

Joint Probability Density Function Defined

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

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

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

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

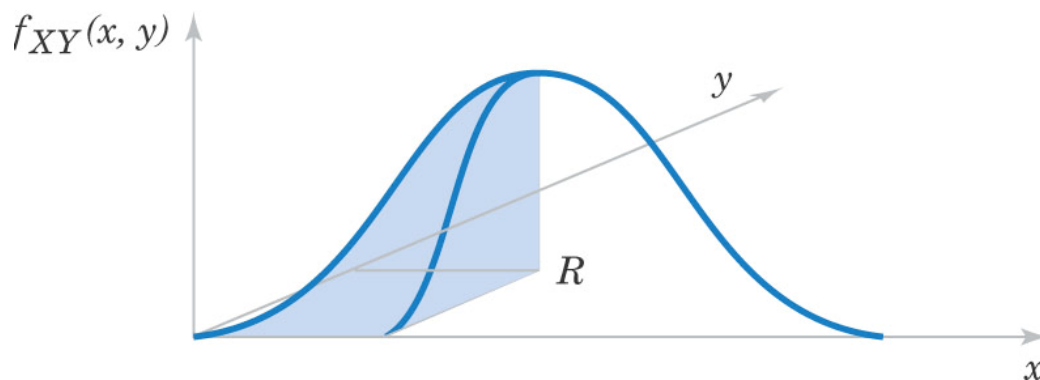


Figure 5-2 Joint probability density function for the random variables X and Y . Probability that (X, Y) is in the region R is determined by the **volume** of $f_{XY}(x,y)$ over the region R .

Joint Probability Density Function Graph

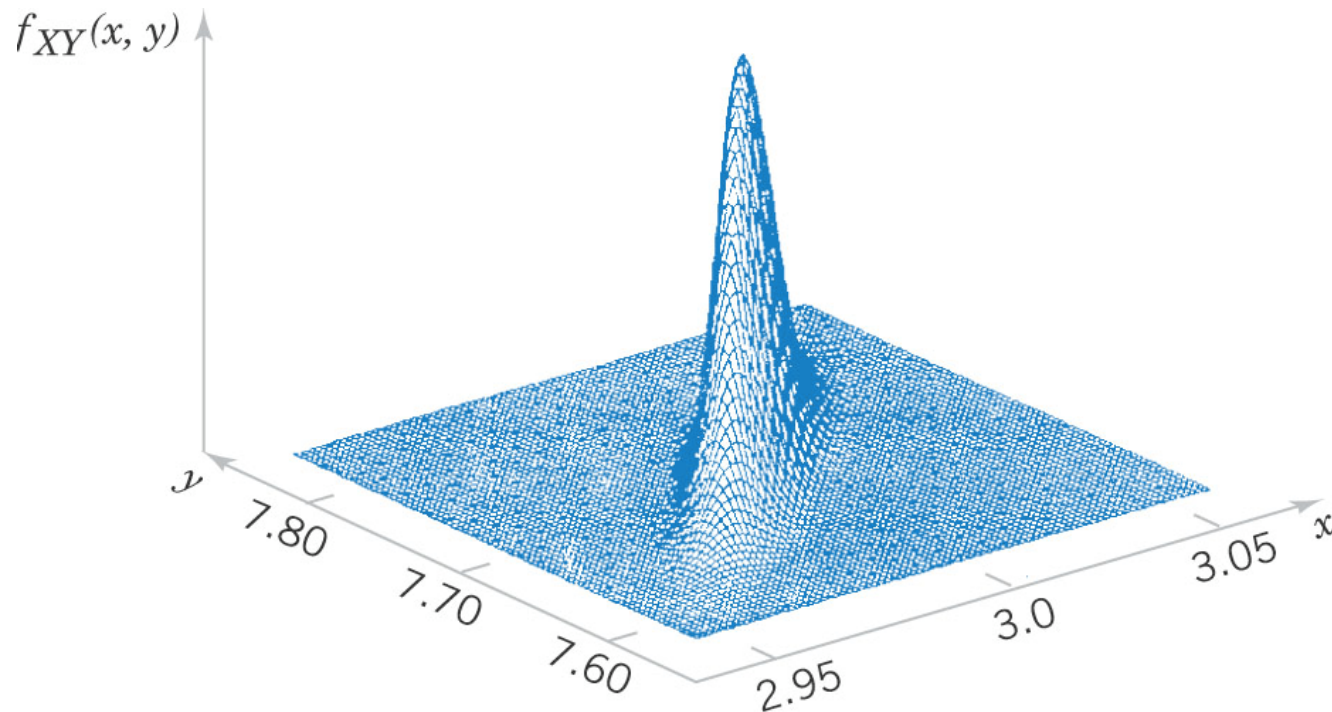


Figure 5-3 Joint probability density function for the continuous random variables X and Y of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the X dimension are more likely to occur when small values in the Y dimension occur.

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:

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

$$f_Y(y) = \int_x f_{XY}(x, y) dx \quad (5-3)$$

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

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

Conditional Probability Density Function Defined

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

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

which satisfies the following properties:

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

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

(3) $P(Y \in B | X = x) = \int_B f_{Y|x}(y) dy$ 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**.

$$(1) f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

$$(2) f_{Y|x}(y) = f_Y(y) \text{ for all } x \text{ and } y \text{ with } f_X(x) > 0$$

$$(3) f_{X|y}(x) = f_X(x) \text{ for all } x \text{ and } y \text{ with } f_Y(y) > 0$$

$$(4) P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B) \text{ for any sets } A \text{ and } B \text{ in the range of } X \text{ and } Y, \text{ respectively.} \quad (5-7)$$

$$P(Y=y|X=x)=P(Y=y) \text{ for any } x \text{ or}$$

$$P(X=x|Y=y)=P(X=x) \text{ for any } y \text{ or}$$

$$P(X=x, Y=y)=P(X=x) \cdot P(Y=y) \text{ for any } x \text{ and } y$$

Covariation, Correlations

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

Covariance - 1 number to
measure dependance
between random variables

$\text{Cov}(X, Y)$ or σ_{xy}

$$\sigma_{xy} = E[(X - \mu_x) \cdot (Y - \mu_y)] = \\ = E(X \cdot Y) - \mu_x \cdot \mu_y$$

- $\text{Var}(X) = \text{Cov}(X, X)$
- If X & Y are independent

$$\text{Cov}(X, Y) = E[X - \mu_x] \cdot E[Y - \mu_y] = 0$$

- $-\infty$ < $\text{Cov}(X, Y)$ < $+\infty$ 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

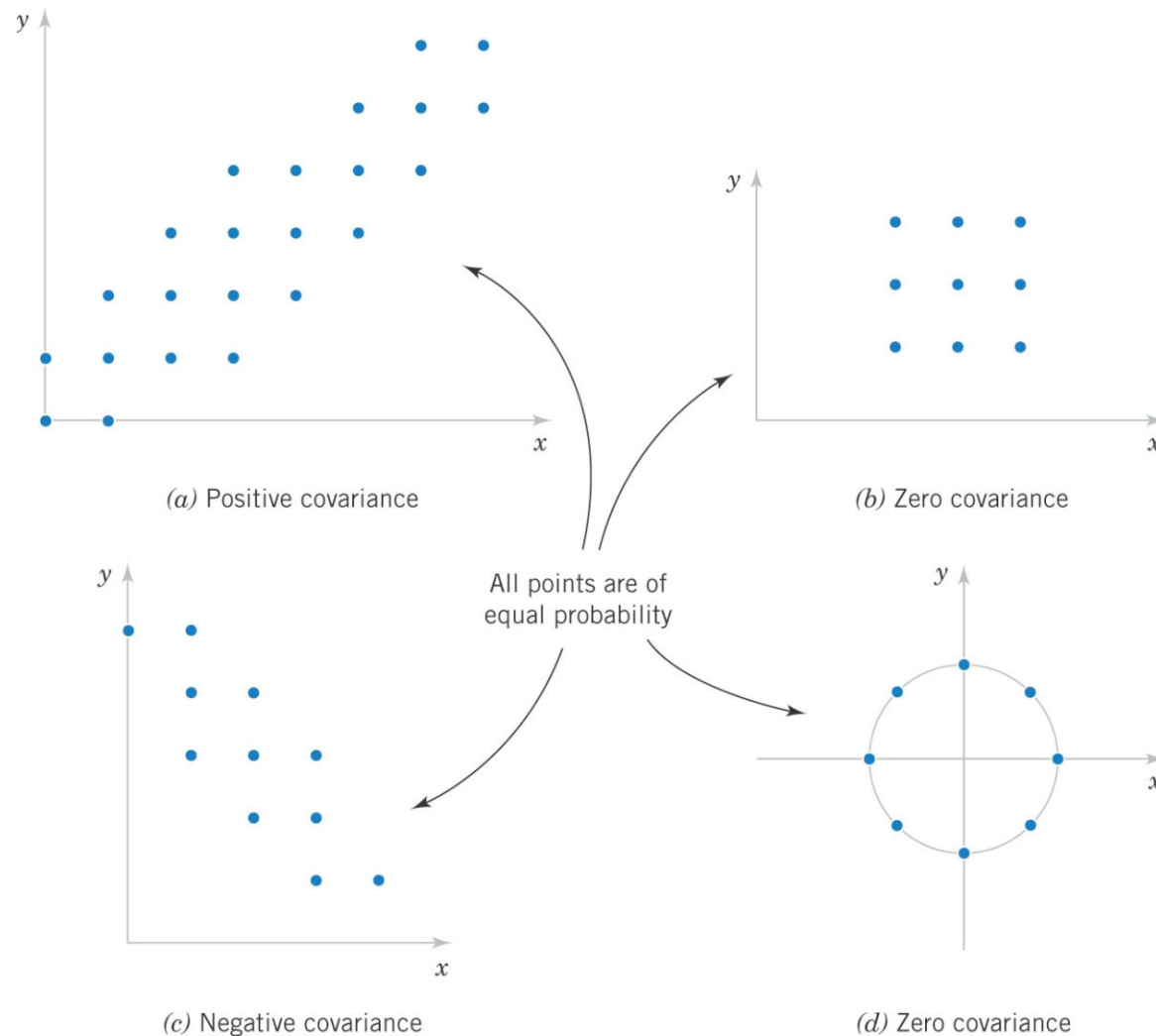
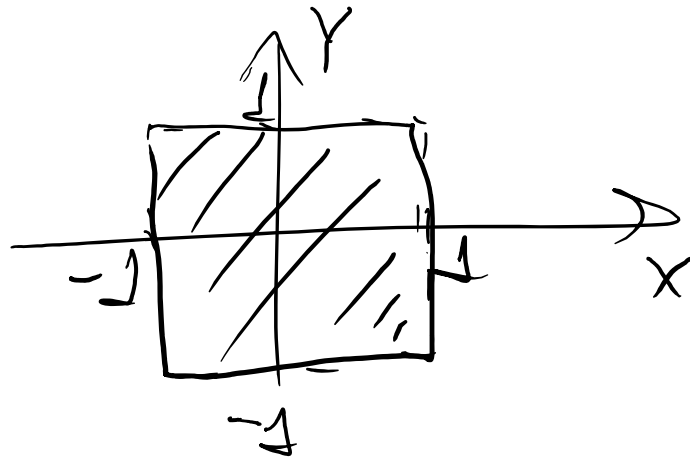


Figure 5-13 Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are **correlated**.

Example 1:

Uniform distribution in the square

$$-1 < X < 1, -1 < Y < 1$$



$$\begin{cases} f_{XY}(x, y) = c & \text{if } -1 < x < 1 \text{ and } -1 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

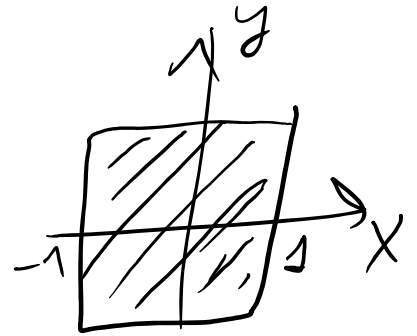
$$1 = \int_{\text{square}} dx dy f_{XY}(x, y) = c \cdot \text{Area} = c \cdot 4 \rightarrow c = \frac{1}{4}$$

Are X and Y independent? Yes they are

Let's test if $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

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

$$= \int_{-1}^1 \frac{1}{4} dy = \frac{1}{2} \text{ if } -1 < x < 1$$

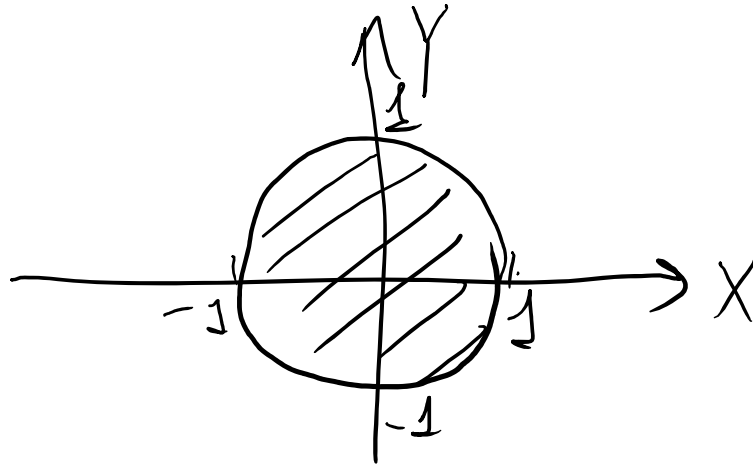


Same for $f_Y(y) = \frac{1}{2}$ if $-1 < y < 1$

$$\frac{1}{4} = f_{XY}(x, y) = \frac{1}{2} \cdot \frac{1}{2} = f_X(x) \cdot f_Y(y)$$

0 otherwise if both x & y are in $[-1, 1]$

X and Y are uniformly distributed in
the disc $x^2+y^2 \leq 1$



Are they independent?

A. yes

☒ B. no

C. I could not figure it out

Get your i-clickers

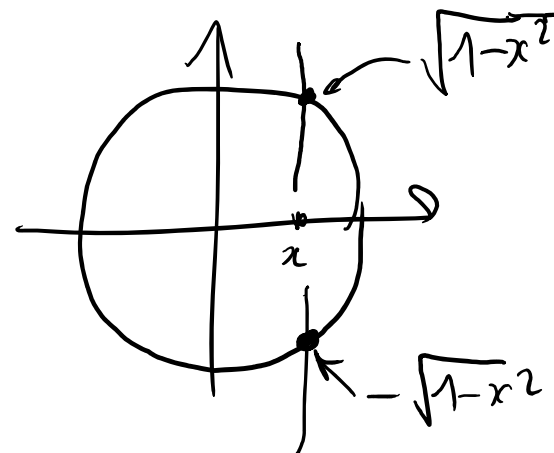
Joint PDF $f_{XY}(x, y) = \frac{1}{\text{area}} = \frac{1}{\pi}$ if x, y in the disc

0 - otherwise

Marginal distributions:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{dy}{\pi} = \frac{2\sqrt{1-x^2}}{\pi}$$

Same for $f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi}$



$$\frac{1}{\pi} = f_{XY}(x, y) \neq \frac{2}{\pi} \sqrt{1-x^2} \cdot \frac{2}{\pi} \sqrt{1-y^2} = f_X(x) \cdot f_Y(y)$$

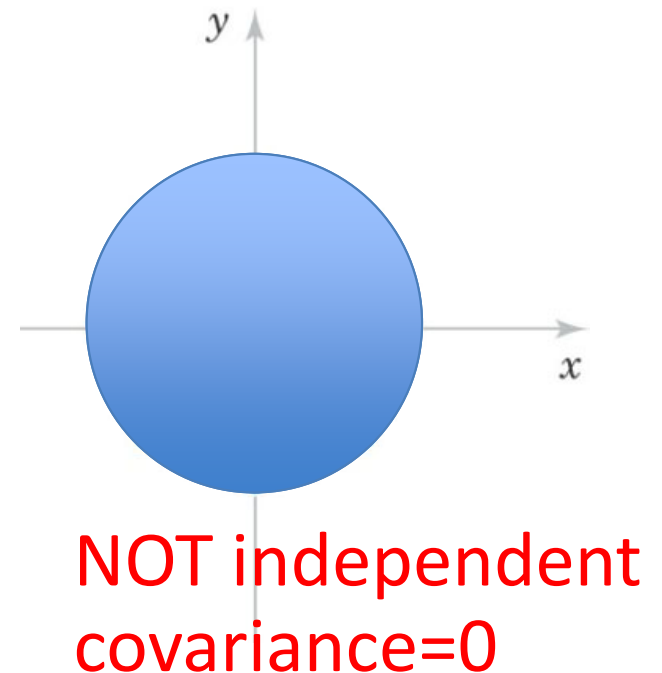
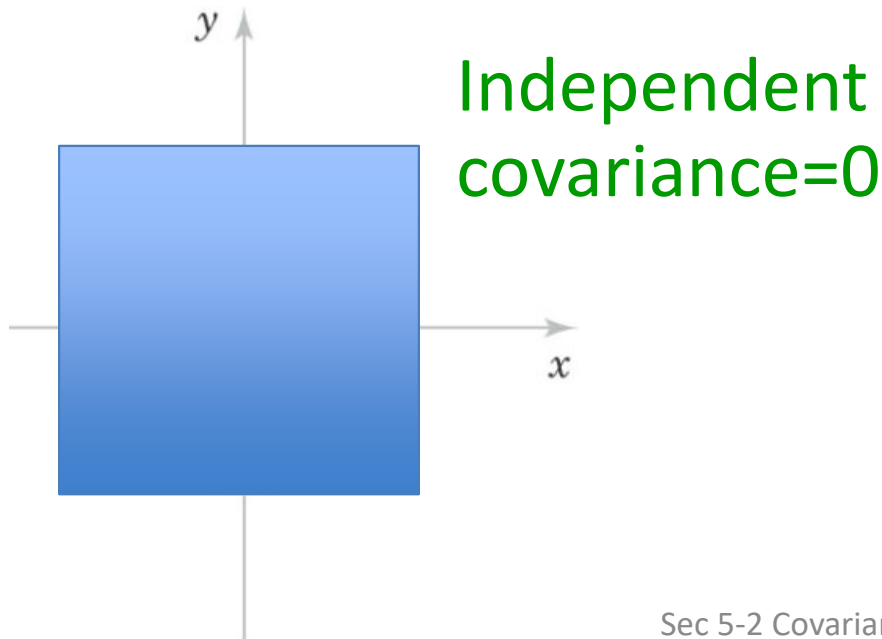
Variables are NOT independent

Independence Implies $\sigma=\rho = 0$ but not vice versa

- If X and Y are independent random variables,

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

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



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 $[-1, 1]$

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

$$\begin{aligned} 0 \leq E((Z_x - Z_y)^2) &= E(Z_x^2) + E(Z_y^2) - \\ &- 2E(Z_x \cdot Z_y) = 2 - 2 \frac{1}{\sigma_x \sigma_y} E((X - \mu_x)(Y - \mu_y)) = \end{aligned}$$

$$2 - 2\rho_{xy} \implies \boxed{\rho_{xy} \leq 1}$$

$$\begin{aligned} 0 \leq E((Z_x + Z_y)^2) &= E(Z_x^2) + E(Z_y^2) + \\ &+ 2E(Z_x \cdot Z_y) = 2 + 2\rho_{xy} \implies \end{aligned}$$

$$\implies \boxed{\rho_{xy} \geq -1}$$

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:
$$\text{Spearman}(X,Y) = \text{Pearson}(r_X, r_Y)$$
- **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)

IMPORTANT

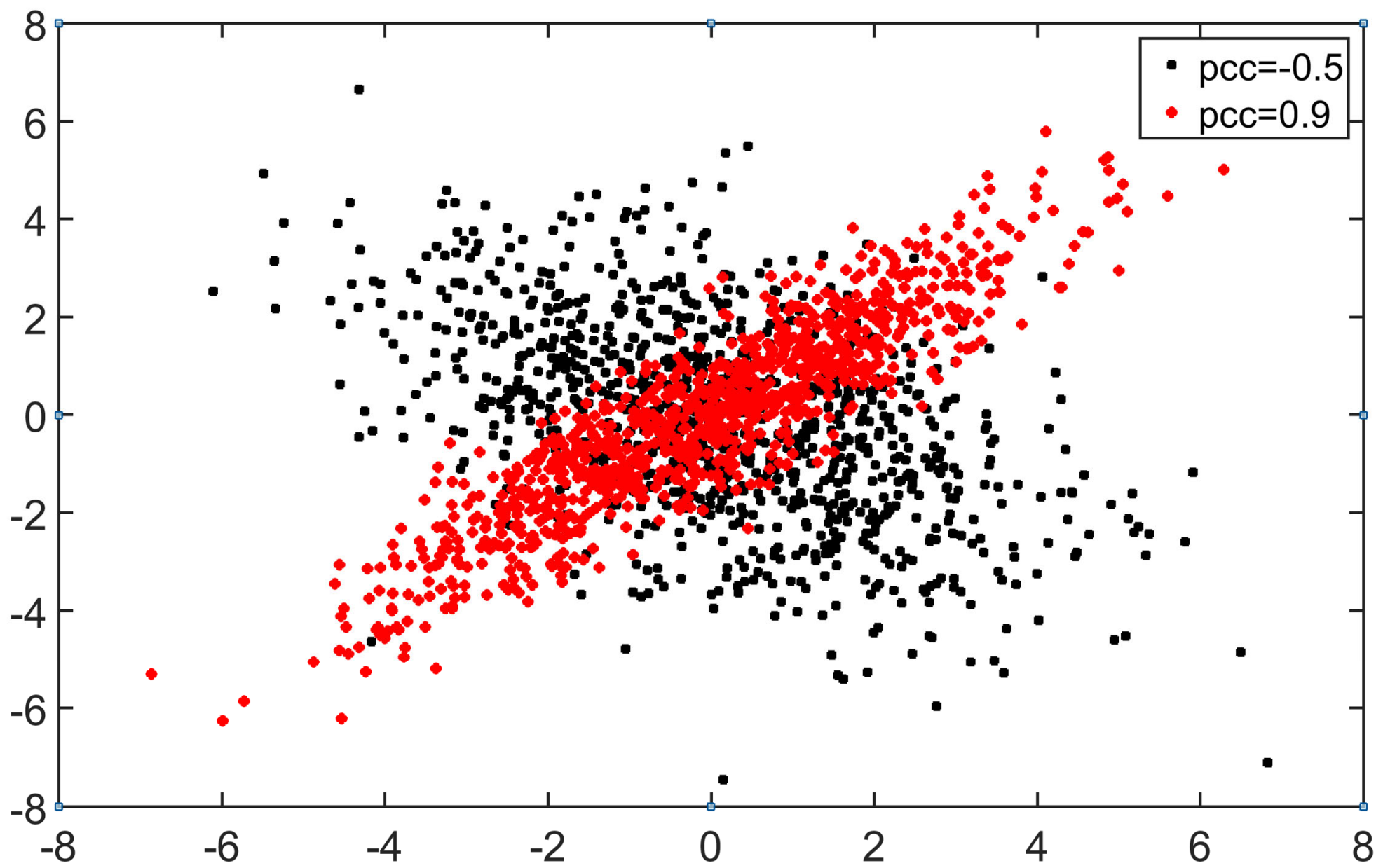


Matlab exercise: Correlation/Covariation

- 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 a new variable: **$r1_{mix} = mix \cdot r2 + (1 - mix^2)^{0.5} \cdot r1$** ; 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: **$plot(r1_{mix}, r2, 'k')$** ;

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



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

WHY AREN'T MY
ARMS GROWING



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY AREN'T ECONOMISTS RICH

WHY DO AMERICANS CALL IT SOCCER

WHY ARE MY EARS RINGING

WHY ARE THERE SO MANY AVENGERS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY IS SPACE BLACK

WHY IS OUTER SPACE SO COLD

WHY ARE THERE PYRAMIDS ON THE MOON

WHY IS NASA SHUTTING DOWN

WHY ARE THERE MALE AND FEMALE BIKES

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 SEX
SO IMPORTANT



WHY ARE THERE
GHOSTS



WHY IS THERE AN OWL IN MY BACKYARD

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY DO OWLS ATTACK PEOPLE

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE GODS

WHY ARE THERE TWO SPOCKS

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 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 AREN'T
THERE GUNS IN
HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

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

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

WHY ARE THERE
SQUIRRELS



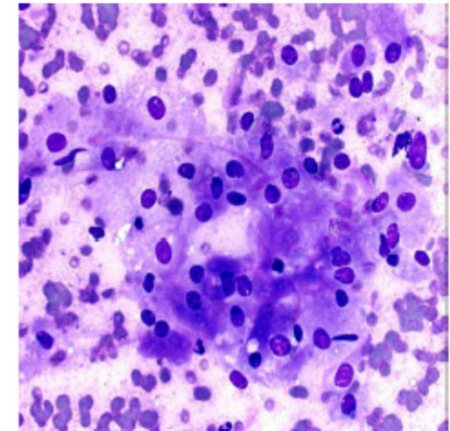
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE

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

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 ($\text{perimeter}^2 / \text{area} - 1.0$)	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 $30 \times 29 / 2 = 435$ 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

