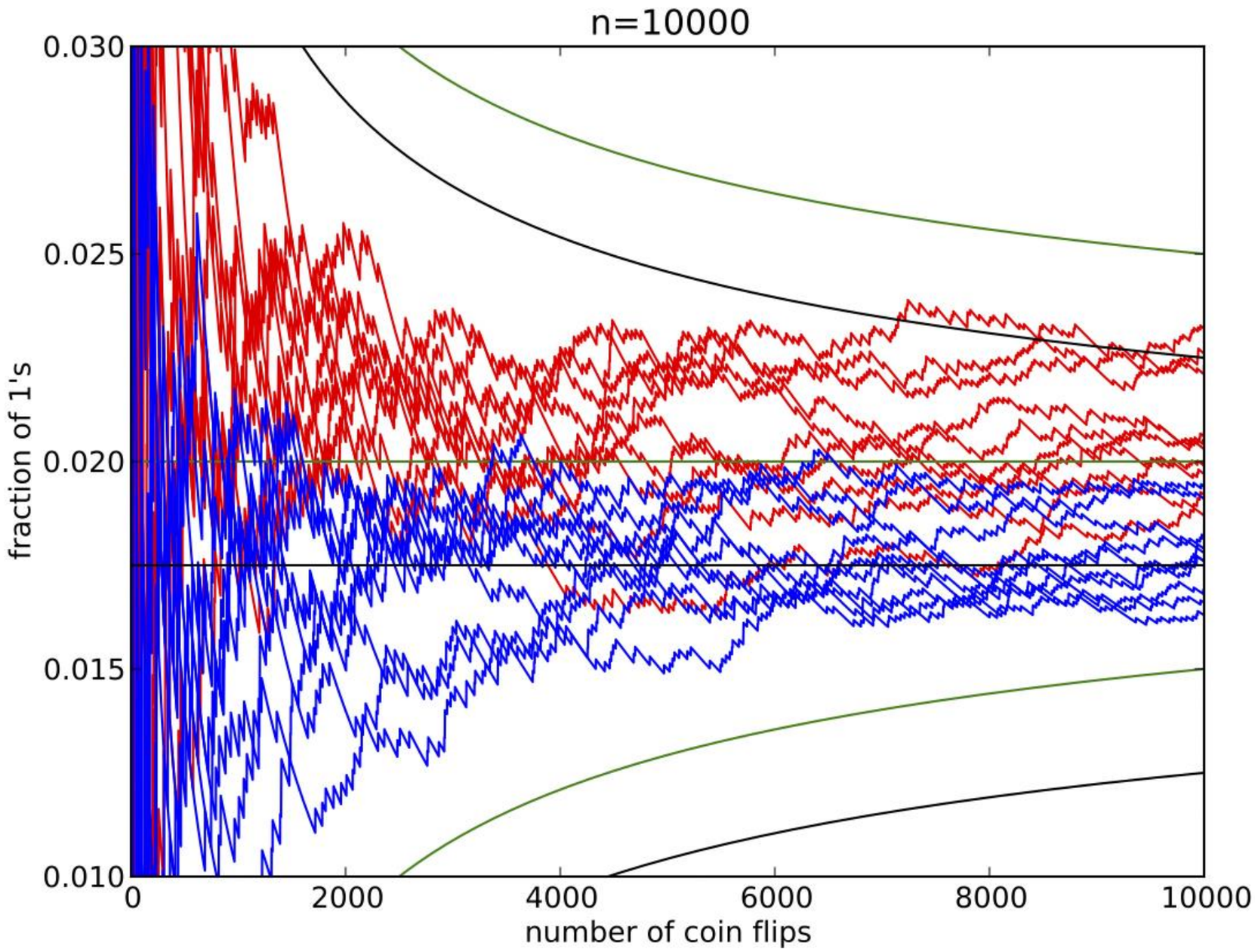
- We alternate presenting the two ads.

- How many presentations are needed in order to know, with confidence, which ad has the higher click-through-rate.

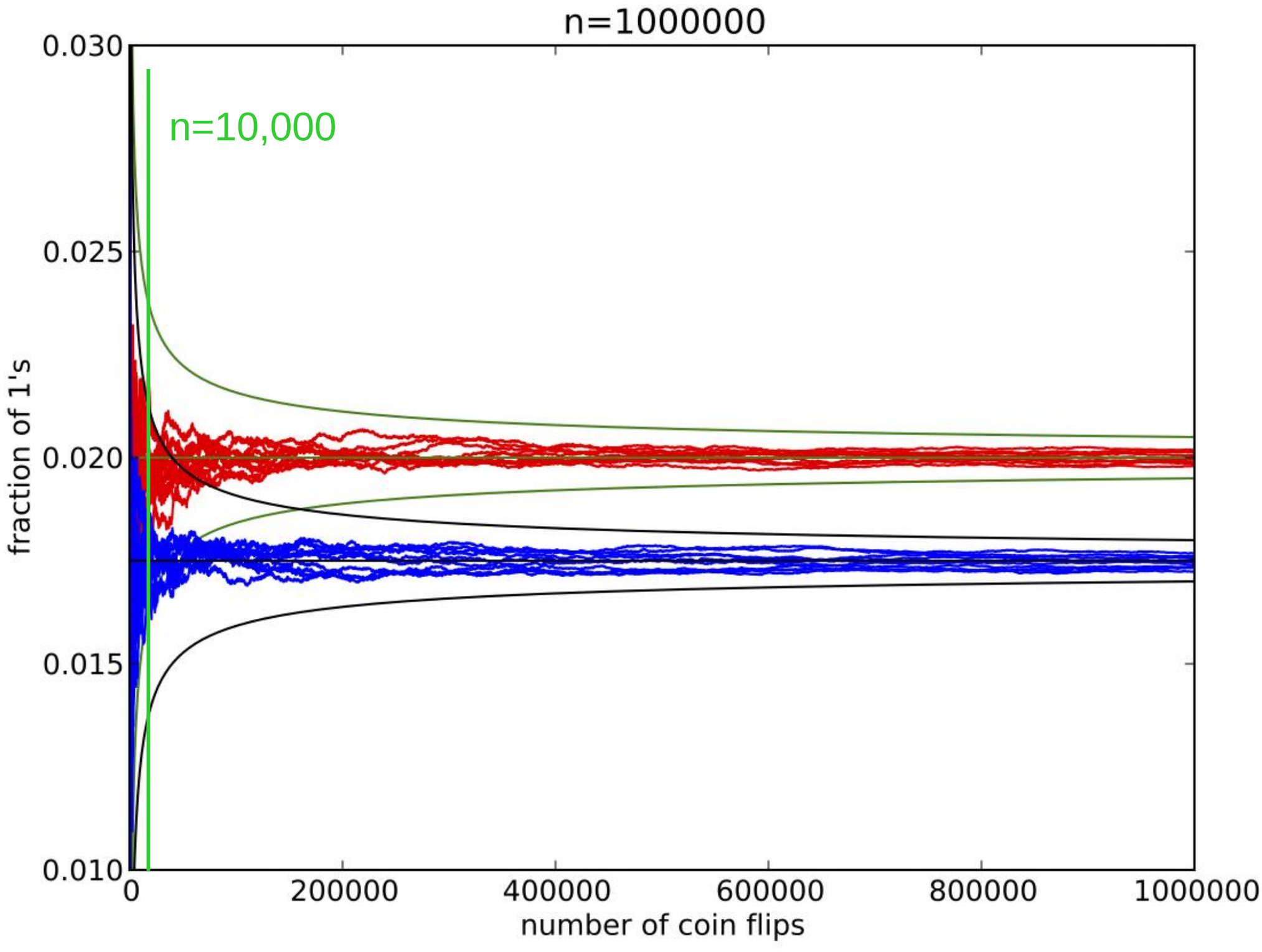
Running averages after 10,000 trials

The smooth green and black curves define the “envelope” of likely sequences

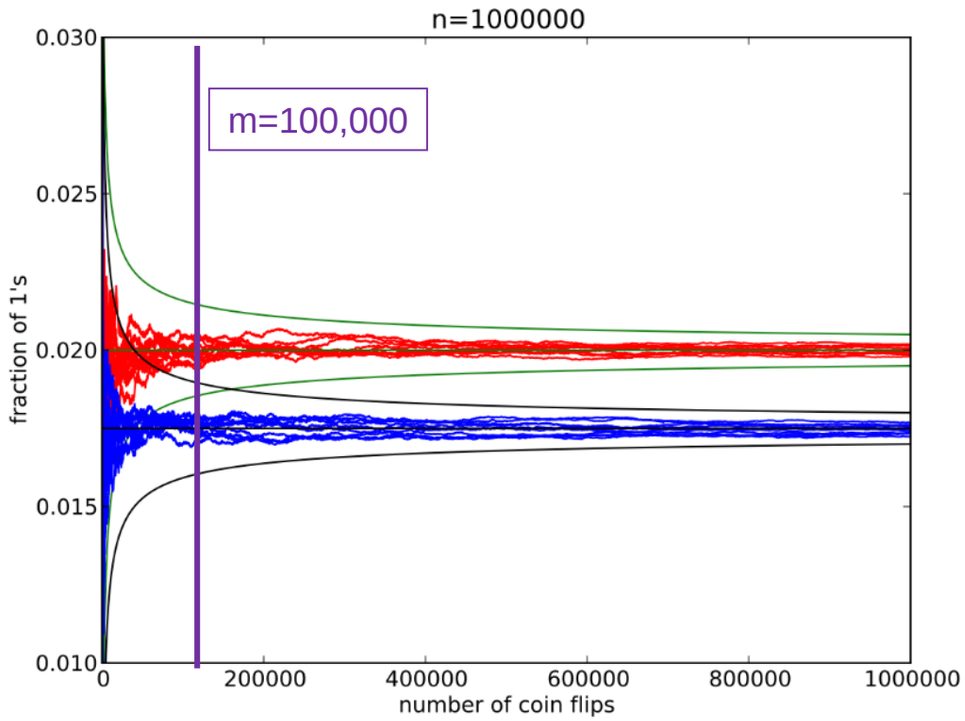
Each jagged line is the running average for one sequence



Running average after 1,000,000 trials



Effect of increasing the gap



- Suppose the two probabilities are 1% and 2%, instead of 1.75% and 2%
- Gap is 4 times larger
- how many examples do we need to get clear separation?

(a) $\mathrm{m} - 4$ (b) m^2

(c) $\frac{m}{4^2}$ (d) $\frac{m}{4}$

Convergence to the Mean
Take 1

Using the variance

The average also called the empirical mean

Suppose $X_1, X_2, \ldots, X_n$ are

independent identically distributed (IID) random variables

$$\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p, \qquad 0 \le p \le 1$$

$$E[X_i] = 1 \times p + 0 \times (1 - p) = p$$

We define the average to be another random variable

$$S_n \doteq \frac{1}{n} \sum_{i=1}^{n} X_i$$

We already know that

$$E[S_n] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} p = p$$

We want to show that S_n tends to be close to p

Approach 1: using the variance

$$S_n \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E[S_n] = E[X_i] = p$$

$$Var[X_i] = p \times (1-p)^2 + (1-p) \times (0-p)^2$$

$$= (1-p+p) \times (1-p) \times p = p(1-p)$$

As X_i are IID:

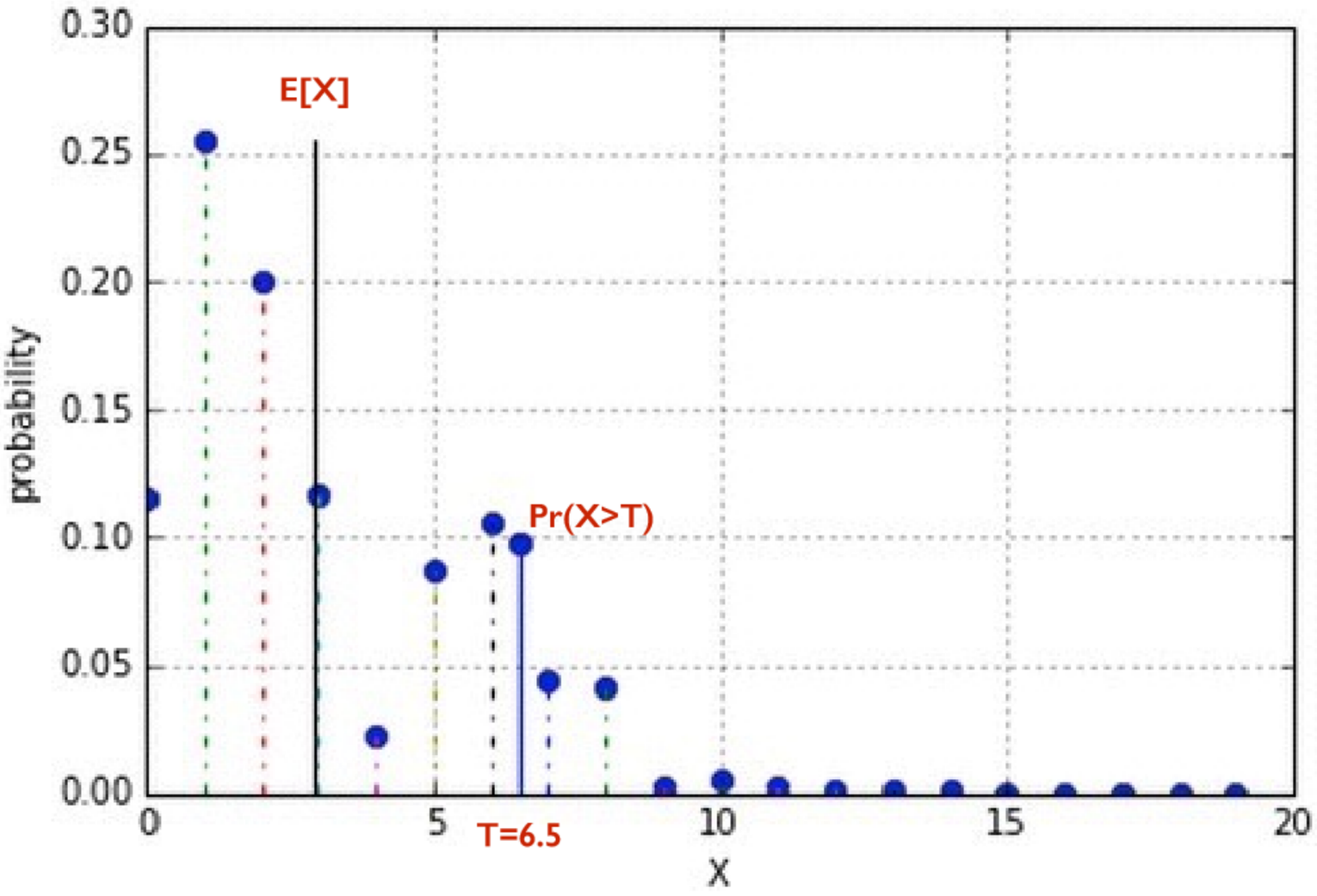
$$Var[S_n] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] =$$

$$\frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$\sigma(S_n) = \sqrt{\frac{p(1-p)}{n}}$$

Detour 1: Markov Bound

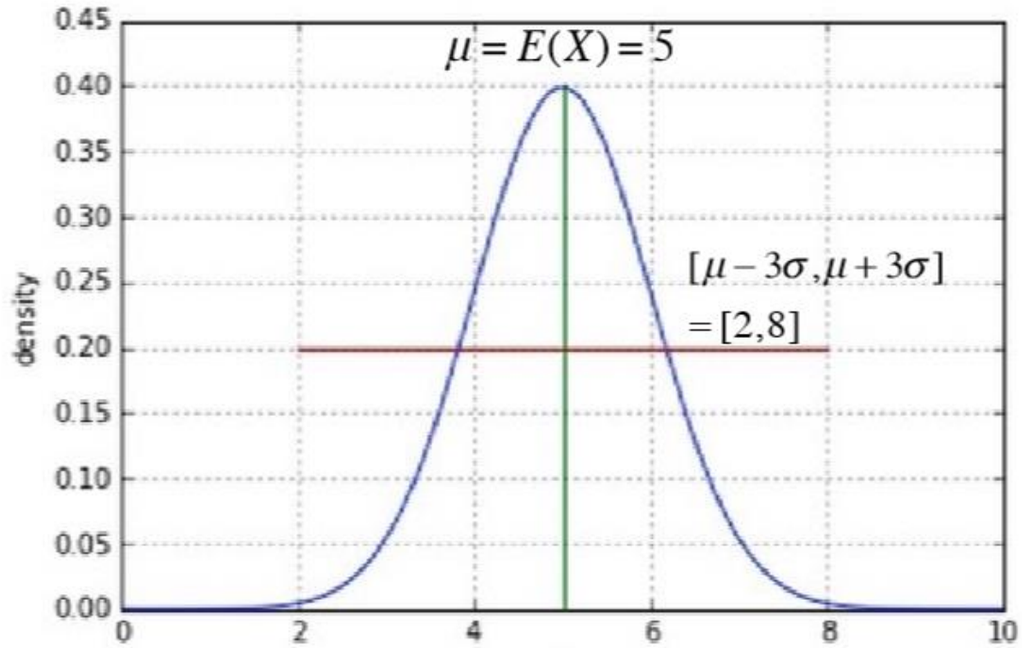
- Suppose the RV X is distributed over the **non-negative** integers 0,…,20
- Suppose we know the mean E[X]. Can we bound the probability that X>T ?



$$E[X] \geq 0 \times \Pr(X < T) + T \times \Pr(X \geq T)$$

$$\Pr(X \geq T) \leq \frac{E(X)}{T}$$

Detour 2: Chebyshev's bound

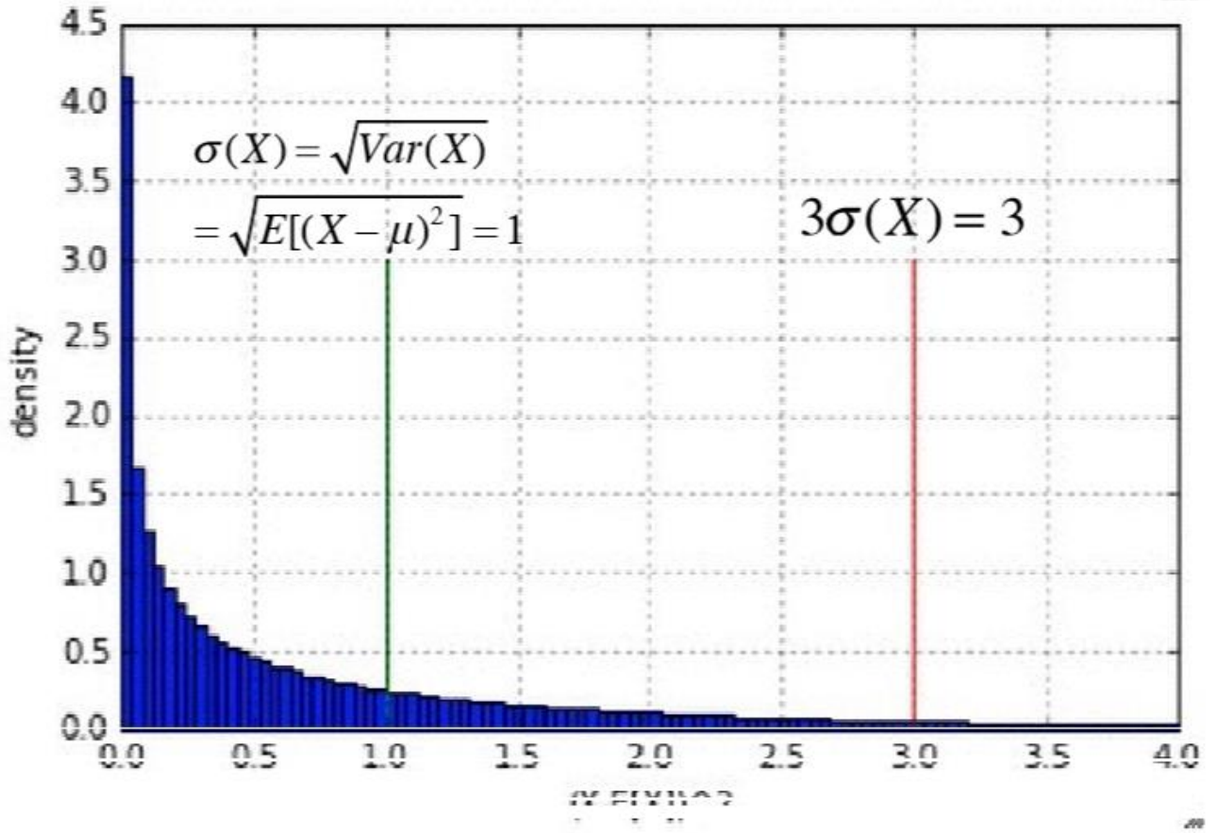


$$\Pr\left((X-\mu)^2 \geq \lambda^2\right) \leq \frac{E\left[(X-\mu)^2\right]}{\lambda^2} = \frac{Var(X)}{\lambda^2}$$

Plugging in $\lambda = k\sigma(X)$

$$\Pr\left[\left|X-\mu\right| \geq k\sigma(X)\right] \leq \frac{\sigma(X)^2}{k^2\sigma(X)^2} = \frac{1}{k^2}$$

In this example:
[\mu-
\mu=E



In the example shown

$$\mu = E(X) = 5$$

$$\sigma = \sqrt{Var(X)} = 1$$

We choose $k = 3$ to get that

$$\Pr\left(\left|X-5\right| \geq 3\right) \leq \frac{1}{k^2} = \frac{1}{9}$$

Applying Chebishev's bound

$$\Pr\left[\left|X-\mu\right| \geq k\sigma(X)\right] \leq \frac{\sigma(X)^2}{k^2\sigma(X)^2} = \frac{1}{k^2}$$

A few slides ago, we found that

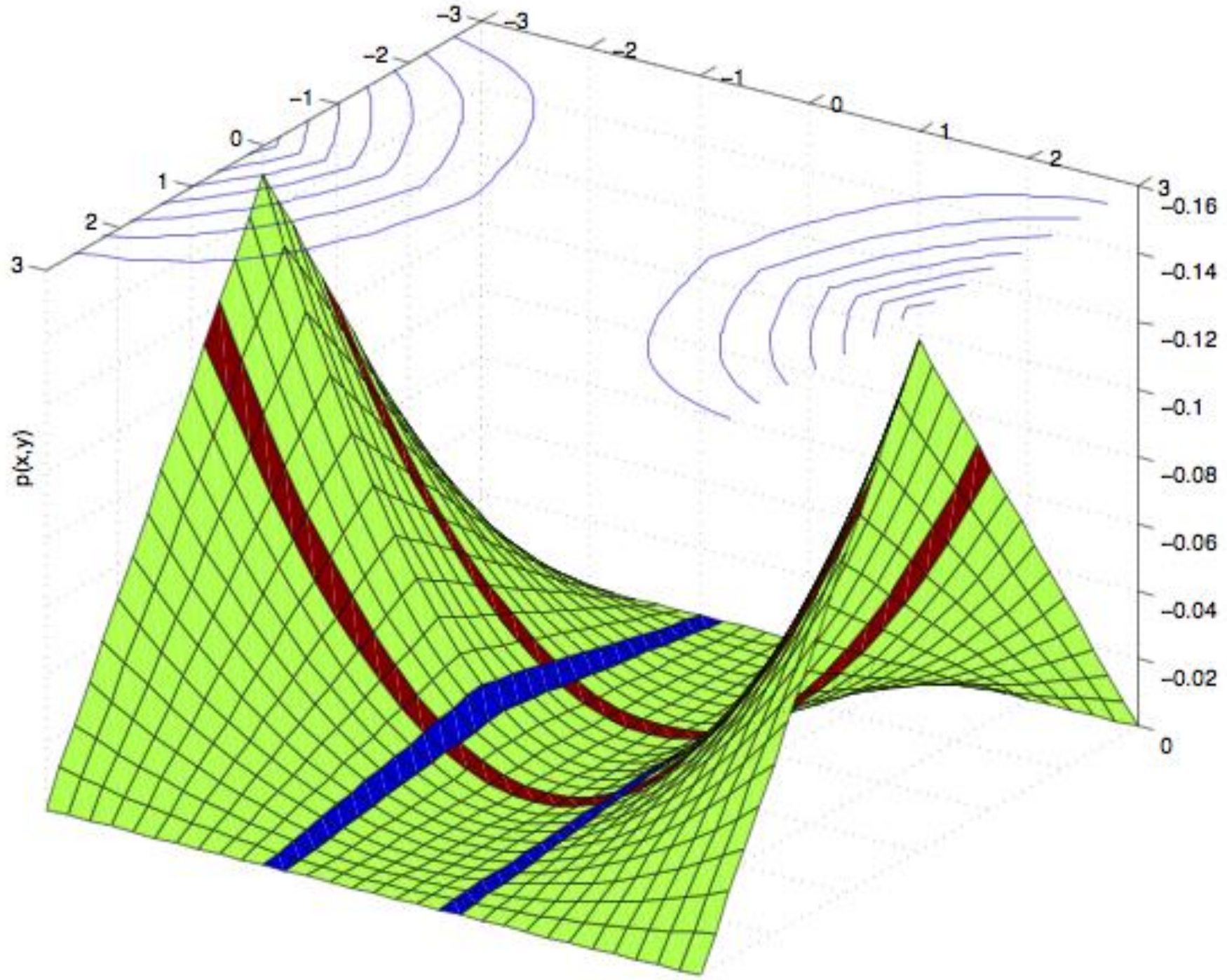
$$\mu(S_n) = p; \quad \sigma(S_n) = \sqrt{\frac{p(1-p)}{n}}$$

$$\Pr\left[\left|S_n - p\right| \geq k\sqrt{\frac{p(1-p)}{n}}\right] \leq \frac{1}{k^2}$$

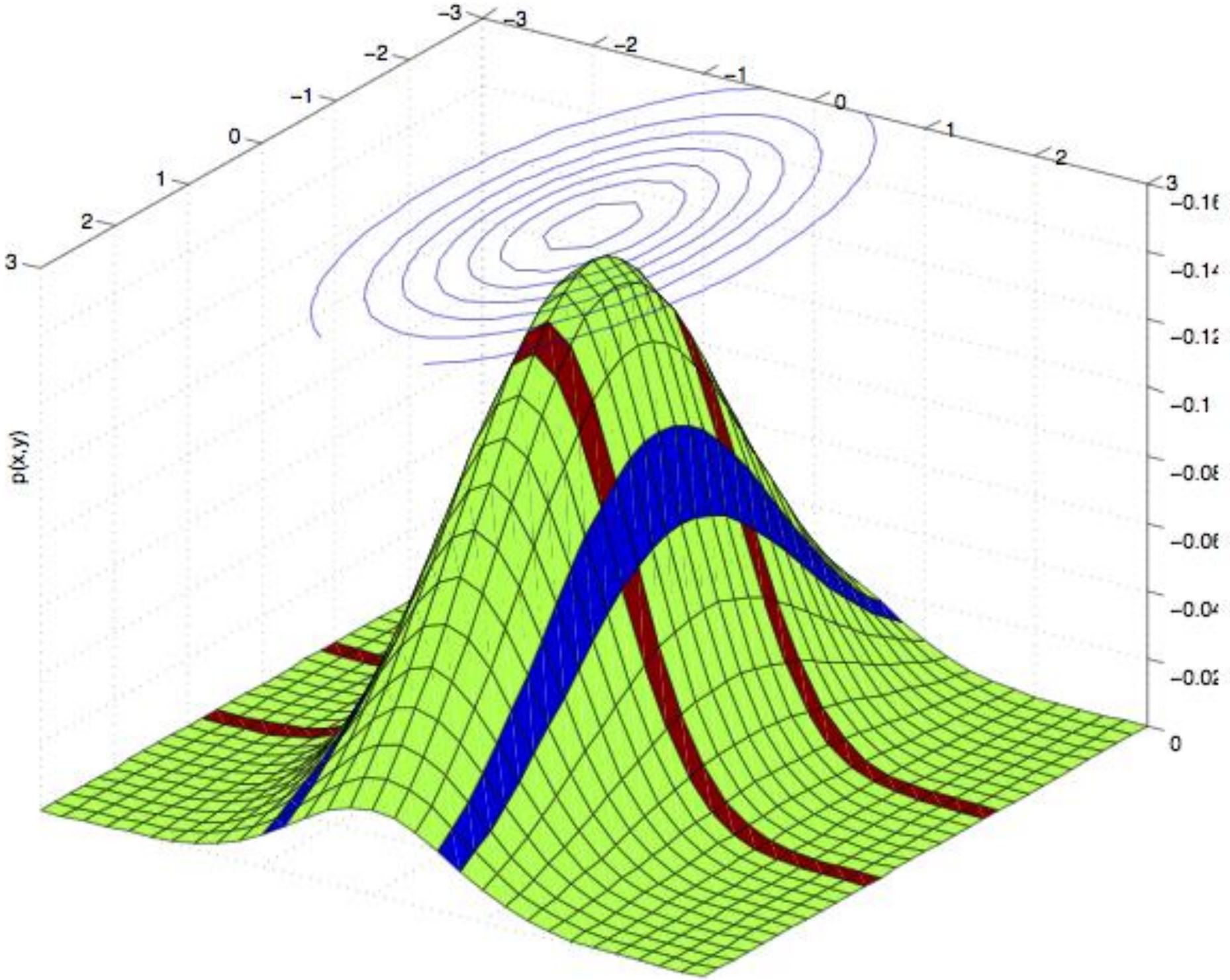
fixing k and letting n increase

Correlation and independence in uncountably infinite spaces.

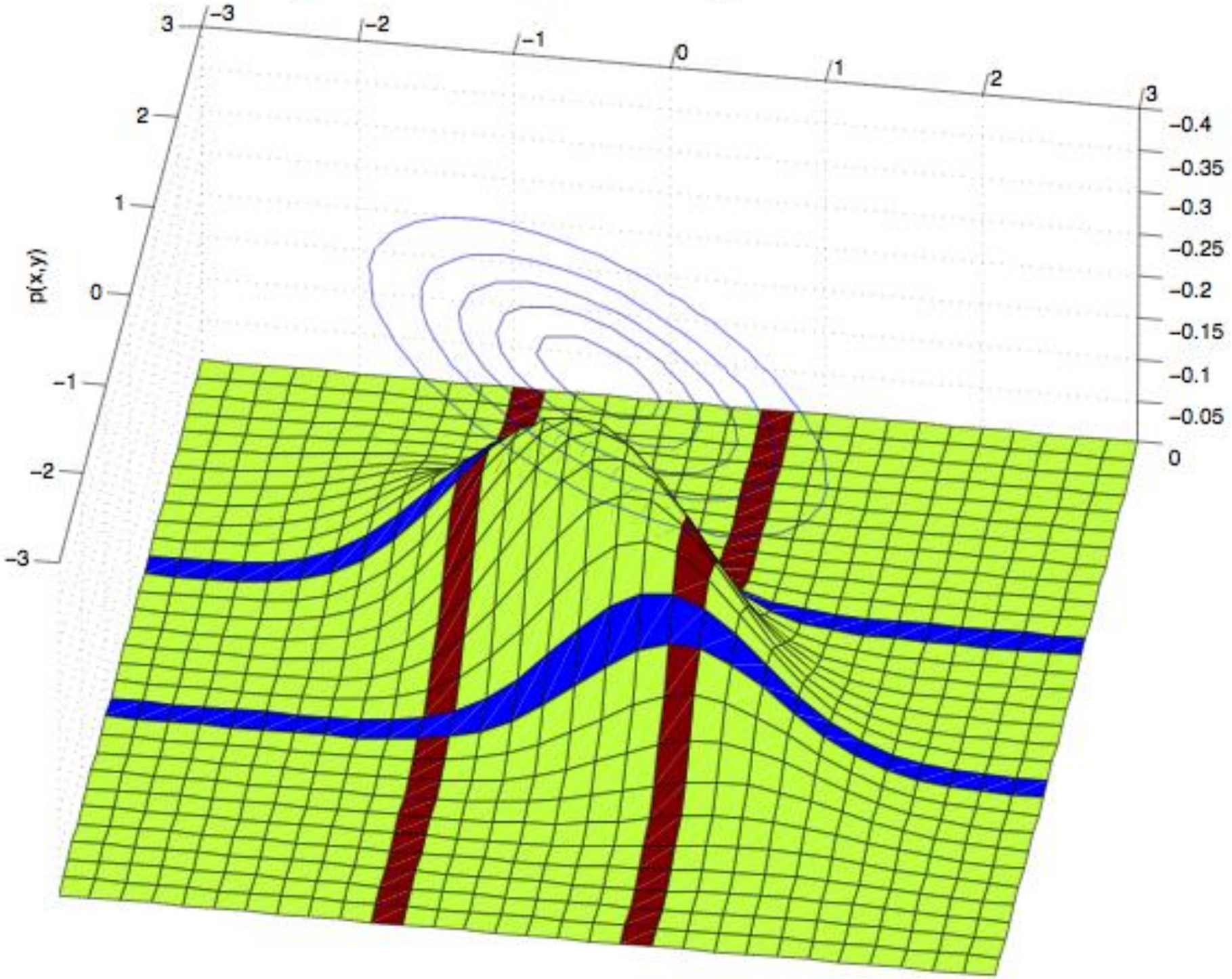
Example I, independent RVs



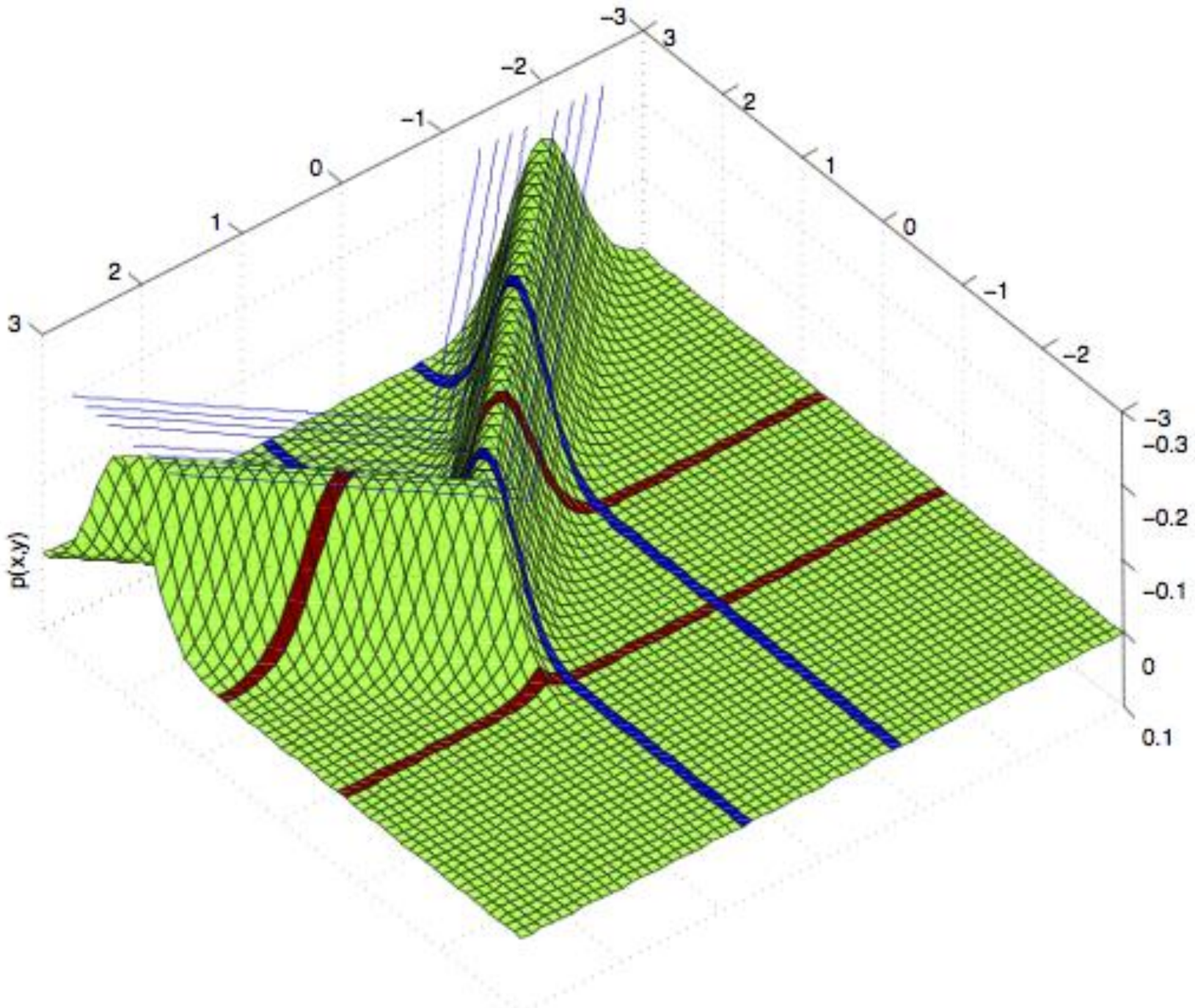
Example 2 independent RVs



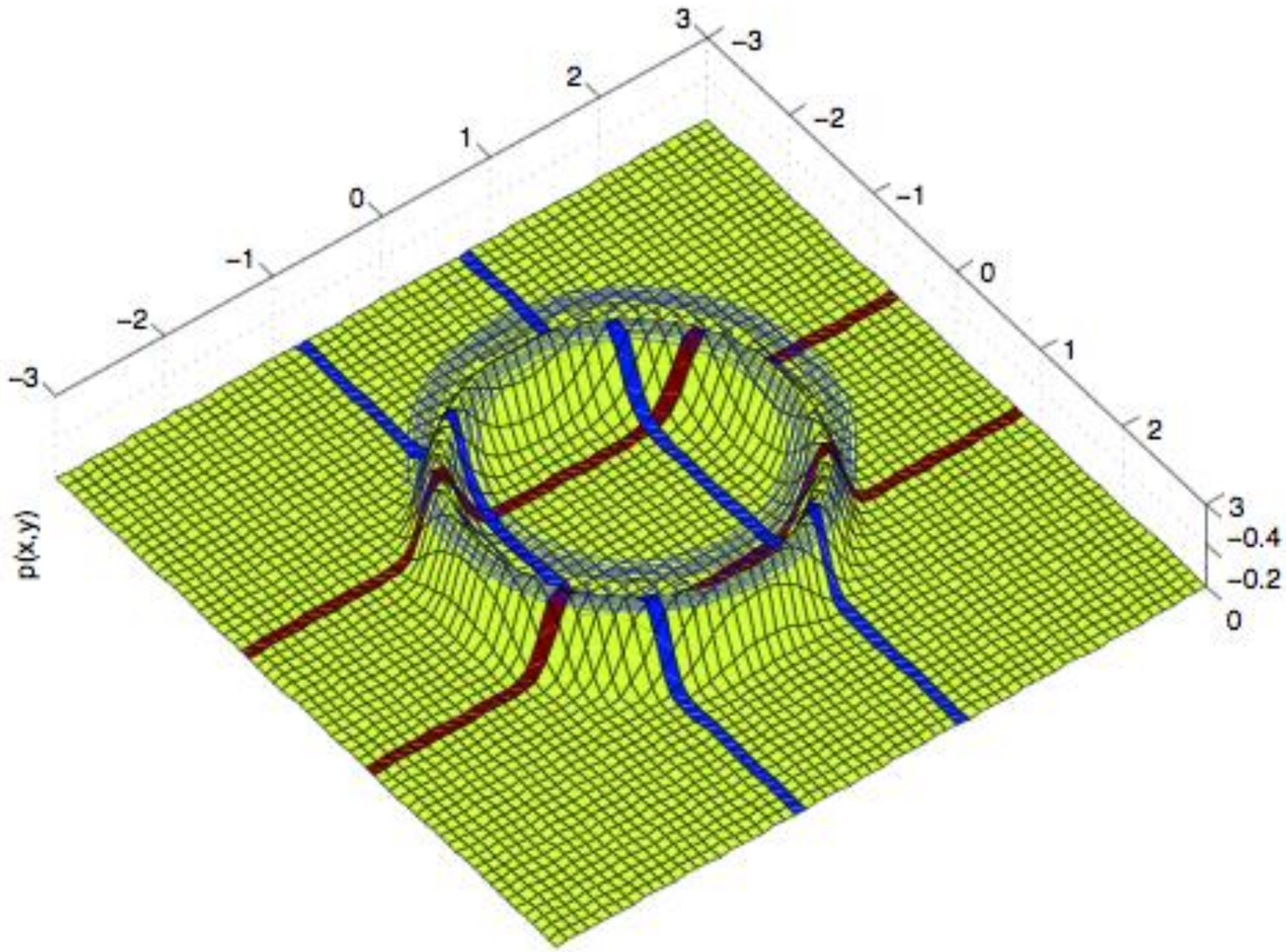
Example 3, Dependent RVs



Example 4, functional dependence



Example 5, Circle



Correlation vs. Dependence

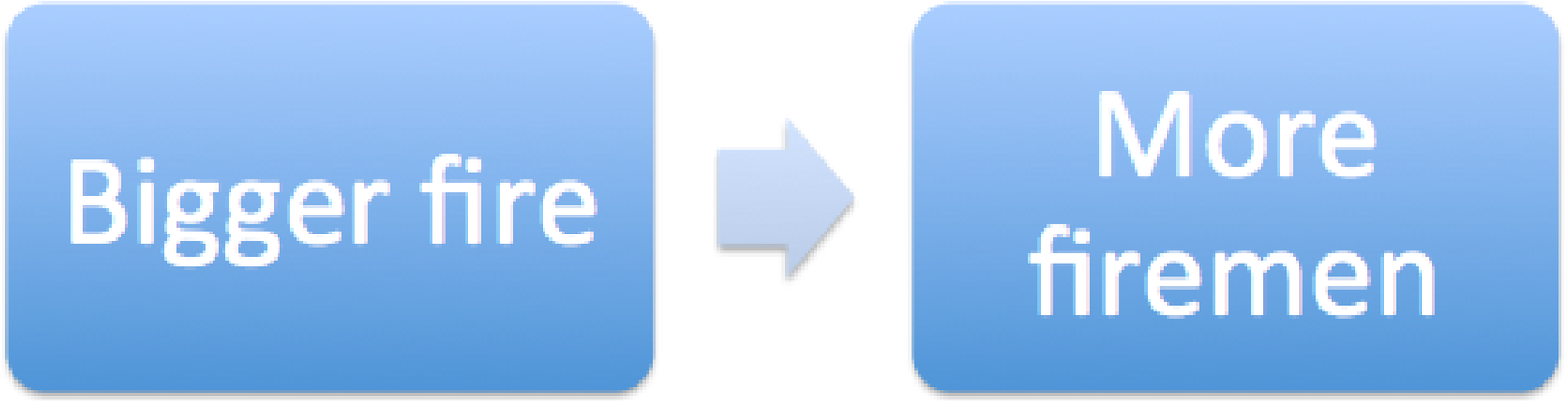
- Non-zero Correlation implies dependence
- Dependence does not imply correlation

Correlation vs Causation

- Using correlation because common, same can be said regarding Dependence vs. causation.
- The simple case is: the number of mosquitoes is correlated with the number of malaria cases. Therefor mosquitoes cause malaria. Which is true.
- However, one can deduce that malaria causes mosquitoes, which is false.

Correlation vs. causation 1

- The more firemen fighting a fire, the bigger the fire.
- Therefore firemen cause an increase in the size of a fire.



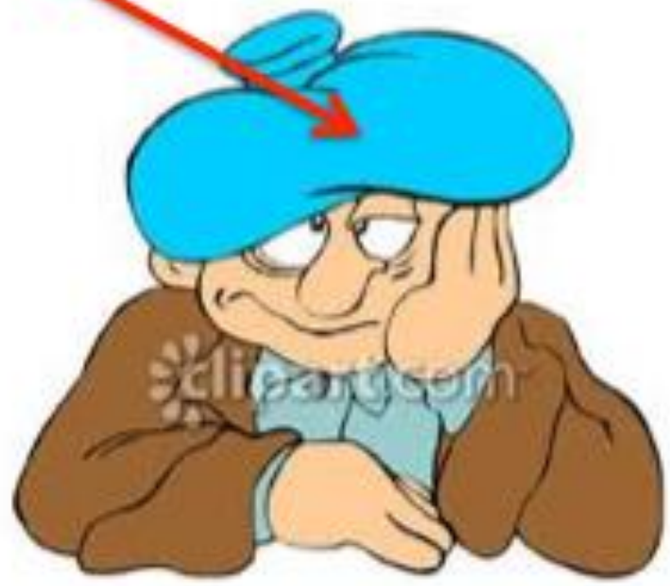
- Causation reversal. Correlation cannot distinguish between **A causes B** and **B causes A**

Correlation vs. causation 2

Excessive Drinking

Sleeping with shoes on

Common Cause



Morning Headache

Correlation vs. Causation 3

- For an ideal gas in a fixed volume, temperature is correlated with pressure.
- Gas, volume and temperature are related by the equation PV=nRT.
- Pressure and Temperature or co-dependent.
- Causation is bi-directional or not well defined.

Determining causation

- Can be very hard.
- Usually required intervention
- How can you do determine whether or not sleeping with shoes causes headaches?
  - A. Stop drinking.
  - B. Flip a coin to decide whether to wear shoes to bed.
  - C. Flip a coin to decide whether or not to drink.
  - D. Observe that every time you drank, you both slept with shoes and got up with a headache.

What is done in practice?

- Given random variables $X_1,\ldots,X_n$ and their joint distribution, we want to identify causal relationships. (For example, the causes for a particular disease).
- We perform a correlation analysis, computing the correlation for each pair Xi,Xj
- Sometimes, we know the causation direction, for example, a mutation in DNA causes a change in the protein and not vice versa.
- We pick the pairs with strongest correlations and use additional experiments to identify the causes.

Empirical contingency tables

n=100
std~1/10

independent

true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2600	0.4800	0.2600
Male	0.3900	0.1000	0.2100	0.0800
Female	0.6100	0.1600	0.2700	0.1800

empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2200	0.3900	0.3900
Male	0.4400	0.1100	0.1400	0.1900
Female	0.5600	0.1100	0.2500	0.2000

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2200	0.5000	0.2800
Male	0.3500	0.1000	0.1200	0.1300
Female	0.6500	0.1200	0.3800	0.1500

dependent

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1	0.2700	0.5200	0.2100
Male	0.4800	0.1200	0.2200	0.1400
Female	0.5200	0.1500	0.3000	0.0700

	marginal	A-average	B-average	C-average
marginal	1	0.2600	0.4300	0.3100
Male	0.4700	0.0700	0.2100	0.1900
Female	0.5300	0.1900	0.2200	0.1200

	marginal	A-average	B-average	C-average
marginal	1	0.1800	0.4000	0.4200
Male	0.5500	0.1000	0.1600	0.2900
Female	0.4500	0.0800	0.2400	0.1300

Empirical contingency tables

n=10,000
std~1/100

independent

true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2473	0.5062	0.2465
Male	0.4021	0.0977	0.2052	0.0992
Female	0.5979	0.1496	0.3010	0.1473

empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2530	0.4943	0.2527
Male	0.3925	0.0982	0.1942	0.1001
Female	0.6075	0.1548	0.3001	0.1526

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2457	0.5005	0.2538
Male	0.3893	0.0936	0.1945	0.1012
Female	0.6107	0.1521	0.3060	0.1526

dependent

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4064	0.2936
Male	0.4991	0.0958	0.2052	0.1981
Female	0.5009	0.2042	0.2012	0.0955

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3058	0.3895	0.3047
Male	0.5044	0.1023	0.1989	0.2032
Female	0.4956	0.2035	0.1906	0.1015

	marginal	A-average	B-average	C-average
marginal	1	0.2984	0.4047	0.2969
Male	0.4970	0.0997	0.1938	0.2035
Female	0.5030	0.1987	0.2109	0.0934

Empirical contingency tables

n=1,000,000
std~1/1,000

independent

true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.4998	0.2502
Male	0.3998	0.1001	0.2000	0.0997
Female	0.6002	0.1499	0.2998	0.1505

empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2504	0.4997	0.2499
Male	0.3991	0.1000	0.1993	0.0999
Female	0.6009	0.1504	0.3004	0.1500

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2504	0.5000	0.2497
Male	0.4005	0.1005	0.2003	0.0996
Female	0.5995	0.1498	0.2997	0.1500

dependent

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1	0.2997	0.4000	0.3002
Male	0.4997	0.0994	0.2000	0.2003
Female	0.5003	0.2003	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1	0.3006	0.3996	0.2997
Male	0.5002	0.1000	0.2001	0.2001
Female	0.4998	0.2006	0.1995	0.0997

	marginal	A-average	B-average	C-average
marginal	1	0.3001	0.3998	0.3001
Male	0.4995	0.1000	0.1998	0.1997
Female	0.5005	0.2001	0.2000	0.1004

1. It is possible to prove dependence, it is impossible to prove independence.
2. Checking for dependence using empirical contingency tables
   requires a good estimate for the probability in each of the cells.
3. The estimation error decreases like 1/sqrt(n) with the number of
   samples.
4. The estimation error increaeses like sqrt(rows x columns)
5. the random variables have density distributions, the probability of getting
   the same value twice is zero and the whole method breaks down.
   ---> one cannot check whether two continuous random variables are
   indpendent.
   ---> In practice, the same holds if the number of cells is moderately high (>100)

We need a different way to check whether continuous random variables are dependent.

Lets look at E(X*Y)