Geospatial Data Science
Content Block II: *Techniques*
Lecture 6
Probability Theory, Spatial Corr.

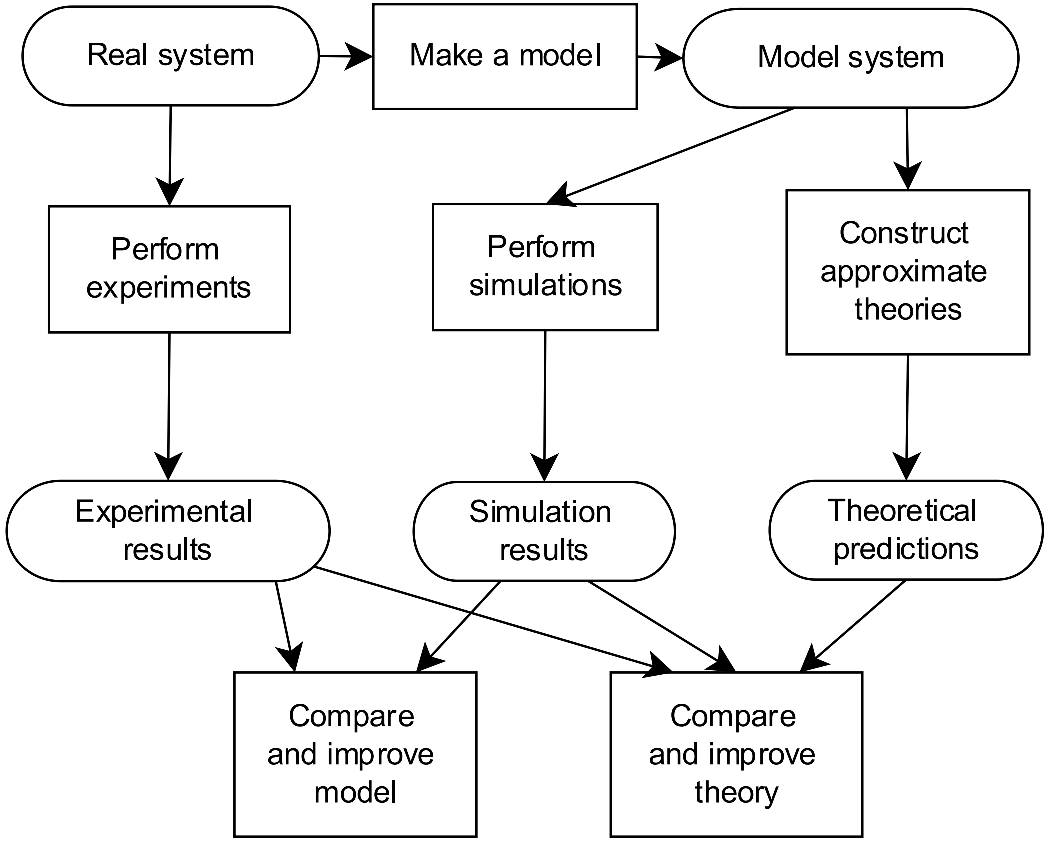
Austin J. Brockmeier, Ph.D.

Monday, March 13th, 2023

Probability and Spatial Statistics for Geospatial Systems and Data

1. Statistical analysis to understand the relationship of attributes in and across time and space (Lecture 6)
2. Generate data from random variables and processes through **computer simulation** (Lab 6&7)
3. Analyze spatial patterns in data (Lab 6&7)
4. Create **models** of data-generating processes (Lecture 7)

- **Lab 6**: generate & analyze spatial random variables
- **Lab 7**: spatial interpolation



Outline

Probability theory and spatial correlation

- Playing with chance: random experiments
- Probability mass function
- Population density
- Statistics: Bias, variance, and error
- Dependence and correlation
- Joint, marginal, and conditional distributions
- Bayes rule

Poll Everywhere

- Go to the website: PollEv.com/ajbrock

Example Motivating Questions:

Abstract:

What is the most likely attribute category in a particular area?

What’s the average and variation of attribute values in a particular area?

Do pairs of relatively nearby points have similarly valued features?

How does the value of one attribute co-relate with the category of another attribute?

Concrete:

- What is the most common crop raised in each section of land?
- What is the average rainfall in the month of July in Newark? How much does the rainfall vary per year?
- How related is the rainfall in Newark, Delaware to the rainfall at the beach in Lewes, Delaware?
- How dependent is the choice of crop on the rainfall across the US?

It all starts with counting

How many objects were observed in a particular time and space?

Examples of counting experiments without attributes

Number of northbound vehicles on I-95 between 7:00–8:00 am on 3/1/2023.

Number of pedestrians crossing on N. College at Delaware Ave between 7:00–8:00 am on 3/1/2023.

Number of popcorns popping after 2 minutes on the stove.

It all starts with counting

Examples of counting experiments without attributes

Number of northbound vehicles on I-95 between 7:00–8:00 am on 3/1/2023.

Number of pedestrians crossing on N. College at Delaware Ave between 7:00–8:00 am on 3/1/2023.

Number of popcorns popping after 2 minutes on the stove.

Assuming the objects were recorded with attributes, what proportion fit a particular description?

Examples of attributes

(type [categorical], mass in kg [continuous], state of license [categorical]) ∈ {car, bus, semi,…}× $\mathbb{R}_{>0}$ ×{AL, AK,..}

(height in m [continuous], back-pack [bool], hat [bool], mask [bool]) ∈ $\mathbb{R}_{>0} \times \{0,1\} \times \{0,1\} \times \{0,1\}$

(elapsed time in seconds [continuous]) ∈ $\mathbb{R}_{>0}$

Examples of proportions

65% of the vehicles were DE-plated cars

20% of the pedestrians were wearing a hat and had a height greater than 1.60 m

5% of the popcorns popped after 15 s

Probability theory: Sample space, outcomes, and events

Sample space: A **set** of possible outcomes for each observation. This set defines the object attributes—what we expect to see and what we will keep track of.

Each **outcome**, in the sample space, is a unique description of a possible object.

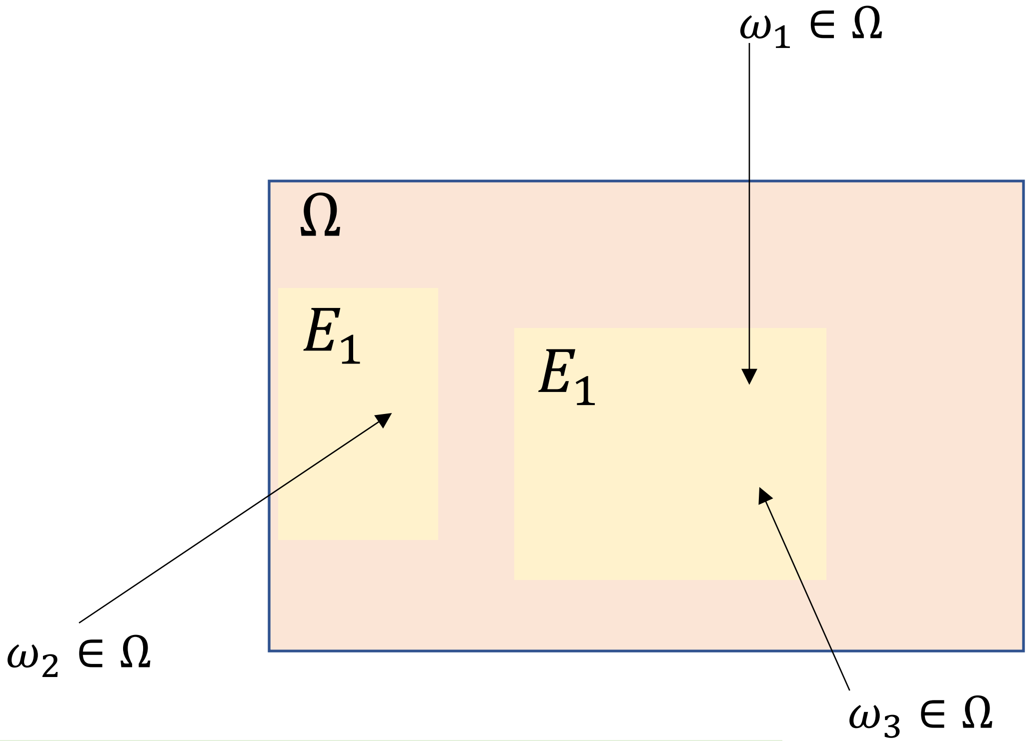
Only one **outcome** per trial.

Event space: A set of all relevant sets of outcomes.

Each **event**, in the event space, <u>is a set of outcomes.</u>

Ω

$\omega \in \Omega$



$E = \{\omega \in \Omega : \text{condition on } \omega\}$

Probability theory: Sample space, outcomes, and events

Sample space: A **set** of possible outcomes for each observation. This set defines the object attributes—what we expect to see and what we will keep track of.

Each **outcome**, in the sample space, is a unique description of a possible object.

Only one **outcome** per trial.

Event space: A set of all relevant sets of outcomes.

Each **event**, in the event space, <u>is a set of outcomes.</u>

<u>Example 1</u>
Consider a particular census tract.

An outcome is the completed census form.

The event (subset of interest) is households with at least one child under 5.

<u>Example 2</u>
Draw a card 🂦 from a shuffled standard deck of 52.
The sample space is the set of cards.

Each outcome corresponds to a particular card drawn.

Let the event of interest be that the card was from the spade suit ♠. This is a set consisting of the outcomes for all 13 cards that are spades.

Q1. The sample space is $\{1,2,3\}$

Which of these is a possible event?

A. 3

B. $\{6\}$

C. $\{\emptyset, 3\}$

D. $\{1,2\}$

E. $\emptyset$

Q2. The sample space is $\mathbb{R}^2$

Which of these is NOT a possible event?

A. $[0,1]^2$

B. $\{1,2\} \times \{3,4\}$

C. $\{1,2\}$

D. $[0,1] \times \mathbb{R}$

E. $\mathbb{R} \times \{3,4\}$

Probability theory: Probability measures

- A **probability measure** assigns a non-negative number [0,1] to each event
- The probability of a union of disjoint events is equal to the sum of the individual probabilities
- If the union of events is the entire set of outcomes then the probability is 1=100%

Given $E_1, \dots, E_N$,
if $E_i \cap E_j = \emptyset$, for $i \neq j$
then $\Pr\left(\bigcup_{i=1}^{N} E_i\right) = \sum_{i=1}^{N} \Pr(E_i)$

if $\bigcup_{i=1}^{N} E_i = \Omega$
$\Pr\left(\bigcup_{i=1}^{N} E_i\right) = 1$

N_O: Total count of objects.
N_E: Count of objects in a particular subset (particular categorical attributes or ranges of values).

The ratio N_E/N_O is a proportion, representing the probability/chance of selecting an object within the particular subset if objects are drawn at random.

Example 1
Consider drawing a household at random from particular census tract. What is the chance that it has children under 5 years of age? Assume there are $N_O = 200$ total households, and $N_E = 30$ households with at least one child under 5. Then this corresponds to a N_E/N_O = 30/200= 15% chance.

Example 2
Draw a card ⁙ from a shuffled standard deck of 52. What is the chance that it is a spade ♠?

There are $N_O = 52$ possible outcomes, but only $N_E =$ 13 cards are spades.

The probability of this event is
$\frac{N_E}{\mathrm{N}_O} = \frac{13}{52} = \frac{1}{4}$

Q3. A sample space is {(A,B),(A,A),(B,A),(B,B)} where A and B are categories.

If each outcome is an equally likely event, then what is the ***probability*** *of the two attributes being equal?*

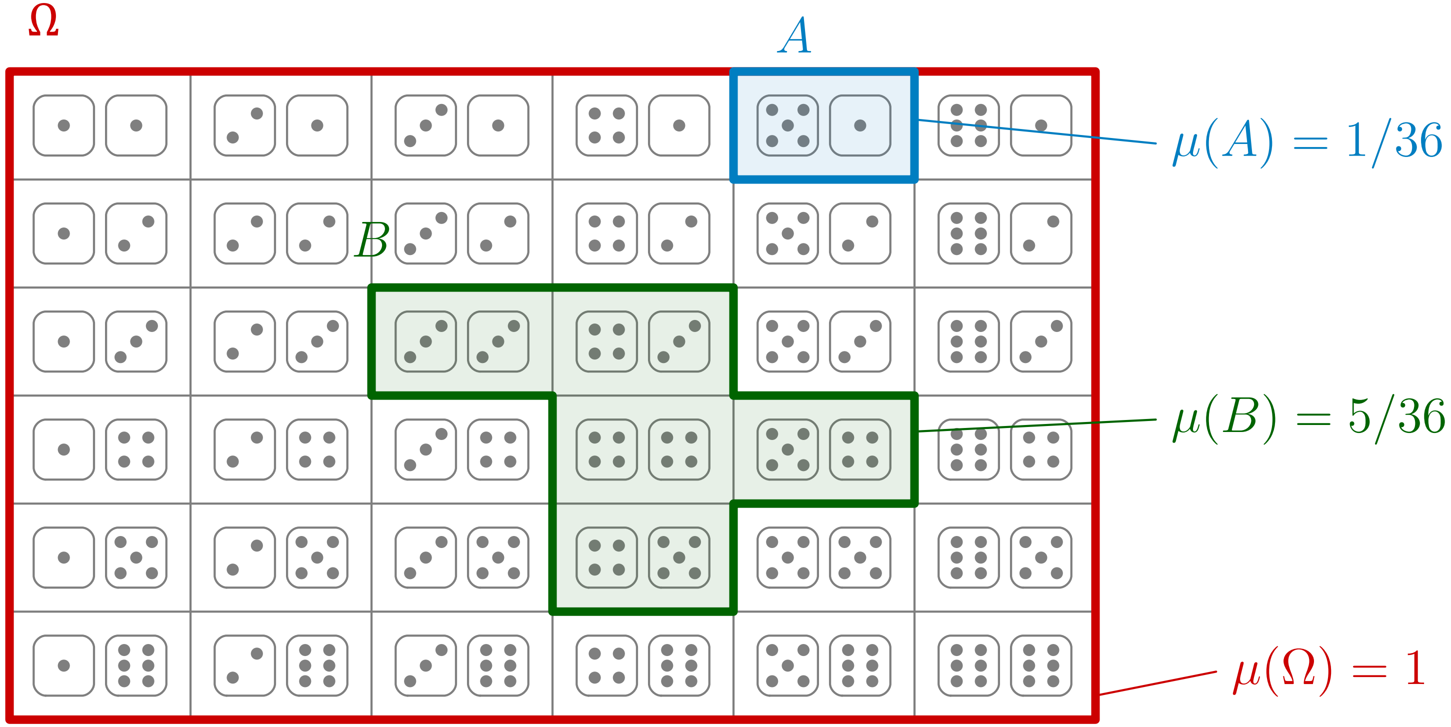
A. $\{(A, A), (B, B)\}$

B. 2

C. 2/4=1/2

D. ¼

Example with two six-sided dice



By Sascha Lill 95 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104138014

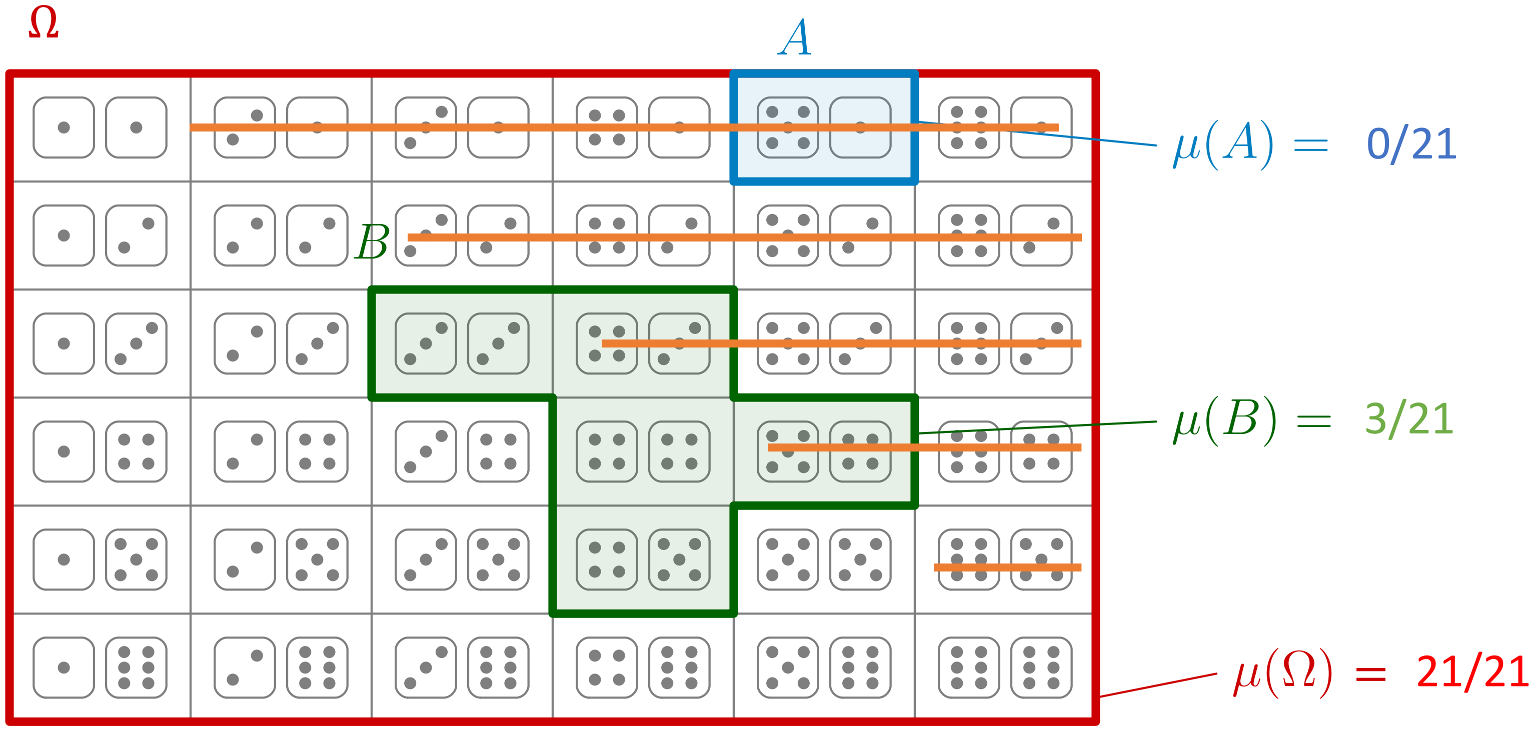
Throwing 2 dice (fair and independent) results in a sample space Ω of cardinality $|\Omega| = 36$.

Let $\mathcal{F}$ denote the event space.

Three different events $A, B, \Omega \in \mathcal{F}$ are outlined.

Their probabilities are denoted using the function $\mu: \mathcal{F} \to [0,1]$

Modified example with two six-sided dice



By Sascha Lill 95 - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=104138014

Throwing 2 dice in sequence **(the 2nd is rerolled until it is equal or greater than the first).**

What are the probabilities of A, B, Ω?

Probability theory: random variable

- A **random variable** X is a function from the sample space Ω to a measurable space (we will assume $\mathbb{R}$)
- Associates a probability to subsets of the measurable space

For $A \subseteq \mathbb{R},\ \Pr(X \in A) = \Pr(\{\omega \in \Omega : X(\omega) \in A\})$

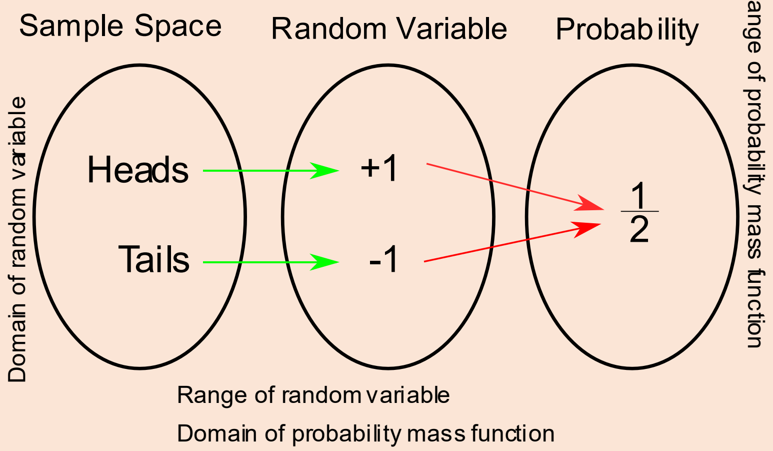
Example 0: X is the time of the first popcorn pop

$$\Pr\big(X \in [0, x)\big) = 1 - e^{-\frac{x}{4}}$$

$$\Pr\big(X \in [0,1.6)\big) = 0.32967$$

Example 1:

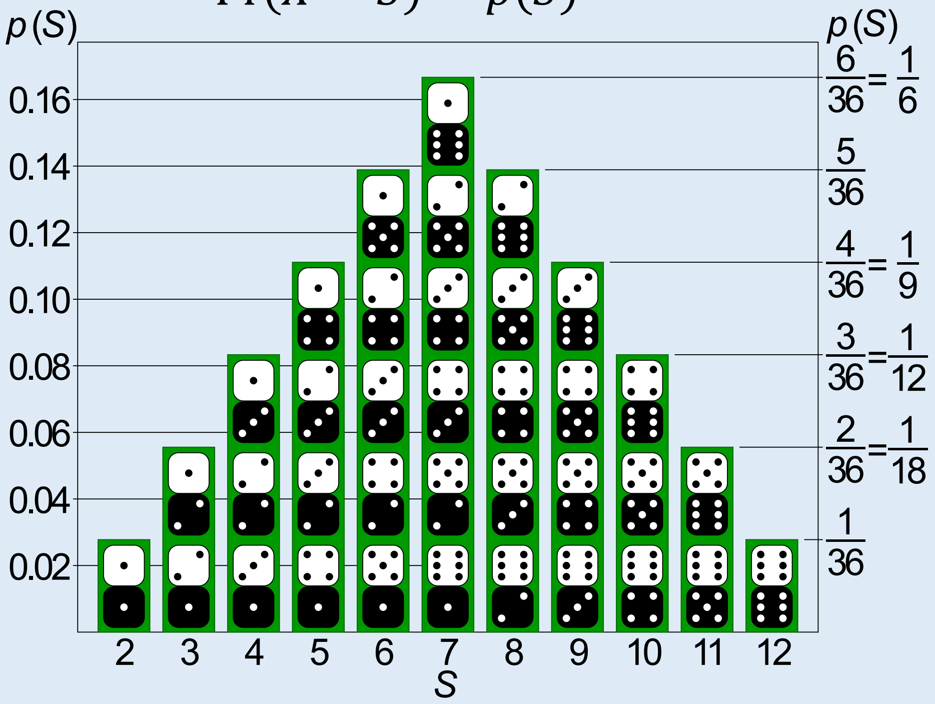
$$X(\omega) = \begin{cases} +1, \omega = \text{Heads} \\ -1, \omega = \text{Tails} \end{cases}$$



Sample Space
Random Variable
Probability
Domain of random variable
Heads
Tails
+1
-1
$\frac{1}{2}$
Range of probability mass function
Range of random variable
Domain of probability mass function

Example 2: X is the sum of pips on 2 fair & ind. dice

$$\Pr(X = S) = p(S)$$



Q4. Let *X* denote a random variable representing the temperature of particular location.

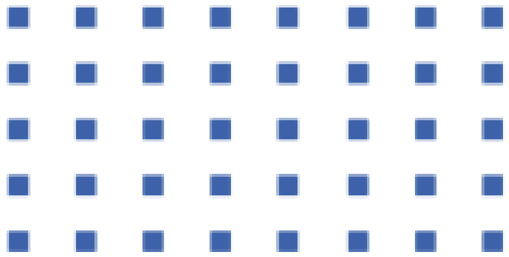
A. *X* is a discrete random variable since the set of outcomes is a discrete and countable set.

B. *X* is a continuous random variable since the temperature varies continuously. The set of outcomes is not countable.

C. *X* is an infinite random variable, because it can take infinite values.

D. *X* is is a finite random variable, because it has to be finite.

Q5. Let *X* denote the row index a geolocation on a regular grid.



A. *X* is a discrete random variable since the set of outcomes is a discrete and countable set.

B. *X* is a continuous random variable since the temperature varies continuously. The set of outcomes is not countable.

C. *X* is an infinite random variable, because it can take infinite values.

D. *X* is is a finite random variable, because it has to be finite.

Probability theory: probability mass function

- For a **discrete random variable** X, the **probability mass function** $p_X : \mathbb{R} \to [0,1]$

$$p_X(x) = \Pr(X = x) = \Pr(\{\omega \in \Omega : X(\omega) \in \{x\}\})$$

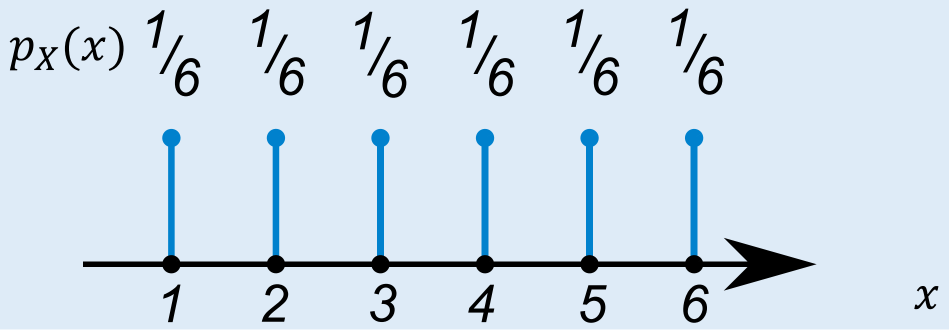
$$p_X(x) \geq 0, \qquad x \in \mathbb{R}$$

Let $\mathcal{X} = \{x \in \mathbb{R} : p_X(x) > 0\}$ be the set of values with non-zero probability.

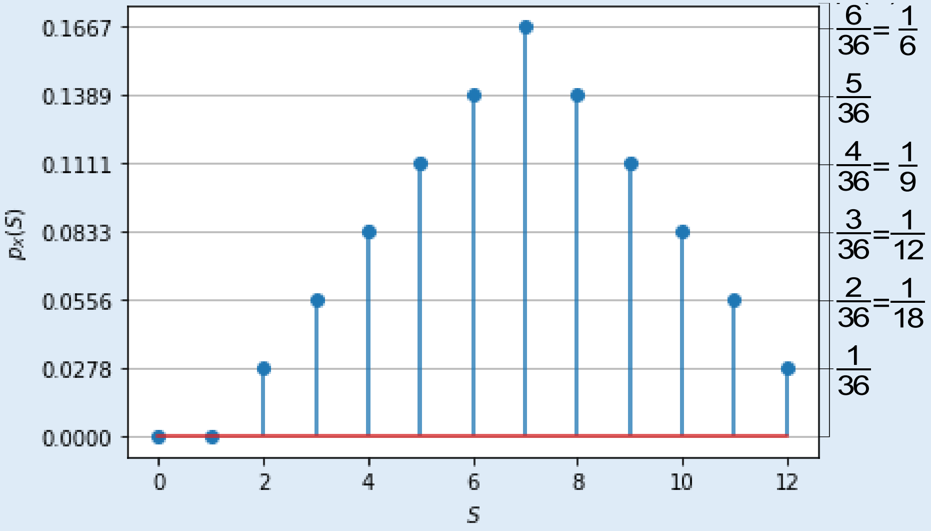
$$\sum_{x \in \mathcal{X}} p_X(x) = 1$$

$$\mathrm{mass}(x) = \frac{\mathrm{count}(x)}{\mathrm{total\ count}}$$

Example 2.0: X is the number rolled on a fair die

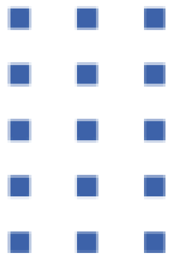


Example 2.1 X is the sum of numbers rolled on 2 dice



https://colab.research.google.com/drive/1sPij0xybc_1sfWiIV9guDphhg-DtEc74?usp=sharing

Q6. If points are uniformly distributed among the squares defined by the blue grid. What is the probability a point will have row index of 1?



A. 1/15

B. 1/3

C. 5

D. 1/5

Probability theory: probability density function

$$\text{density}(x) = \lim_{\text{area}\to 0} \frac{\text{count in area } x}{\text{area surrounding x}}$$

- For an **absolutely continuous random variable** X, the **probability density function** $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$

$$f_X(x) \geq 0, \qquad x \in \mathbb{R}$$

$$\Pr(X \in \ [a, b]\ ) = \Pr(a \leq X \leq \ b) = \int_a^b f_X(x)dx$$

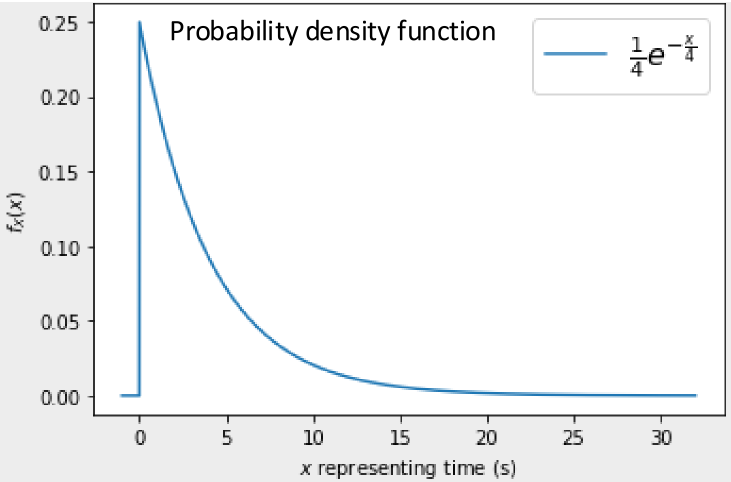
Define the cumulative distribution function $F_X(x) = \int_{-\infty}^{x} f_X(u)du$, $\qquad F_X(\infty) = 1$

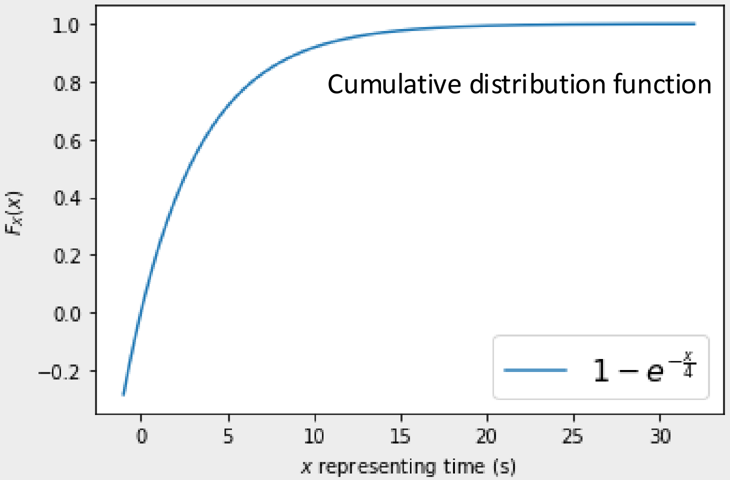
$$f_X(x) = \frac{d}{dx} F_X(x)$$

Example 0: X is the time of the first popcorn pop

$$F_X(x) = \Pr\big(X \in [0, x)\big) = 1 - e^{\frac{-x}{4}}$$

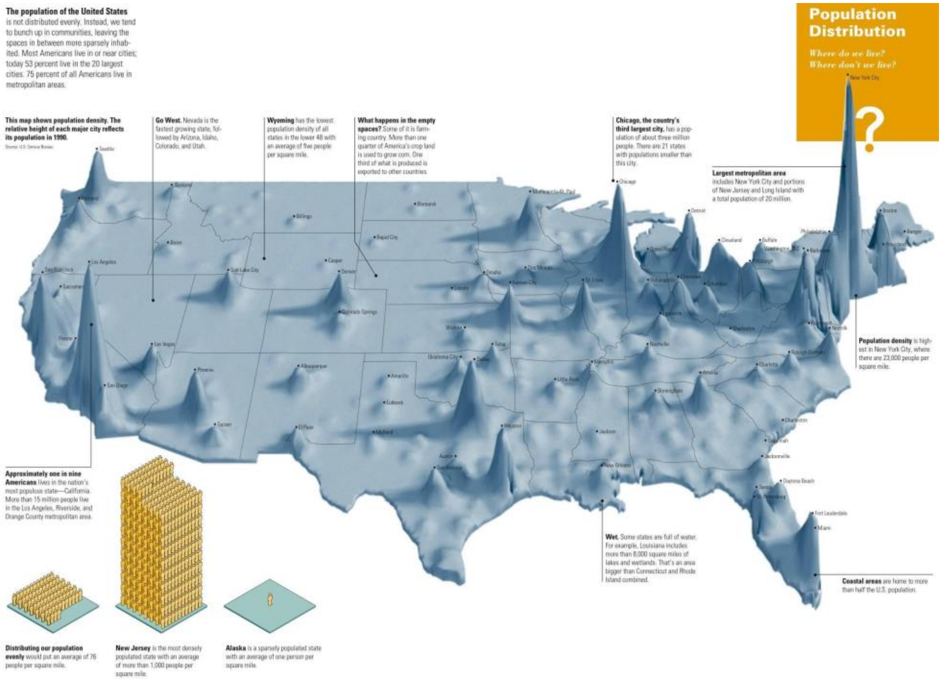
$$f_X(x) = \frac{1}{4} e^{-\frac{x}{4}}$$





Probability density function

- Joint random variable represent coordinates of a random person
- Population density = Total population × probability density for people



The population of the United States is not distributed evenly. Instead, we tend to bunch up in communities, leaving the spaces in between more sparsely inhabited. Most Americans live in or near cities; today 53 percent live in the 20 largest cities. 75 percent of all Americans live in metropolitan areas.

This map shows population density. The relative height of each major city reflects its population in 1990.

Source: U.S. Census Bureau

Go West. Nevada is the fastest growing state, followed by Arizona, Idaho, Colorado, and Utah.

Wyoming has the lowest population density of all states in the lower 48 with an average of five people per square mile.

What happens in the empty spaces? Some of it is farming country. More than one quarter of America's crop land is used to grow corn. One third of what is produced is exported to other countries.

Chicago, the country's third largest city, has a population of about three million people. There are 21 states with populations smaller than this city.

Largest metropolitan area includes New York City and portions of New Jersey and Long Island with a total population of 20 million.

Population density is highest in New York City, where there are 23,000 people per square mile.

Approximately one in nine Americans lives in the nation's most populous state—California. More than 15 million people live in the Los Angeles, Riverside, and Orange County metropolitan area.

Wet. Some states are full of water. For example, Louisiana includes more than 8,000 square miles of lakes and wetlands. That's an area bigger than Connecticut and Rhode Island combined.

Coastal areas are home to more than half the U.S. population.

Distributing our population evenly would put an average of 76 people per square mile.

New Jersey is the most densely populated state with an average of more than 1,000 people per square mile.

Alaska is a sparsely populated state with an average of one person per square mile.

https://i.redd.it/en5j44gfokf21.jpg

Mean: expected value

- Continuous **random variable** X, $\bar{X} = m_X = \mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)dx$
- Discrete **random variable** X, $\bar{X} = m_X = \mu_X = \mathbb{E}[X] = \sum_{x\in\mathcal{X}} x\, p_X(x)$

  Uniform probability case (say from data collected $\{x_1, x_2, \dots, x_n\}$)

  Sample average: $\widehat{m_X} = \sum_i x_i \frac{1}{n}$

Example 0: X is the time of the first popcorn pop

$$f_X(x) = \frac{1}{4}e^{-\frac{x}{4}}, \quad m_X = \frac{1}{\lambda} = 4$$

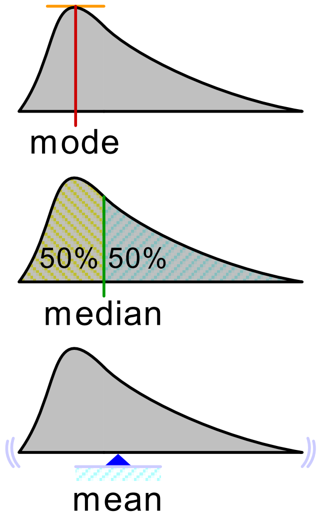
General case, $f_X(x;\lambda) = \lambda e^{-\lambda x}$

Example 2.0: X is the number rolled on a fair die

$$m_X = \sum_{x\in\{1,\dots,6\}} x\frac{1}{6} = 1\frac{1}{6} + 2\frac{1}{6} + \cdots + 6\frac{1}{6} = \frac{21}{6} = 3.5$$

Example 2.1 X is the sum of numbers rolled on 2 dice

$$m_X = 7$$



By Cmglee - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=38969094

Variance: expected value of the squared difference from the mean

Standard deviation: $\sigma_X = \sqrt{\mathrm{var}(X)}$

- Continuous **random variable** X, $\mathrm{var}(X) = \sigma_X^2 = \mathbb{E}[(X - m_X)^2] = \int_{-\infty}^{\infty} (x - m_X)^2 f_X(x) dx$
- Discrete **random variable** X,$\mathrm{var}(X) = \sigma_X^2 = \mathbb{E}[(X - m_X)^2] = \sum_{x \in \mathcal{X}} (x - m_X)^2 \, p_X(x)$

  Uniform probability case (say from data collected $\{x_1, x_2, \ldots, x_n\}$)

  An **unbiased** estimate: $\widehat{\mathrm{var}(X)} = \frac{1}{n-1} \sum_i (x_i - \bar{X})^2$

<u>**Example 0:**</u> X is the time of the first popcorn pop

$$f_X(x) = \frac{1}{4} e^{-\frac{x}{4}}$$

$$\sigma_X^2 = 16$$

https://en.wikipedia.org/wiki/Exponential_distribution

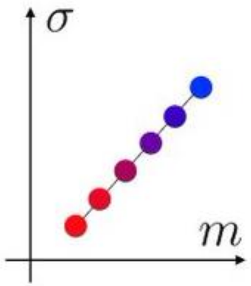
<u>Example 2.0:</u> X is the number rolled on a fair die

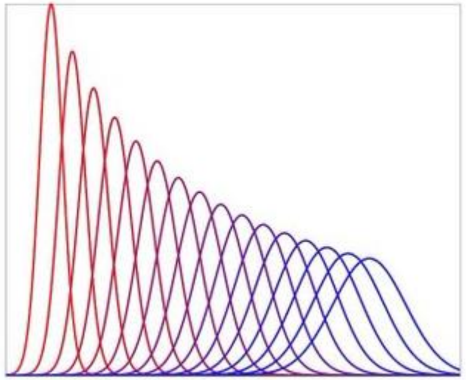
$$\sigma_X^2 = \sum_{x \in \{1,\ldots,6\}} (x - 3.5)^2 \frac{1}{6} = 2.91666$$

<u>Example 2.1</u> X is the sum of numbers rolled on 2 dice

$$\sigma_X^2 = 5.833$$

Gaussian distribution



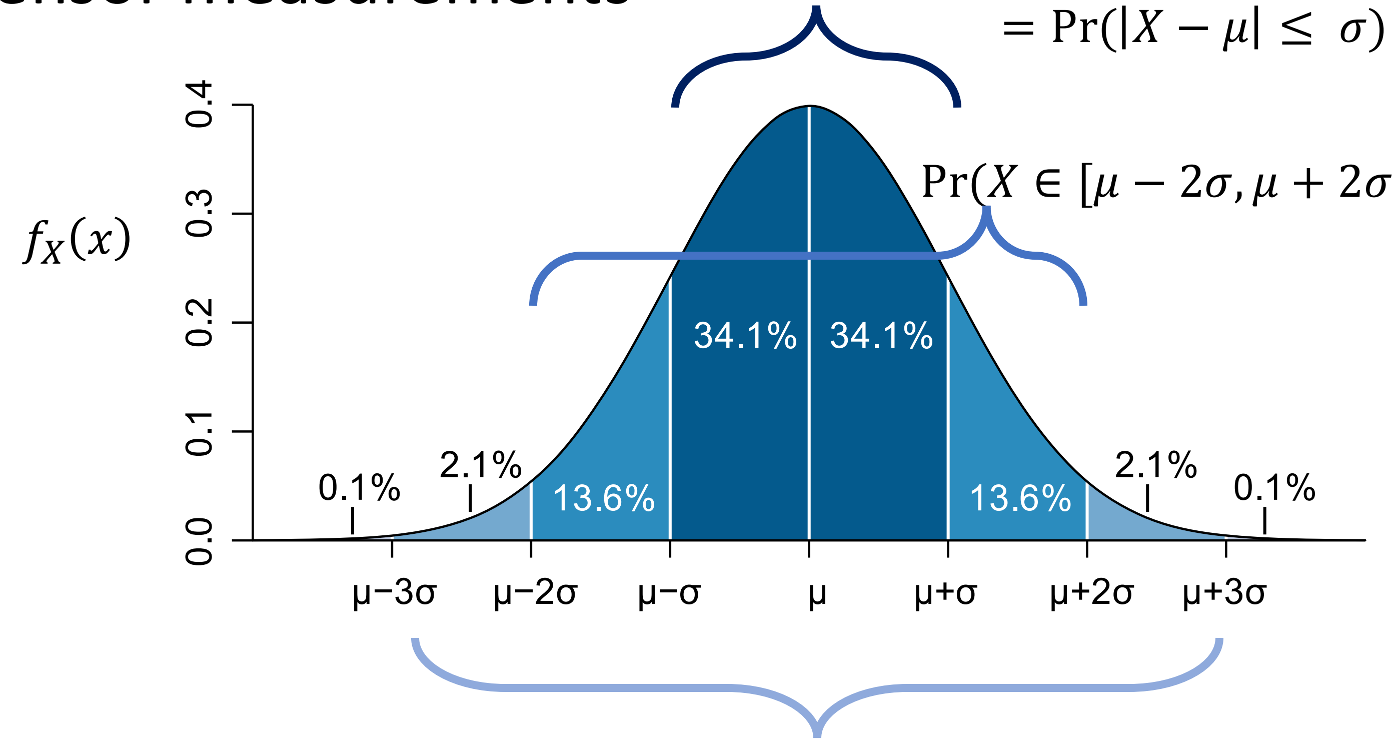


Gabriel Peyré
@gabrielpeyre

- Naturally describes sums of other random variables
- Error in sensor measurements

$$\Pr(X \in [\mu - \sigma, \mu + \sigma]) = \Pr(\mu - \sigma \leq X \leq \mu - \sigma)$$
$$= \Pr(|X - \mu| \leq \ \sigma) \approx 68.2\%$$

$$\Pr(X \in [\mu - 2\sigma, \mu + 2\sigma]) = \Pr(|X - \mu| \leq 2\sigma) \approx 95.4\%$$



$$\Pr(|X - \mu| \leq 3\sigma) \approx 99.6\%$$

By Ainali - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3141713

Estimators: bias, variance, mean squared error

How far off do I expect my estimate from limited data to be?

Expected bias of a parameter estimate: $\mathbb{E}\left[\widehat{\theta(X)}\right] - \theta^*$

Bias of sample mean: $\mathbb{E}[\widehat{m_X}] - m_X$

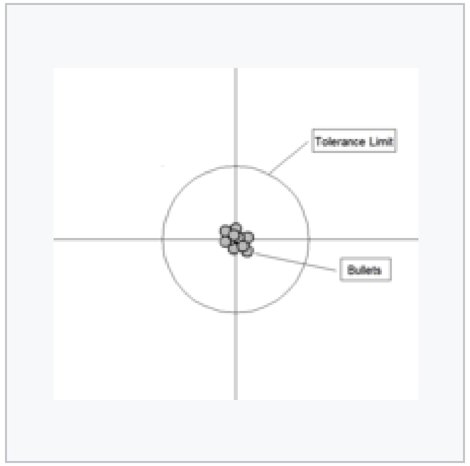
Bias of sample variance: $\mathbb{E}\left[\widehat{\mathrm{var}\,(X)}\right] - \mathrm{var}(X)$

Variance of a parameter estimate:

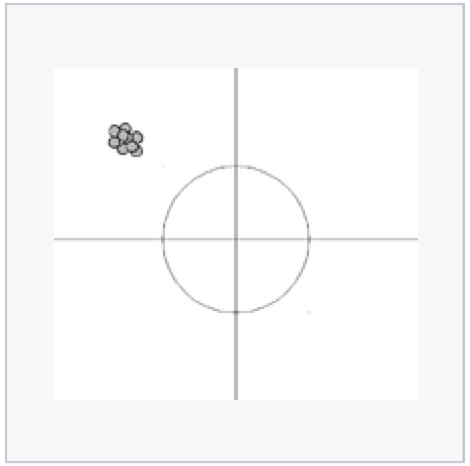
$$\mathbb{E}\left[\left(\widehat{\theta(X)} - \mathbb{E}[\widehat{\theta(X)}\,]\right)^2\right]$$

Variance of mean: $\mathbb{E}[(\widehat{m_X} - \mathbb{E}[\widehat{m_X}])^2]$

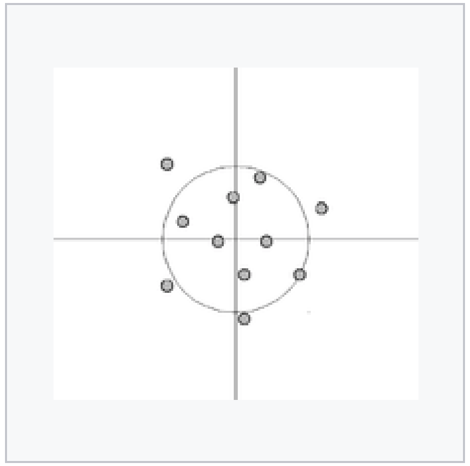
MSE of a parameter estimate: $\mathbb{E}\left[\left(\widehat{\theta(X)} - \theta^*\right)^2\right]$



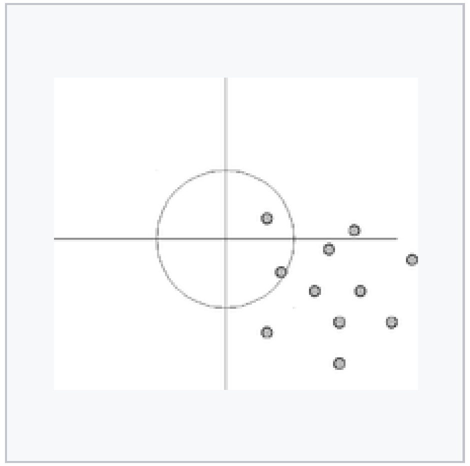
bias low, variance low



bias high,
variance low

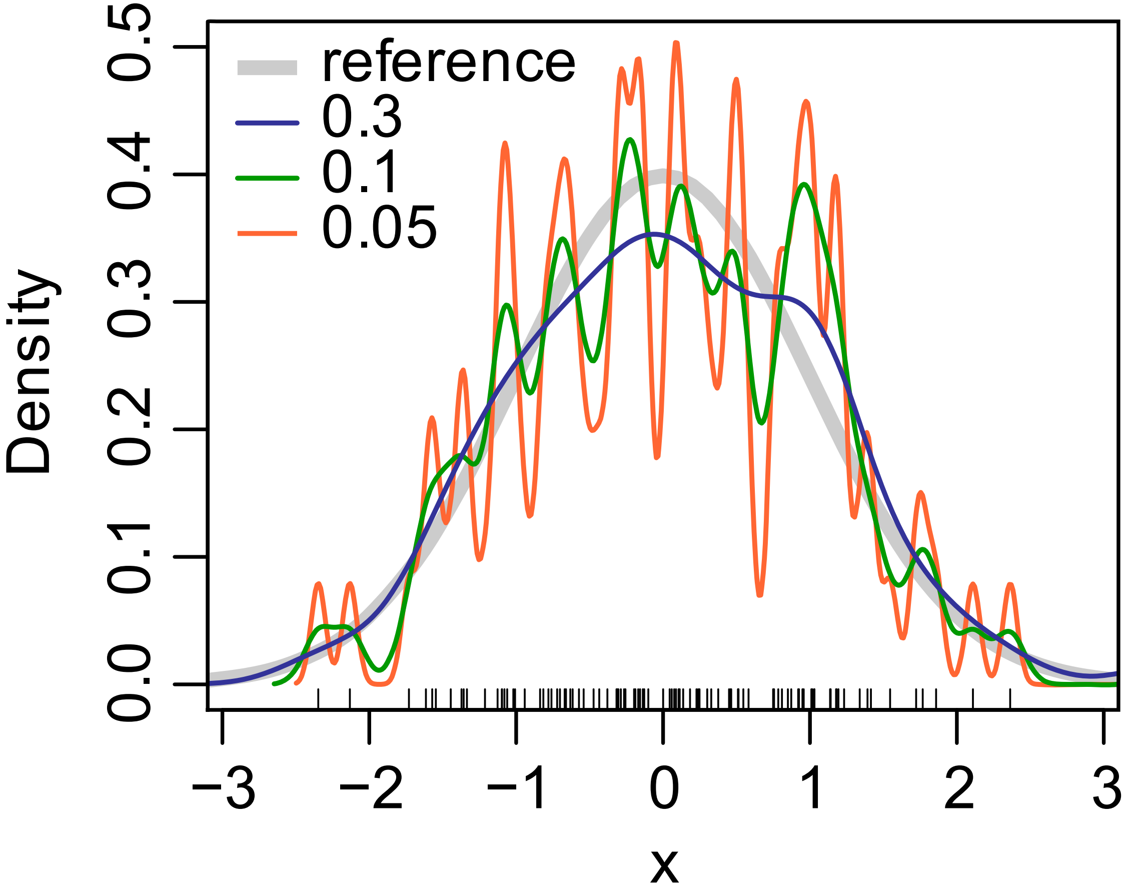


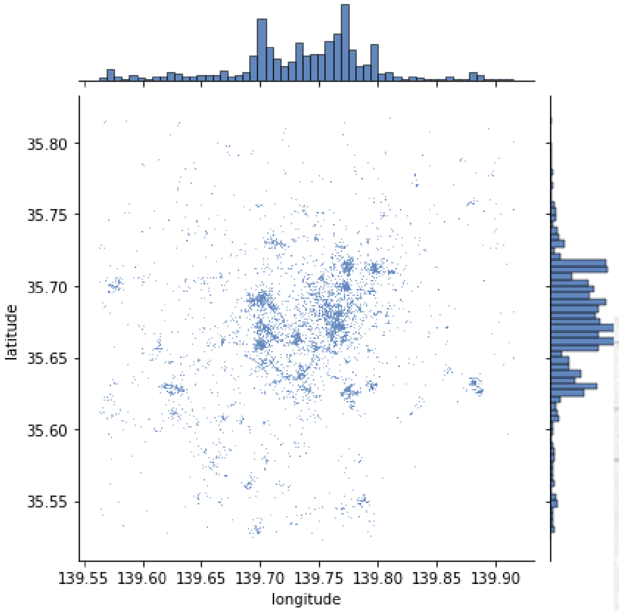
bias low,
variance high

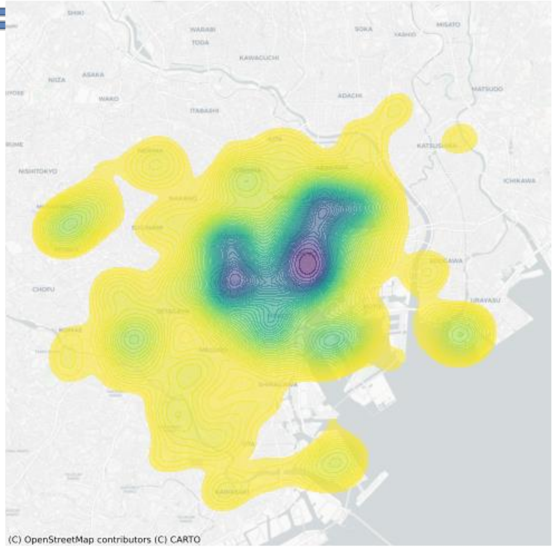


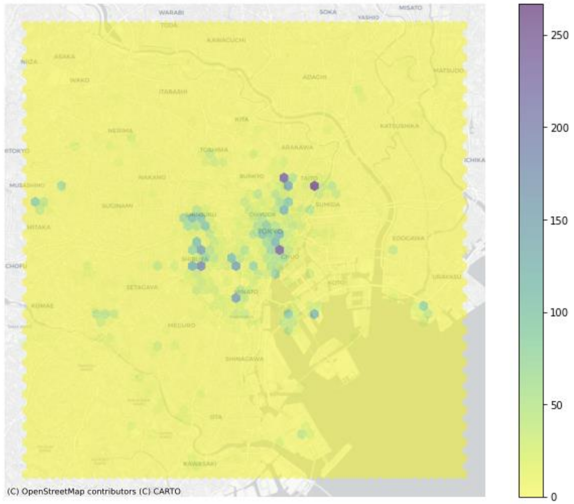
bias high,
variance high

Density as a parameter to estimate









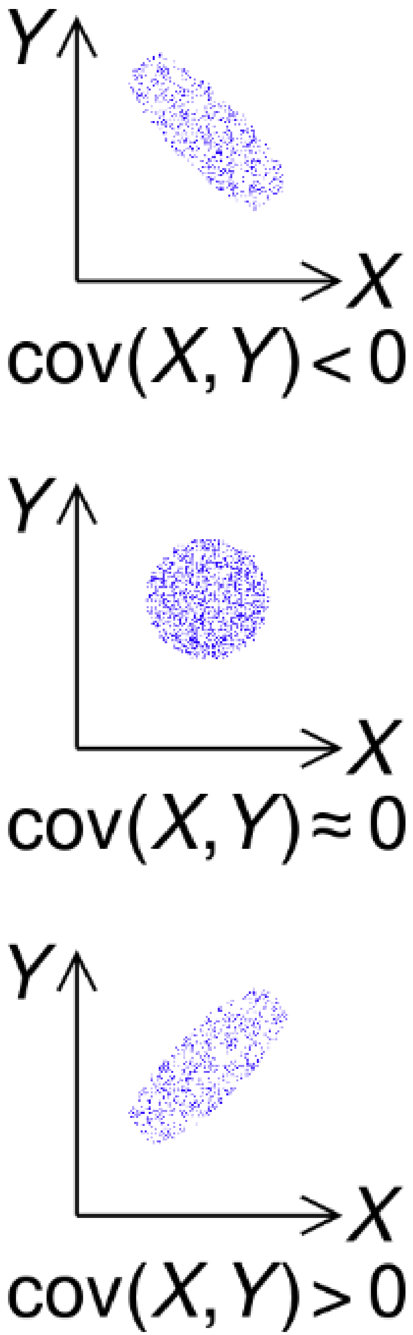
By M. W. Toews - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=2377345

Covariance: expected value of the product of two **centered** (difference from their mean) random variables

$$\mathrm{cov}(X,Y) = \mathbb{E}[(X - m_X)(Y - m_Y)]$$

An **unbiased** estimate is

$$\mathrm{cov}(X,Y) \ = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

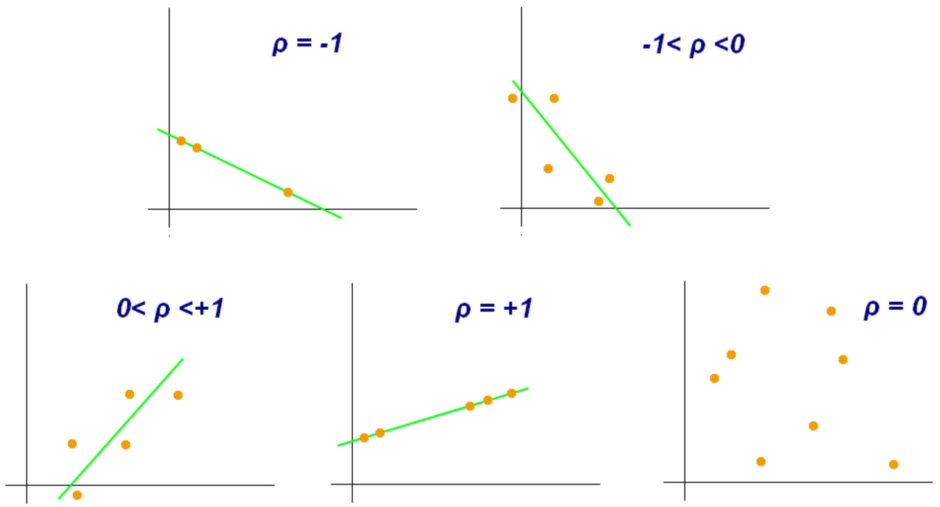


By Cmglee - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=90452334

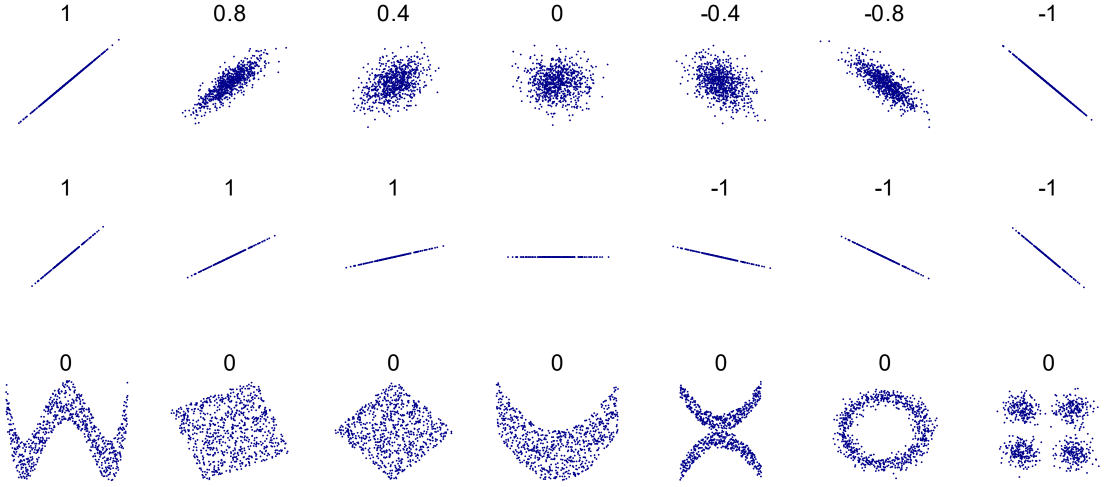
Linear correlation: normalized covariance

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \in [-1,1]$$

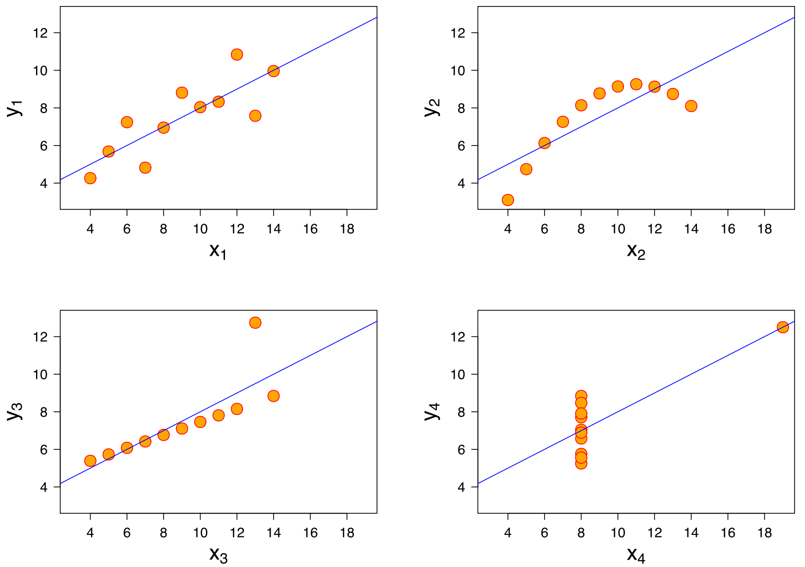
$$\boxed{\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}}$$



By Kiatdd - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=37108966

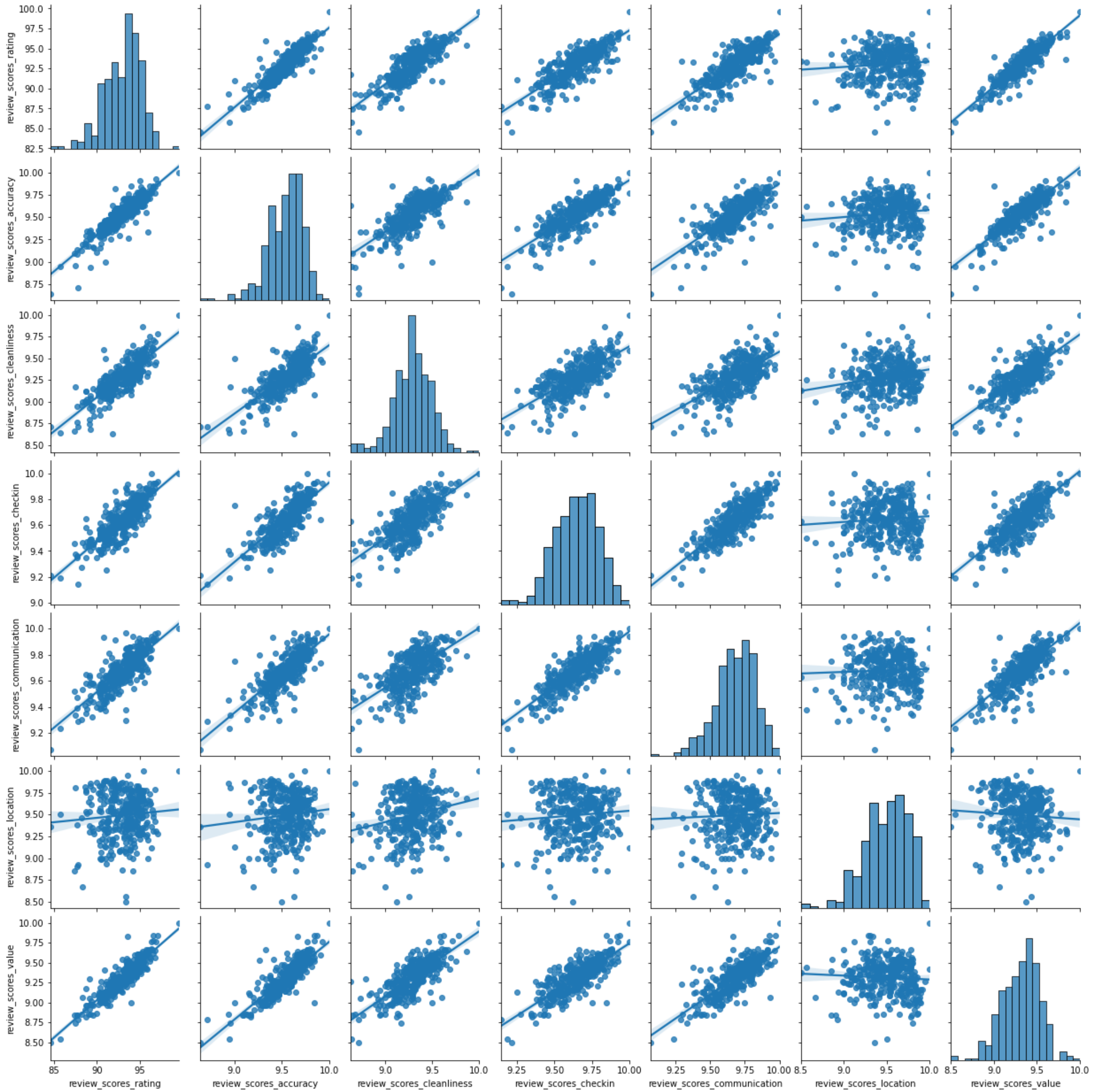


[Anscombe's quartet](): four sets of data with the same correlation of 0.816



Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

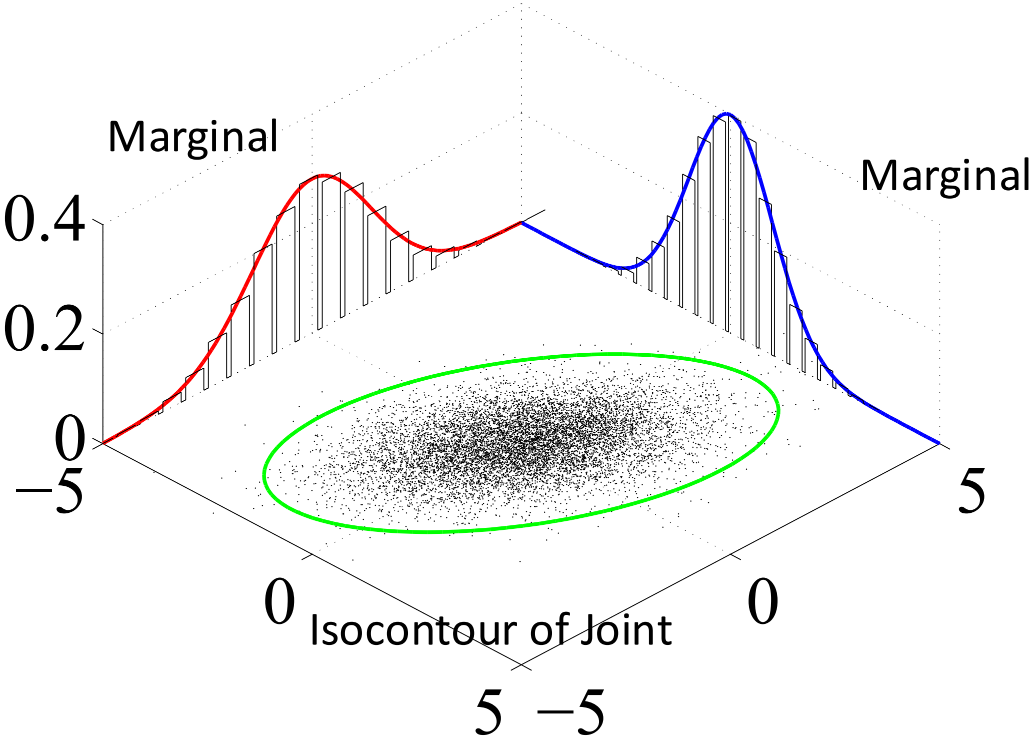
Comparing pairs of attributes

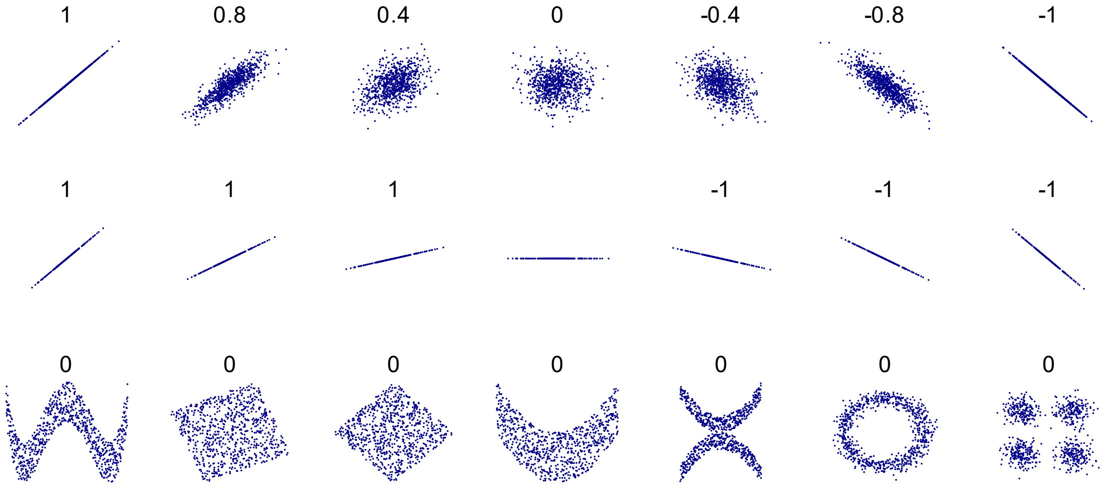


https://darribas.org/gds_course/content/bG/lab_G.html

Independence: no correlation only implies independence for Gaussian

If X and Y are independent,
Joint is equal to product of distribution/densities marginals
Discrete: $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y)$
or
Continuous: $f_{XY}(x, y) = f_X(x)\ f_Y(y)$

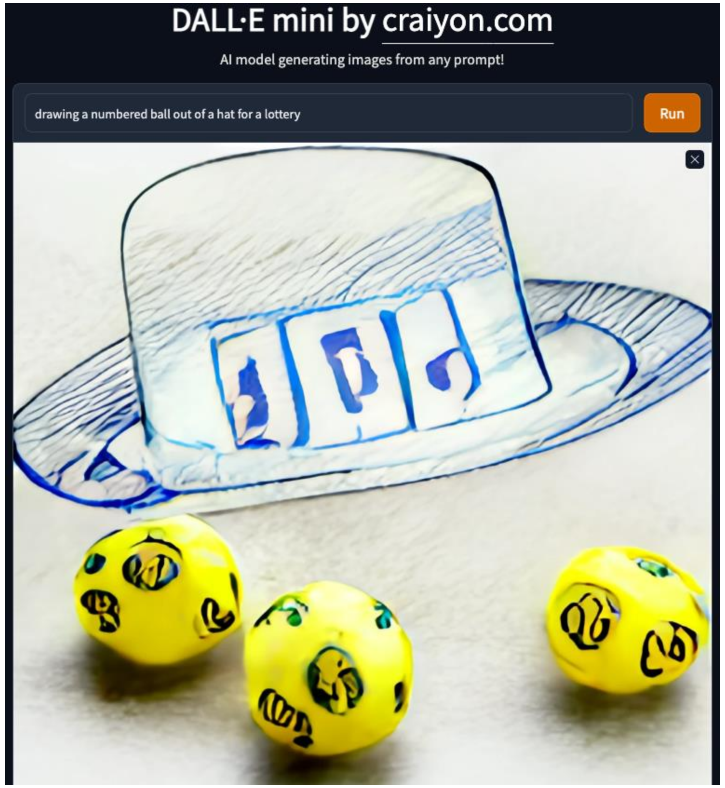




By IkamusumeFan - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=30432580

Upcoming in Lab 6

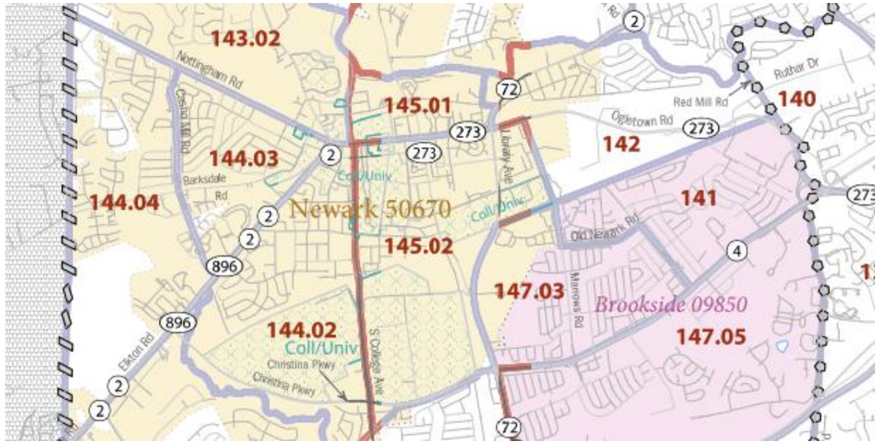
- Generating random numbers
- Basic statistics
- Bayes rule: joint and conditional
- Spatial correlation
- Counting: binning, histograms

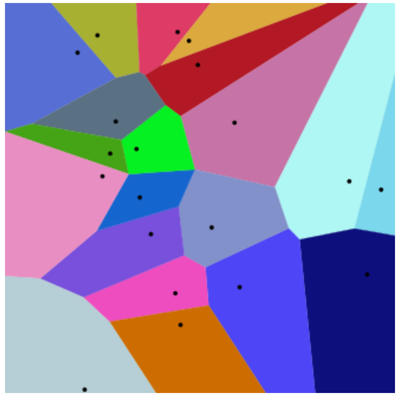


DALL·E mini by craiyon.com

AI model generating images from any prompt!

drawing a numbered ball out of a hat for a lottery



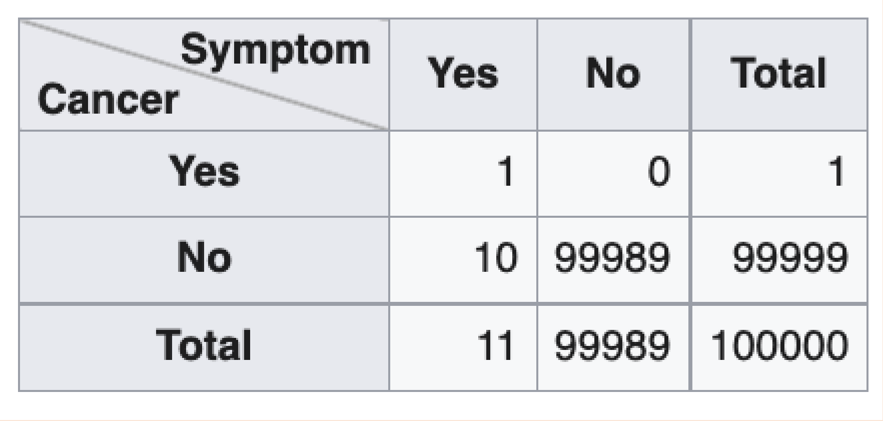


Bayes theorem

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

$$\boxed{f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y) f_X(x)}{f_Y(y)}}$$

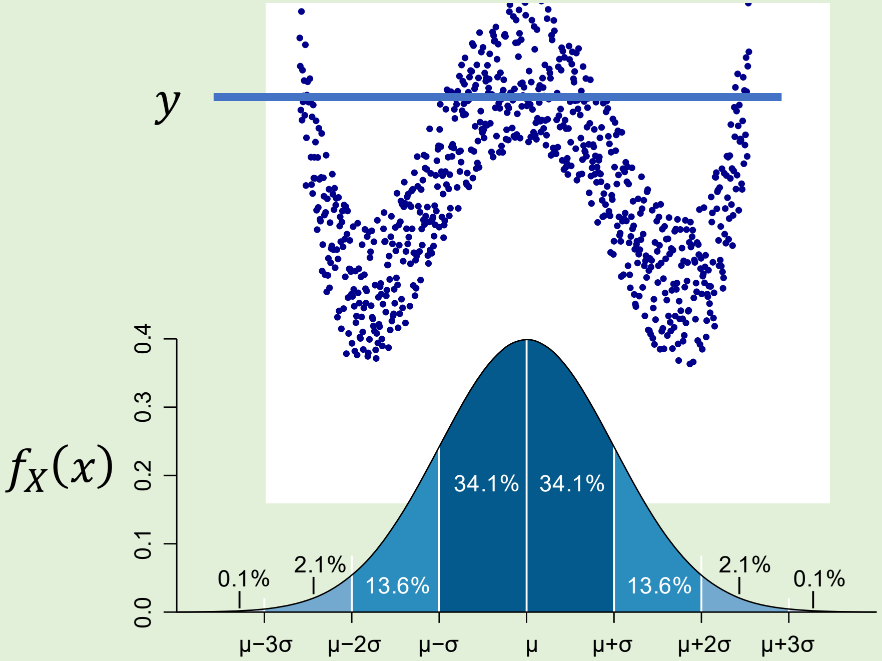
Example 3:



Cancer \ Symptom	Yes	No	Total
Yes	1	0	1
No	10	99989	99999
Total	11	99989	100000

$P(\text{Cancer}|\text{Symptom}) =$

Example 4:



By Ainali - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3141713

Co-relation of variables: autocorrelation and crosscorrelation/covariance in space and time

variable: attribute/measurement/feature

Applicable to **discrete objects** or **points in time or a spatial field**

- Autocorrelation
  - Is there a (linear) relationship in the value of a variable for objects at relatively nearby locations l_1, l_2 ?

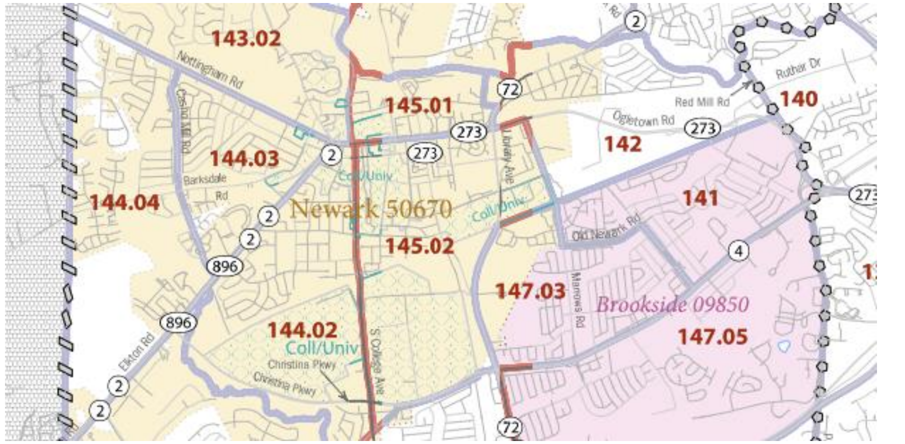
$$R_{XX}(l_1, l_2) = \rho_{X(l_1),X(l_2)}$$

- Cross-correlation
  - Is there a (linear) relationship in the value of the variables for relatively nearby objects/points?
  - $R_{XY}(l_1, l_2) = \rho_{X(l_1),Y(l_2)}$

How does the strength (variance explained) of the relationship vary as a function of the distance between the objects/points?

$$R_{XX}(d) = \mathbb{E}[R_{XX}(L_1, L_2) | \|L_1 - L_2\| = d], \quad d(l_1, l_2) = ||l_1 - l_2||$$

US census tracts designed to help understand the relationship between attributes

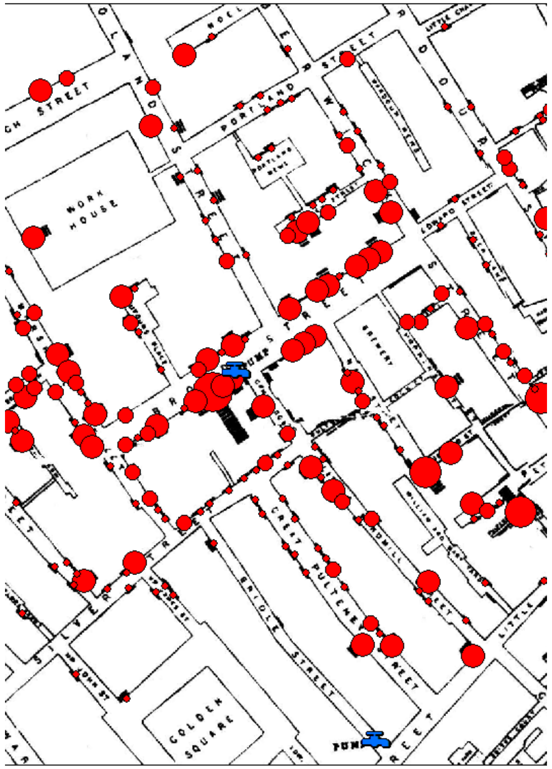


https://www2.census.gov/geo/maps/DC2020/PL20/st10_de/censustract_ma
10003_new_castle/DC20CT_C10003.pdf

"1st recorded […] delineation of small geographic entities based on population, topography, and housing characteristics were the sanitary districts […in the …] the 1890 census"

"sanitary districts […] used to **analyze** […] the **effect** of population, topography, and housing on the **mortality rate** of the inhabitants."

- FYI: In 1854 John Snow used data visualization and map to identify the source of cholera in London, England:



https://blog.rtwilson.com/wp-content/uploads/2012/01/SnowMap_Points.png