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.

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

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

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

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

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

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*.

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
E. I don't know

Get your i-clickers

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 conditional P(X=0|Y=1)?

A. 2/6
B. 1/2
C. 1/6
D. 4/6
E. I don’t know

Get your i-clickers

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 $E_{i_1}, E_{i_2}, \ldots, E_{i_k}$,

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

Independence of Random Variables X and Y

- **Random variable independence** means that knowledge of any of the values of *X* does not change probabilities of any of the values of *Y*

- Opposite: **Dependence** implies that some values of *X* influence the probability of some 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 **<u>all events</u>** 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) $P(X=x, Y=y)=P(X=x)\cdot P(Y=y)$
  <u>for every pair</u> x **<u>and</u>** 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

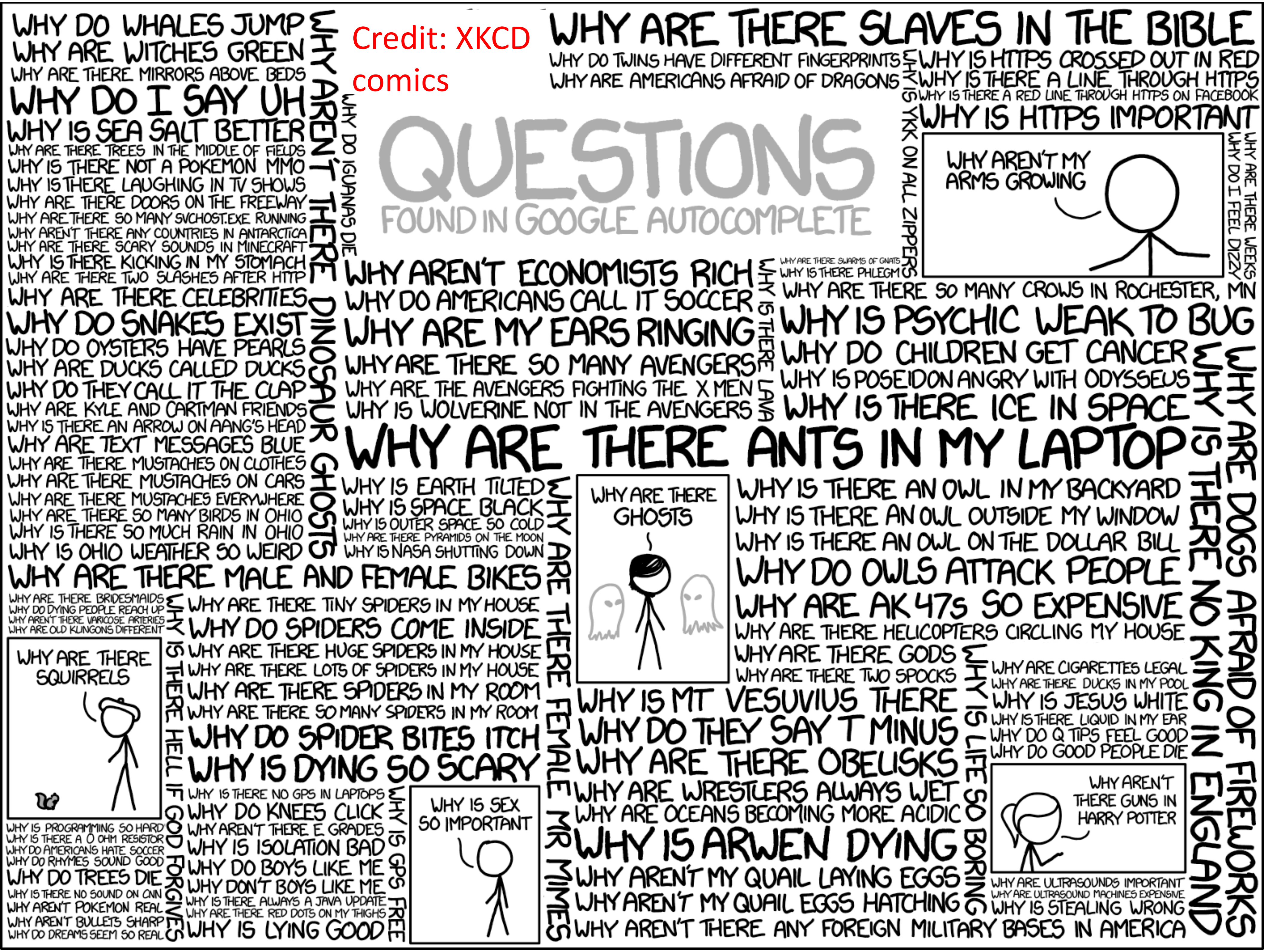
X and Y are Bernoulli variables

	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



Credit: XKCD comics

Joint Probability Density Function Defined

The joint probability density function for the continuous random variables X and Y, denotes as $f_{XY}(x,y)$, satisfies the following properties:

(1) $f_{XY}(x, y) \geq 0 \text{ for all } x, y$

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\,dx\,dy = 1$

(3) $P\big((X, Y) \subset R\big) = \iint_R f_{XY}(x, y)\,dx\,dy \quad (5\text{–}2)$

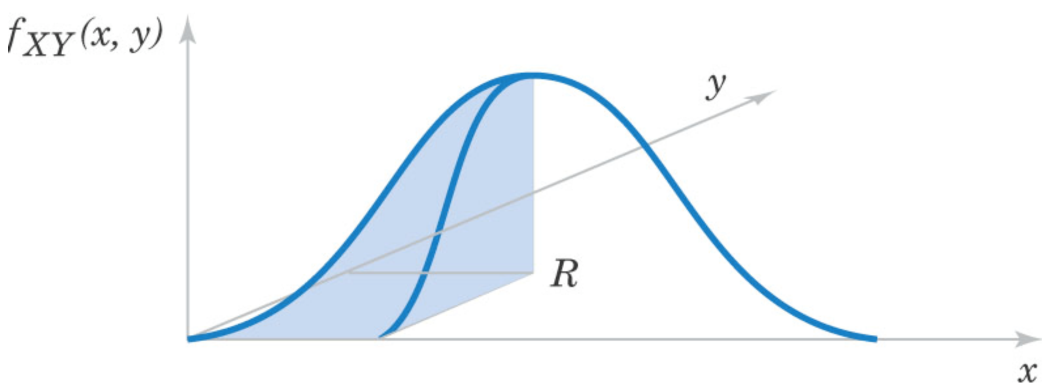


Figure 5-2 Joint probability density function for the random variables X and Y. Probability that (X, Y) is in the region R is determined by the volume of $f_{XY}(x,y)$ over the region R.

Joint Probability Density Function Graph

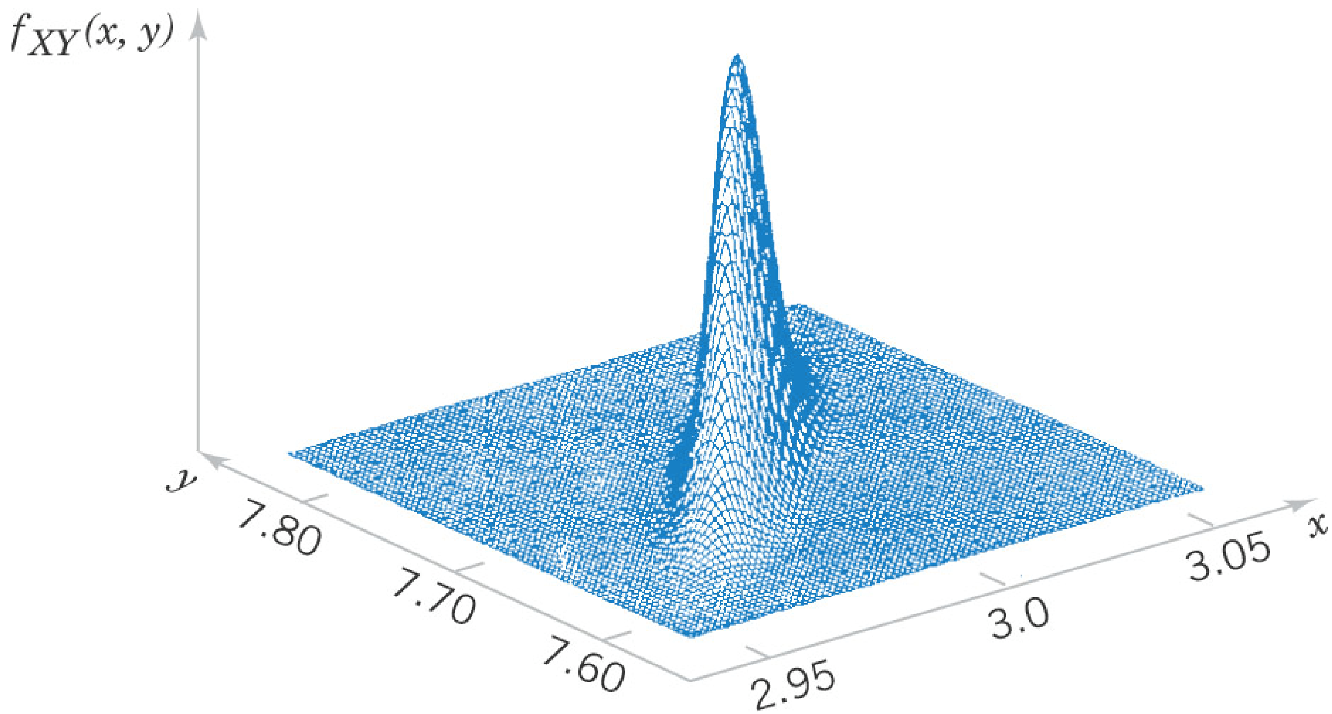


Figure 5-3 Joint probability density function for the continuous random variables *X* and *Y* of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the *X* dimension are more likely to occur when small values in the *Y* dimension occur.

Marginal Probability Distributions (continuous)

- Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
- The marginal PDFs are used to make probability statements about one variable.
- If the joint probability density function of random variables *X* and *Y* is $f_{XY}(x,y)$, the marginal probability density functions of *X* and *Y* are:

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

$$f_Y(y) = \int_x f_{XY}(x,y)\,dx \qquad (5\text{−}3)$$

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

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

Conditional Probability Density Function Defined

Given continuous random variables X and Y with
joint probability density function $f_{XY}(x, y)$,
the conditional probability densiy function of Y given X=x is

$$f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_y f_{XY}(x,y)\,dy} \text{ if } f_X(x) > 0 \qquad (5-4)$$

which satifies the following properties:

(1) $f_{Y|x}(y) \geq 0$

(2) $\int f_{Y|x}(y)dy = 1$

(3) $P(Y \subset B|X = x) = \int_B f_{Y|x}(y)dy$ for any set B in the range of Y

Compare to discrete: *$P(Y=y|X=x)=f_{XY}(x,y)/f_X(x)$*

Conditional Probability Distributions

- Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.
- Suppose *p* = 5 and we wish to find the distribution of X_1, X_2 and X_3 conditional on $X_4=x_4$ and $X_5=x_5$.

$$f_{X_1X_2X_3|x_4x_5}(x_1,x_2,x_3)=\frac{f_{X_1X_2X_3X_4X_5}(x_1,x_2,x_3,x_4,x_5)}{f_{X_4X_5}(x_4,x_5)}$$

for $f_{X_4X_5}(x_4,x_5)>0$.

Independence for Continuous Random Variables

For random variables *X* and *Y*, if any one of the following properties is true, the others are also true. Then *X* and *Y* are independent.

(1) $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$
(2) $f_{Y|x}(y) = f_Y(y)$ for all x and y with $f_X(x) > 0$
(3) $f_{X|y}(y) = f_X(x)$ for all x and y with $f_Y(y) > 0$
(4) $\mathrm{P}(X \subset A, Y \subset B) = P(X \subset A) \cdot P(Y \subset B)$ for any sets A and B in the range of X and Y, respectively. (5−7)

P(Y=y|X=x)=P(Y=y) **for any *x*** or
P(X=x|Y=y)=P(X=x) **for any *y*** or
P(X=x, Y=y)=P(X=x)·P(Y=y) **for any *x* and *y***

Covariation, Correlations

Quick and dirty check for
linear (in)dependence
between variables

Covariance – 1 number to measure dependance between random variables

$Cov(X,Y)$ or σ_{XY}

$$\sigma_{XY} = E\left[(X-\mu_X)\cdot(Y-\mu_Y)\right] =$$

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

- $Var(X) = Cov(X,X)$
- If X & Y are independent $Cov(X,Y) = E[X-\mu_X]\cdot E[Y-\mu_Y] = 0$
- $-\infty < Cov(X,Y) < +\infty$ Can be negative!

Covariance Defined

Covariance is a number quantifying the average ***linear*** dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\text{cov}(X,Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

Montgomery, Runger 5th edition Eq. (5−14)

The units of σ_{XY} are the units of X times the units of Y.

Unlike the range of the variance, covariance can be negative: $-\infty < \sigma_{XY} < \infty$.

Covariance and PMF tables

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

The probability distribution of Example 5-1 is shown.

By inspection, note that the larger probabilities occur as *X* and *Y* move in opposite directions. This indicates a negative covariance.

Covariance and Scatter Patterns

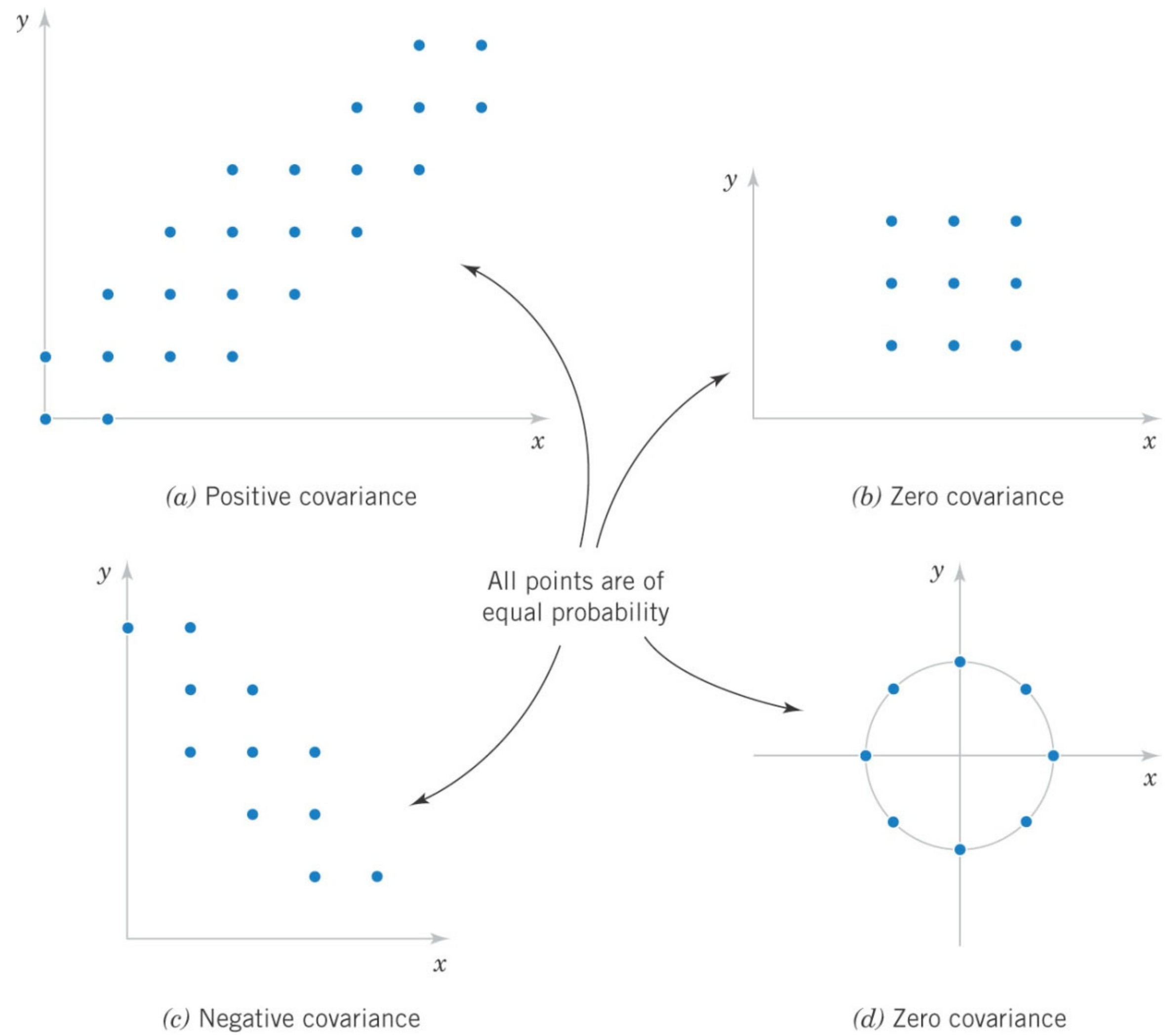


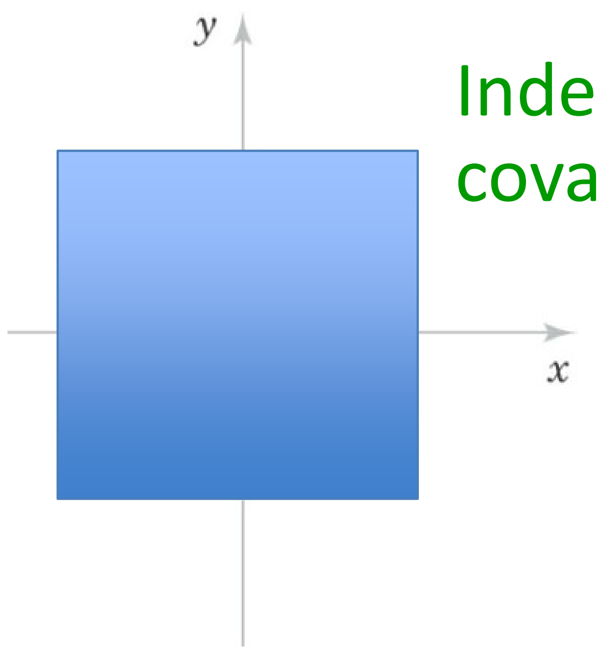
Figure 5-13 Joint probability distributions and the sign of cov(X, Y). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

Independence Implies σ=ρ = 0 but <u>not vice versa</u>

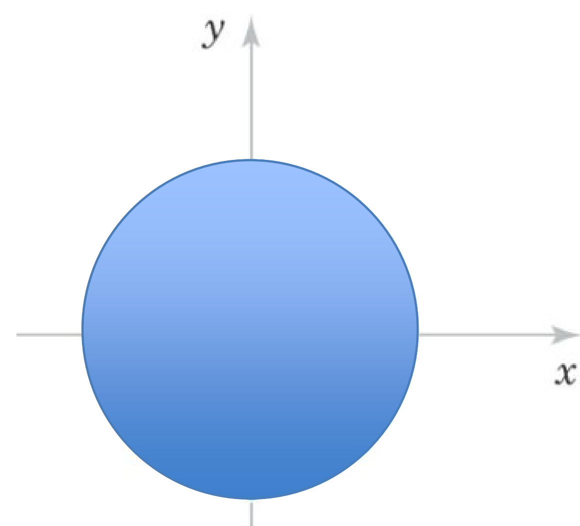
- If *X* and *Y* are independent random variables,

$$\sigma_{XY} = \rho_{XY} = 0 \qquad (5\text{-}17)$$

- $\rho_{XY} = 0$ is necessary, but not a sufficient condition for independence.



Independent
covariance=0



NOT independent
covariance=0

Correlation is “normalized covariance”

- Also called:
  Pearson correlation coefficient

  $\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
  is the covariance normalized to be $-1 \leq \rho_{XY} \leq 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Prove that ρ_{xy} is in $[-1, 1]$

$$Z_x = \frac{X-\mu_x}{\sigma_x}\ ; \quad Z_y = \frac{Y-\mu_y}{\sigma_y}$$

$$0 \le E\left((Z_x - Z_y)^2\right) = E(Z_x^2) + E(Z_y^2) -$$

$$-2E(Z_x \cdot Z_y) = 2 - 2\frac{1}{\sigma_x \sigma_y} E\left((X-\mu_x)(Y-\mu_y)\right) =$$

$$2 - 2\rho_{xy} \Longrightarrow \boxed{\rho_{xy} \le 1}$$

$$0 \le E\left((Z_x + Z_y)^2\right) = E(Z_x^2) + E(Z_y^2) +$$

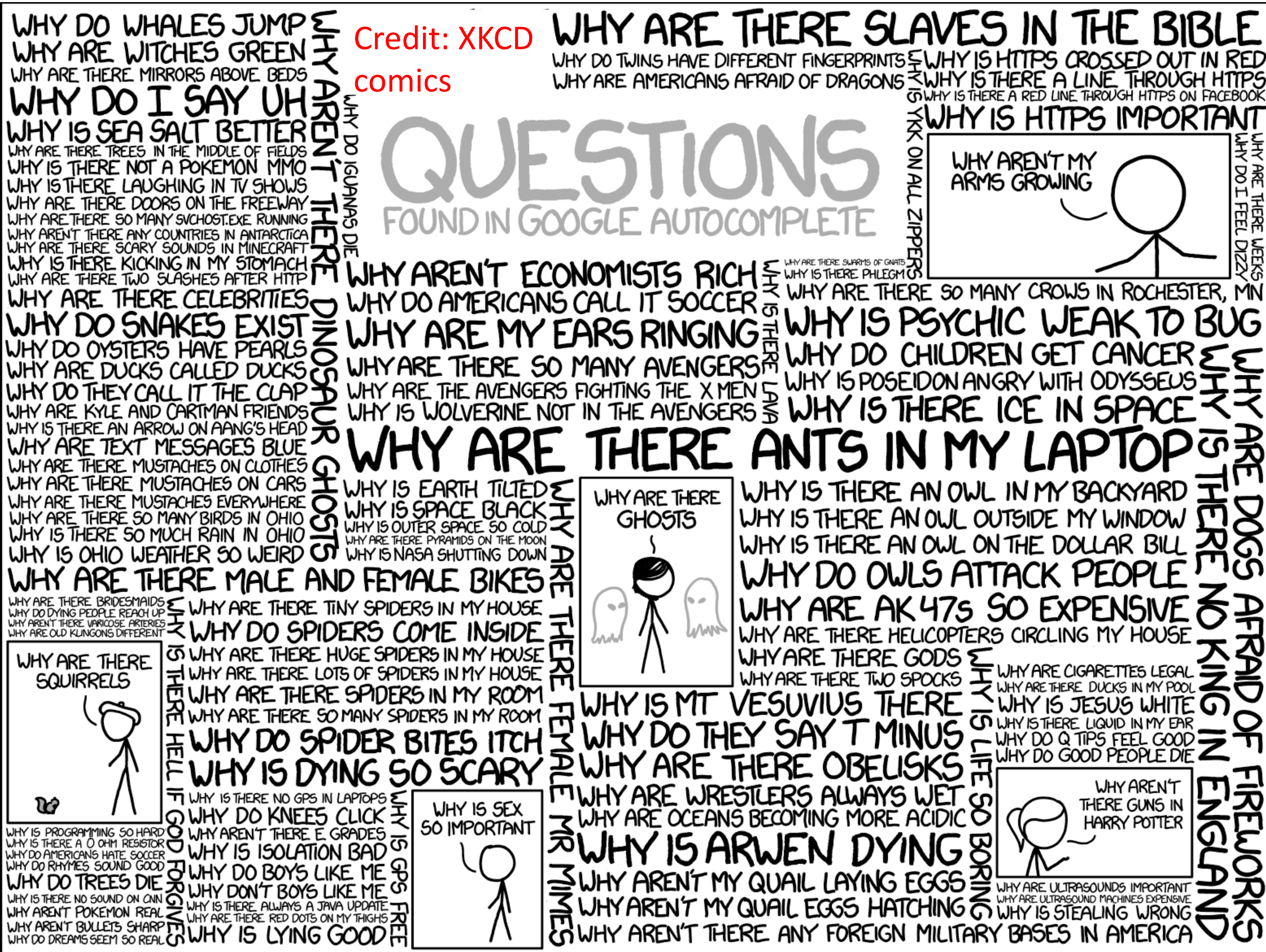
$$+2E(Z_x \cdot Z_y) = 2 + 2\rho_{xy} \Longrightarrow$$

$$\Longrightarrow \boxed{\rho_{xy} \ge -1}$$

Spearman rank correlation

- Pearson correlation tests for linear relationship between X and Y
- Unlikely for variables with broad distributions → non-linear effects dominate
- Spearman correlation tests for any monotonic relationship between X and Y
- Calculate ranks (1 to n), $r_X(i)$ and $r_Y(i)$ of variables in both samples. Calculate Pearson correlation between ranks: Spearman(X,Y) = Pearson(r_X, r_Y)
- Ties: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.
- If lots of ties: use Kendall rank correlation (Kendall tau)

Credit: XKCD comics

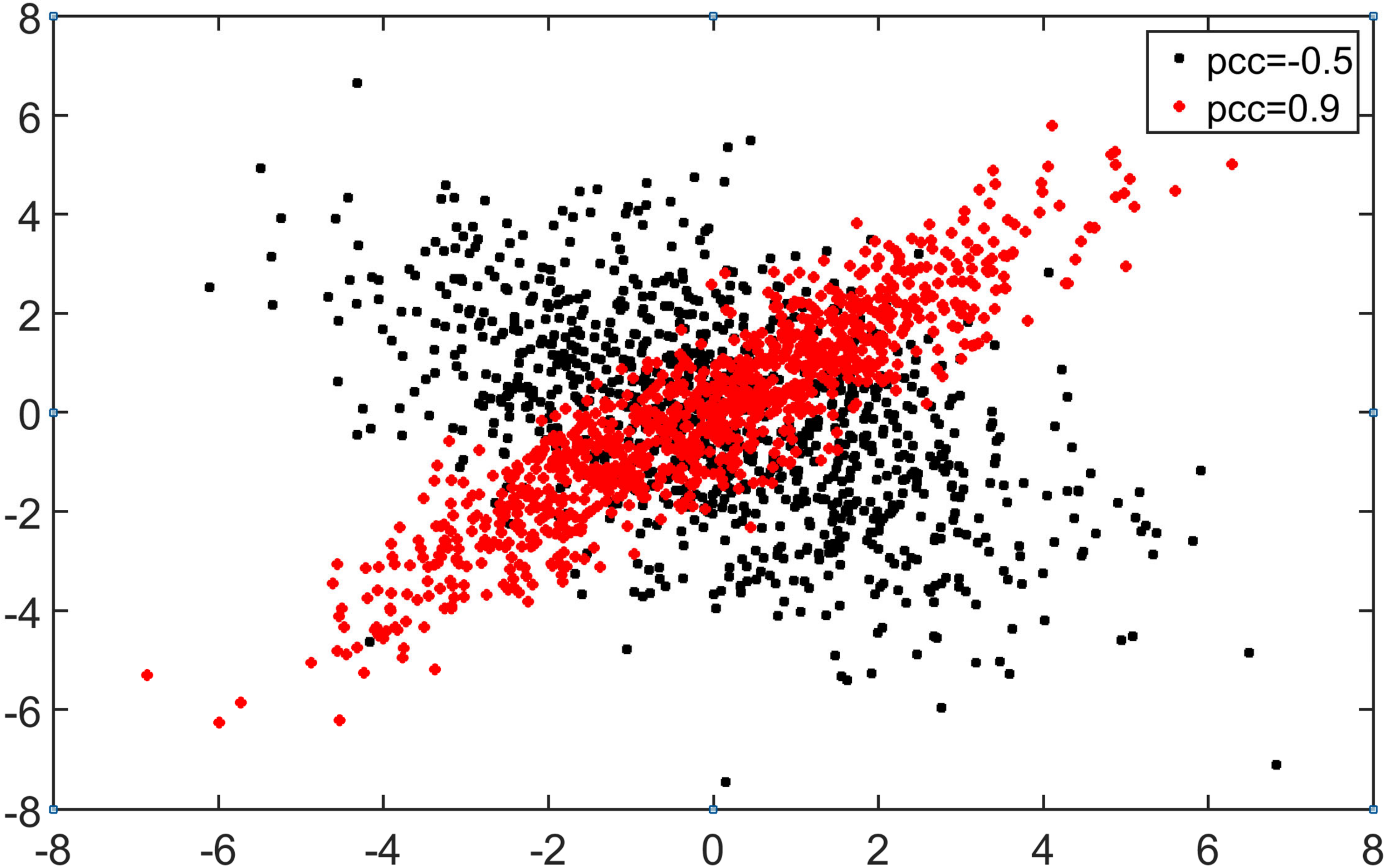


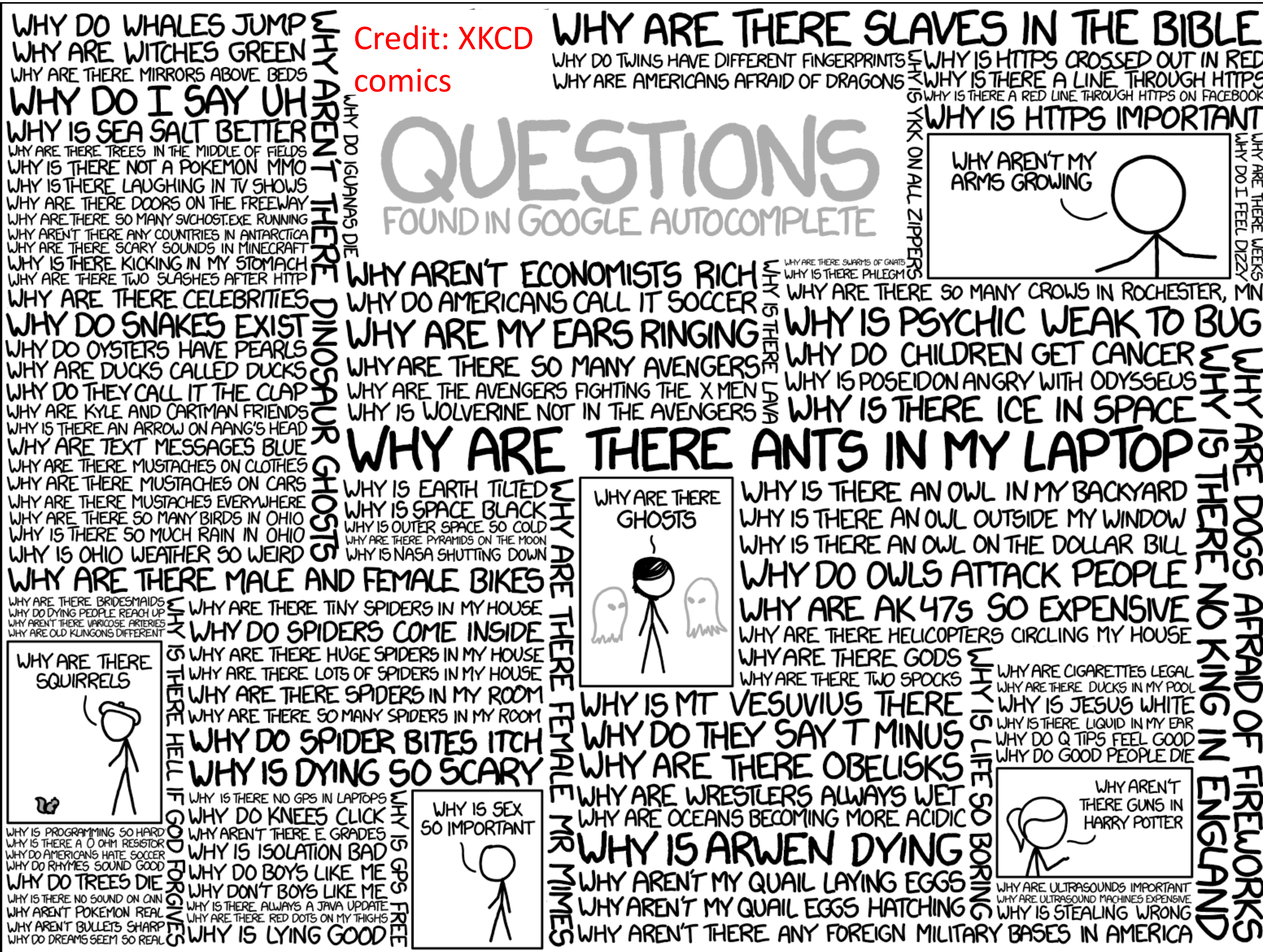
Matlab exercise: Correlation/Covariation

- Generate a sample with Stats=100,000 of two Gaussian random variables r1 and r2 which have mean 0 and standard deviation 2 and are:
  - Uncorrelated
  - Correlated with correlation coefficient 0.9
  - Correlated with correlation coefficient -0.5
  - Trick: first make uncorrelated r1 and r2. Then make anew variable: r1mix=mix.*r2+(1-mix.^2)^0.5.*r1; where mix= corr. coeff.
- For each value of mix calculate covariance and correlation coefficient between r1mix and r2
- In each case make a scatter plot: plot(r1mix,r2,'k.');

Matlab exercise: Correlation/Covariation

1. Stats=100000;
2. r1=2.*randn(Stats,1);
3. r2=2.*randn(Stats,1);
4. disp('Covariance matrix='); disp(cov(r1,r2));
5. disp('Correlation=');disp(corr(r1,r2));
6. figure; plot(r1,r2,'k.');
7. mix=0.9; %Mixes r2 to r1 but keeps same variance
8. r1mix=mix.*r2+sqrt(1-mix.^2).*r1;
9. disp('Covariance matrix='); disp(cov(r1mix,r2));
10. disp('Correlation=');disp(corr(r1mix,r2));
11. figure; plot(r1mix,r2,'k.');
12. mix=-0.5; %REDO LINES 8-11

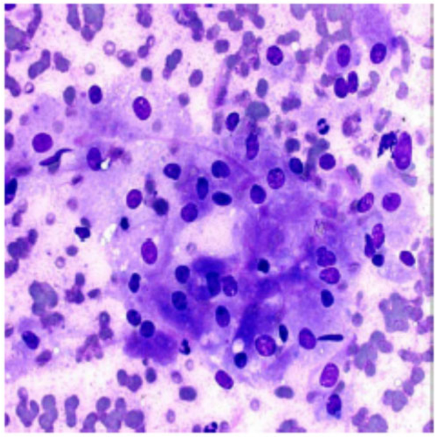

pcc=-0.5
pcc=0.9
8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8



Credit: XKCD comics

Let’s work with real cancer data!

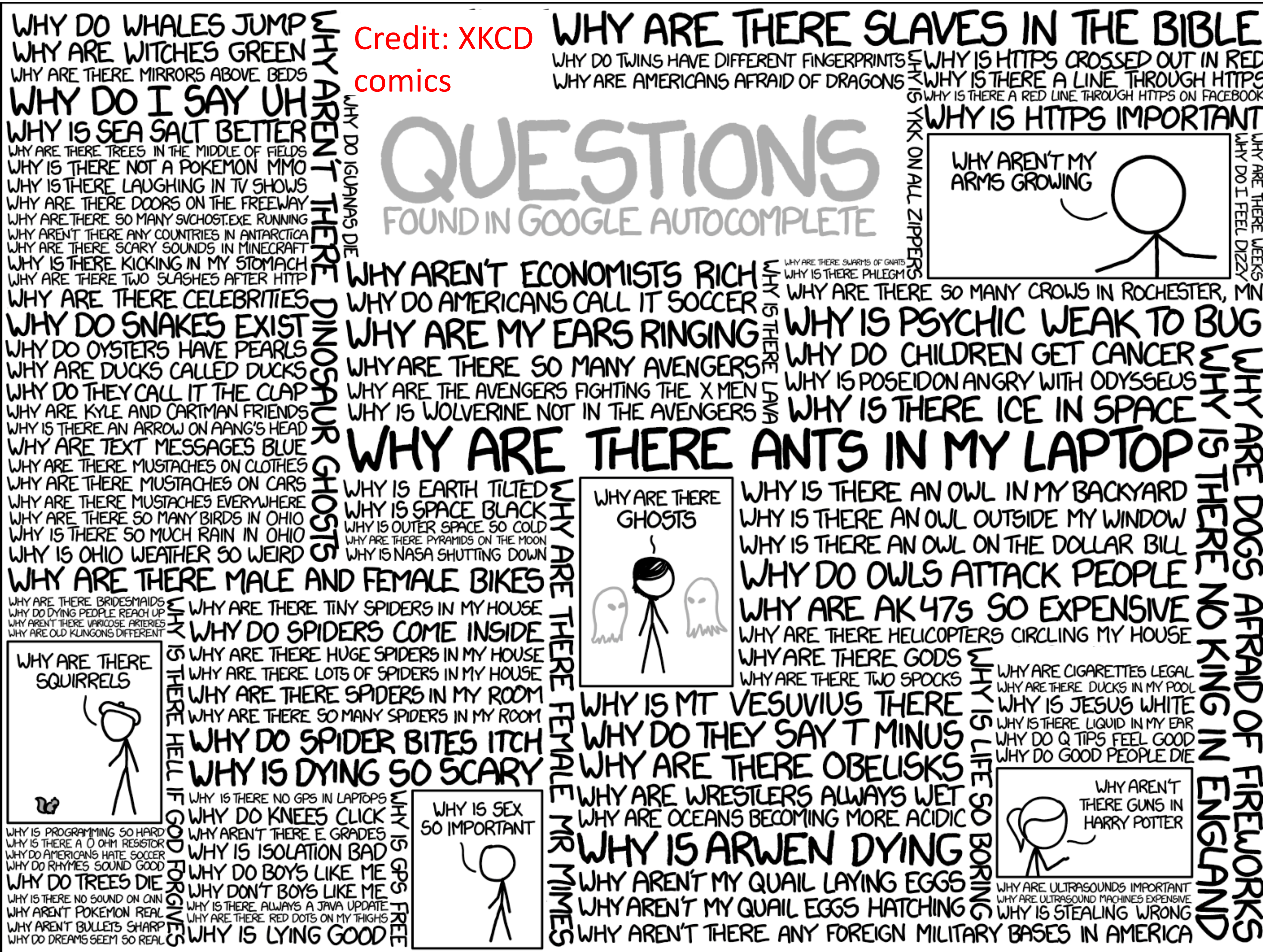
- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer
- Statistics of all cells in the image
- 212 cancer patients and 357 healthy individuals (column 1)
- 30 other properties (see table)



Variable	Mean	S.Error	Extreme
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness (perimeter^2 / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension (“coastline approximation” - 1)	Col 11	Col 21	Col 31

Matlab exercise #2

- Download cancer data in cancer_wdbc.mat
- Data in the file cancerwdbc.mat (569x30). First 357 patients are healthy. The remaining 569-357=212 patients have cancer.
- Make scatter plots of radius vs perimeter and texture vs radius.
- Calculate Pearson and Spearman correlations in both cases
- Calculate the correlation matrix of all-against-all variables: there are 30*29/2=435 correlations.
  Hint: corr_mat=corr(cancerwdbc);
- Plot the histogram of these 435 correlation coefficients.
  Hint: use [i,j,v]=find(corr_mat); then find all i>j and analyze v evaluated on this subset of 435 matrix elements



Credit: XKCD comics