Statistical Data Analysis Lab

Anna Tkachev

Requirements

- ➔ Basic programming skills (I will be using Python)
- ➔ Basic knowledge of statistics and probability

Course overview

- Probability and distributions
- Statistics testing and assumptions
- Statistical significance
- Data transformation
- Basic predictive modeling

❖ Real-world datasets
❖ Synththically generated datasets
❖ Hands on: scripting and testing statistical tools and approaches

Grading

Only homework, according to Canvas schedule

Probability and Distributions

(one class)

Probability, statistics, randomness

Definition [edit]

The requirements for a set function μ to be a probability measure on a σ-algebra are that:

- μ must return results in the unit interval $[0, 1]$, returning 0 for the empty set and 1 for the entire space.
- μ must satisfy the *countable additivity* property that for all countable collections $E_1, E_2, \ldots$ of pairwise disjoint sets:

$$\mu\left(\bigcup_{i\in\mathbb{N}} E_i\right) = \sum_{i\in\mathbb{N}} \mu(E_i).$$

Probability, statistics, randomness

Definition [edit]

The requirements for a set function μ to be a probability measure on a σ-algebra are that:

- μ must return results in the unit interval $[0,1]$, returning 0 for the empty set and 1 for the entire space.
- μ must satisfy the *countable additivity* property that for all countable collections $E_1, E_2, \dots$ of pairwise disjoint sets:

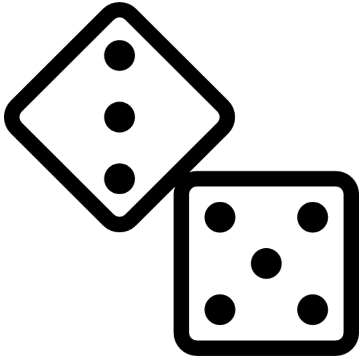
$$\mu\left(\bigcup_{i\in\mathbb{N}} E_i\right) = \sum_{i\in\mathbb{N}} \mu(E_i).$$

Now suppose that $(B, \mathcal{B}, \mu)$ is a measure space equipped with the counting measure μ. The probability density function f of X with respect to the counting measure, if it exists, is the Radon–Nikodym derivative of the pushforward measure of X (with respect to the counting measure), so $f = dX_*P/d\mu$ and f is a function from B to the non-negative reals. As a consequence, for any $b \in B$ we have

$$P(X=b) = P(X^{-1}(b)) = X_*(P)(b) = \int_b f d\mu = f(b),$$

demonstrating that f is in fact a probability mass function.

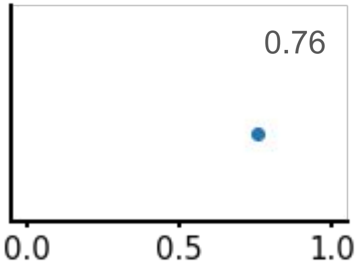
Probability, randomness, statistics



P(sum = 8) = P(2 and 6) + P(3 and 5) +P(4 and 4) +
P(5 and 3) + P(6 and 2) = 5/12

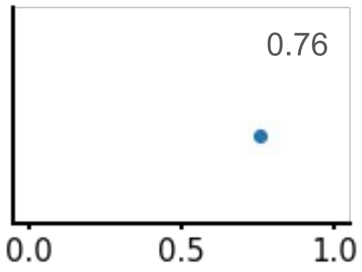
Probability, **randomness**, statistics

(random variable)

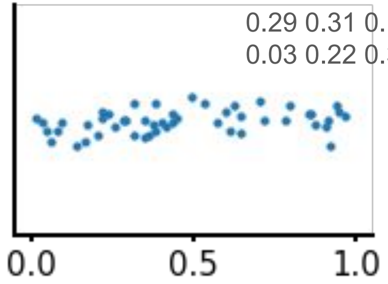


Probability, **randomness**, statistics

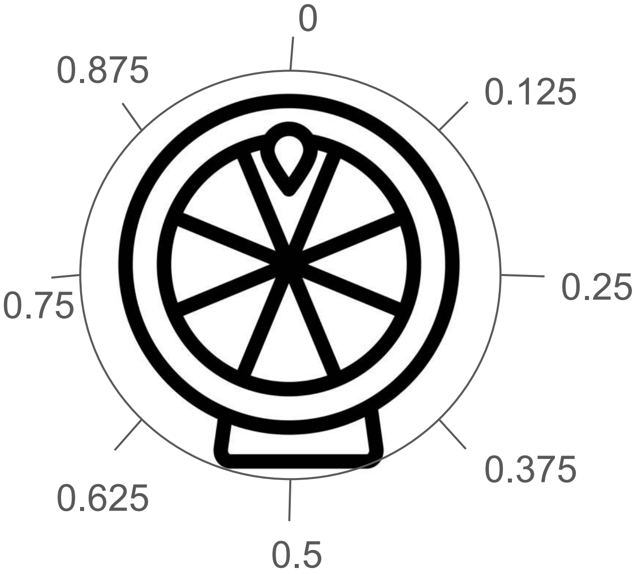
(random variable)

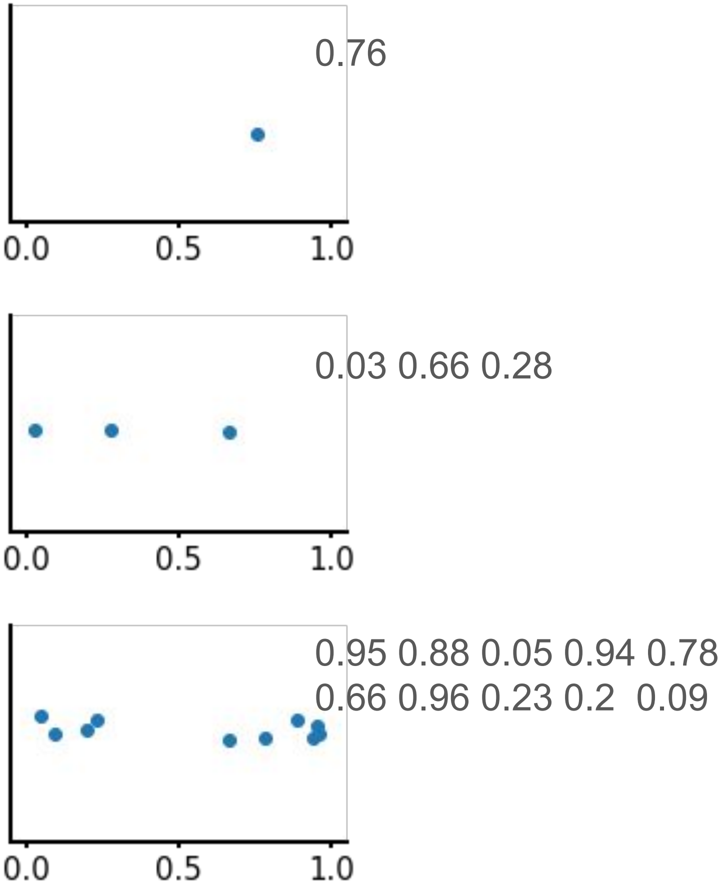


0.4 0.71 0.6 0.36 0.97 0.64 0.04 0.38 0.44 0.91 0.65
0.78 0.2 0.31 0.23 0.61 0.94 0.91 0.45 0.53 0.92 0.37
0.06 0.09 0.57 0.08 0.16 0.28 0.17 0.79 0.21 0.38
0.29 0.31 0.14 0.86 0.49 0.94 0.34 0.7 0.02 0.86 0.62
0.03 0.22 0.35 0.41 0.43 0.25 0.87

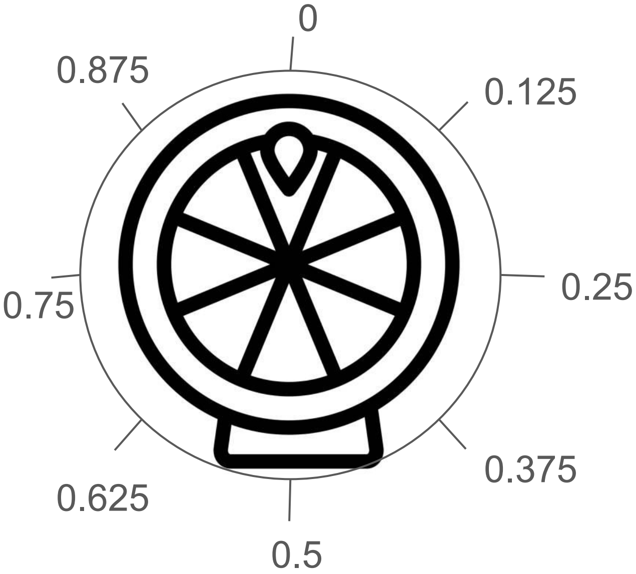


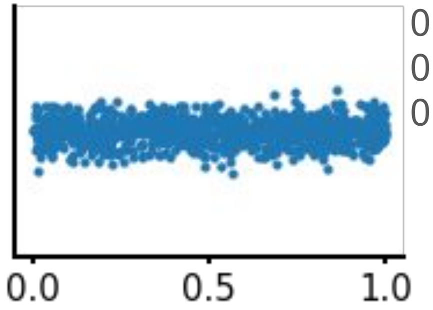
Probability, randomness, statistics

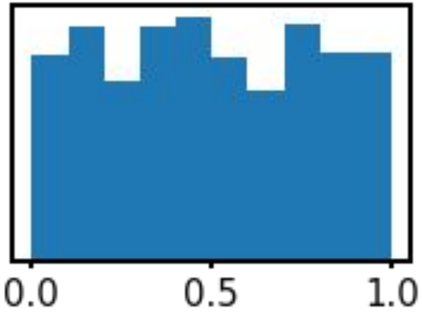




Probability, randomness, statistics

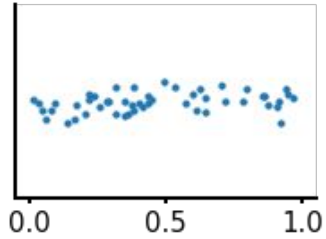


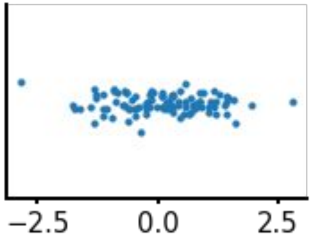


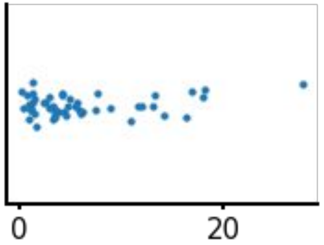


0.25	0.94	0.81	0.33	0.57	0.71
0.7	0.03	0.75	0.92	0.73	0.63
0.13	0.51	0.99	0.01	0.32	0.47
0.85	0.68	0.19	0.56	0.15	0.57
0.31	0.71	0.95	0.1	0.38 ...	

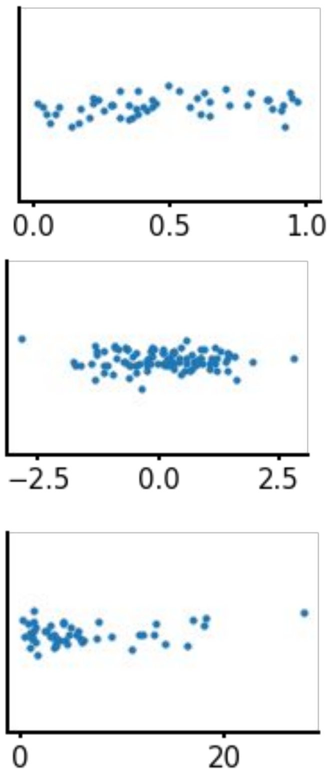
Probability, randomness, **statistics**







Probability, randomness, **statistics**



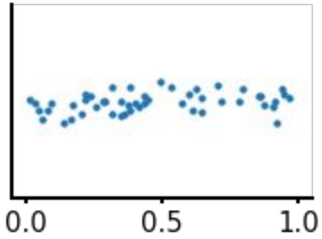
random

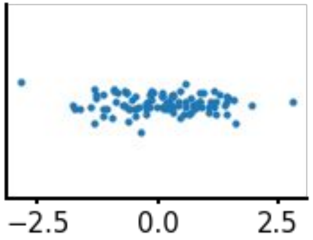
$X_1, X_2, X_3, \dots X_n$

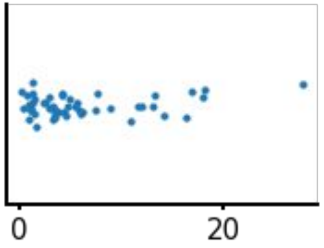
Laws of nature:

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

Probability, randomness, **statistics**







random

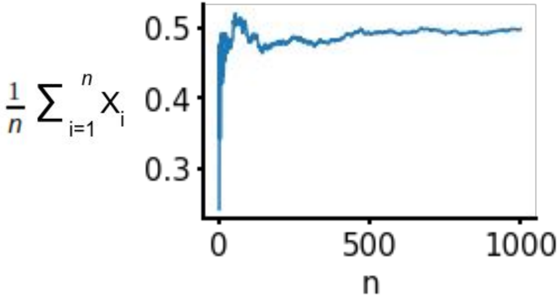
$X_1, X_2, X_3, \dots X_n$

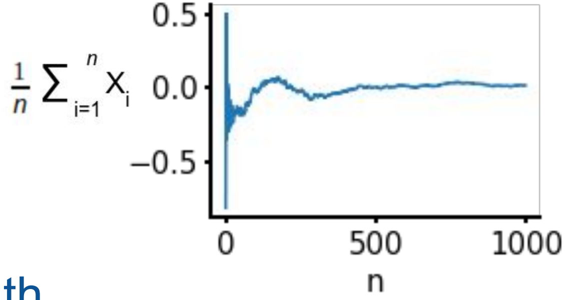
Law of nature:

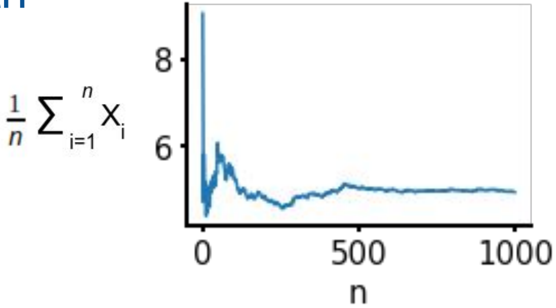
$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

will get more certain with larger *n*

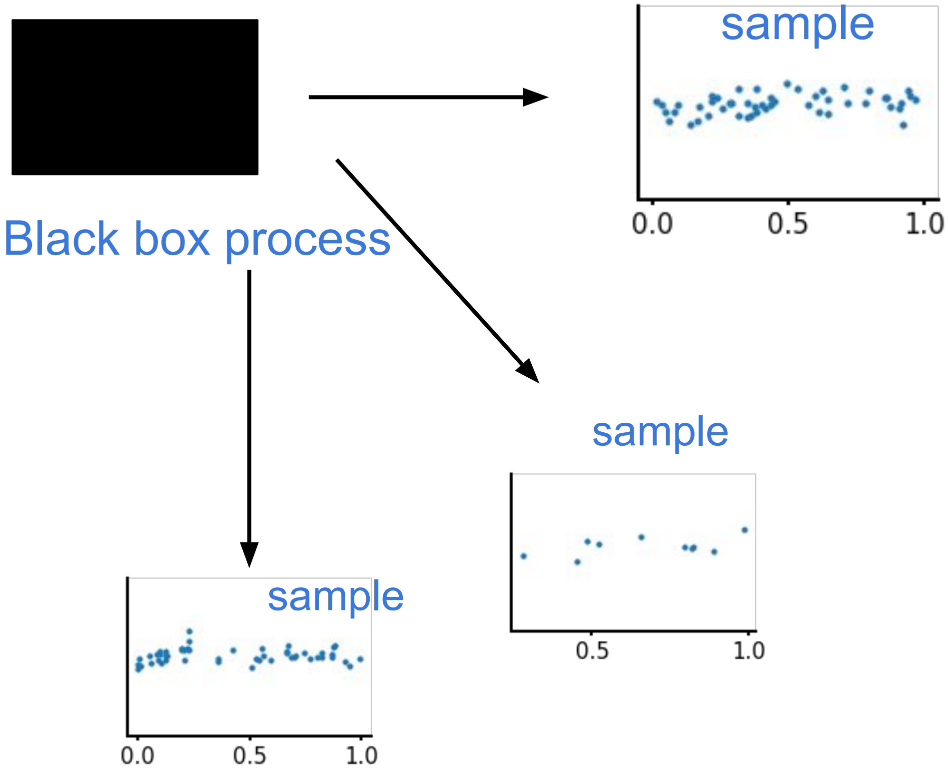
LLN



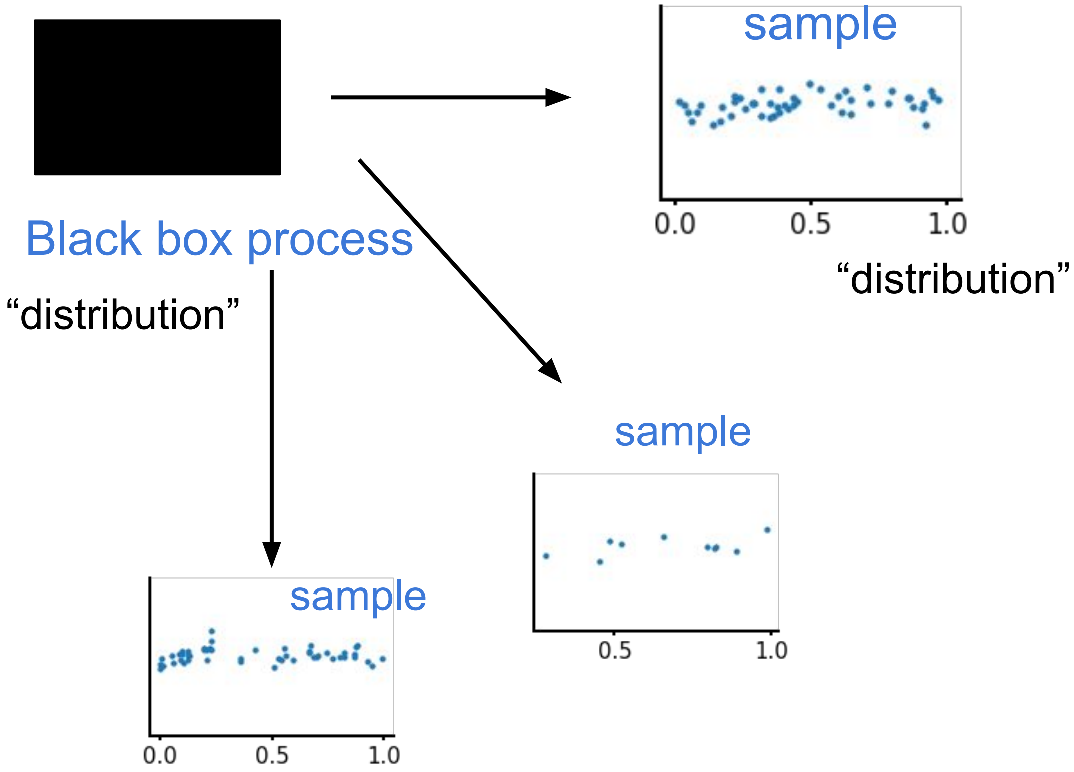




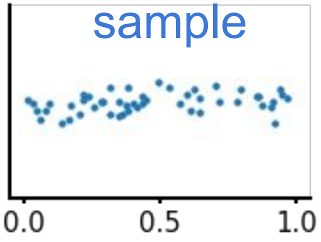
Describing distributions



Describing distributions

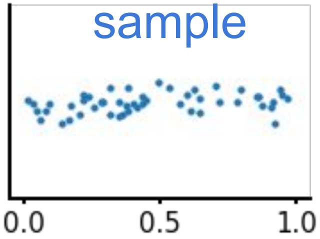


Describing distributions



“distribution”

Describing distributions



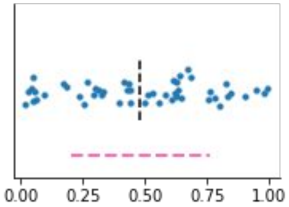
"distribution"

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

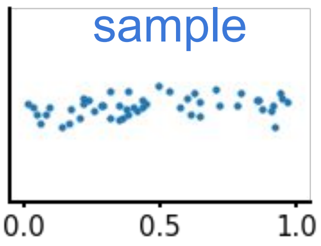
"center"

$$\sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n - 1}}$$

"spread"



Describing distributions



“distribution”

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

“center”

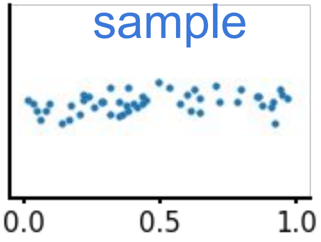
$$\sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n - 1}}$$

“spread”

General shape of distribution

Describing distributions

Black box process



"distribution"

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

"center"
mean

μ

$$\sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n - 1}}$$

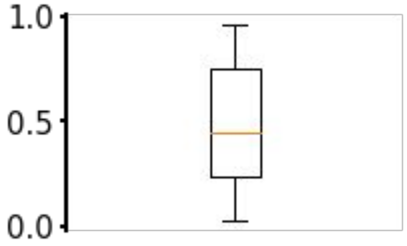
"spread"
standard deviation
std

σ

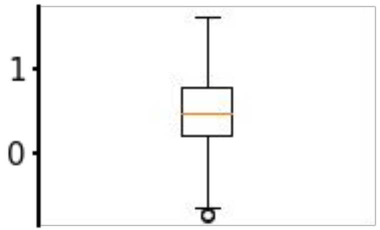
General shape of distribution

parametric families, ex. normal, uniform, exponential, etc.

Describing distributions

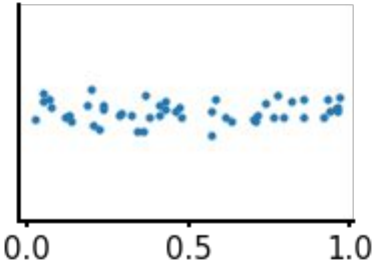


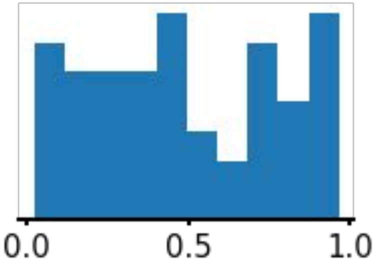
mean = 0.49
std = 0.29

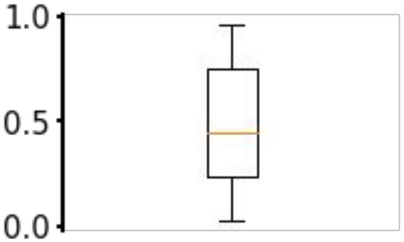


mean = 0.47
std = 0.48

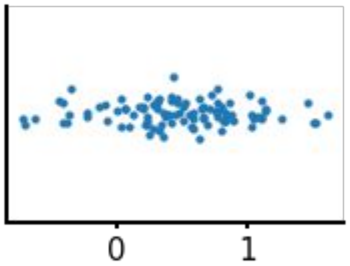
Describing distributions

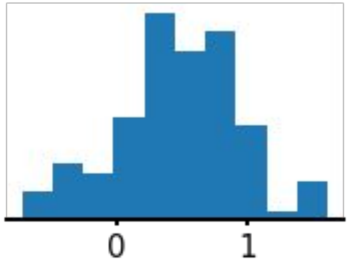


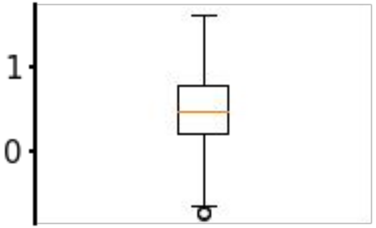




mean = 0.49
std = 0.29

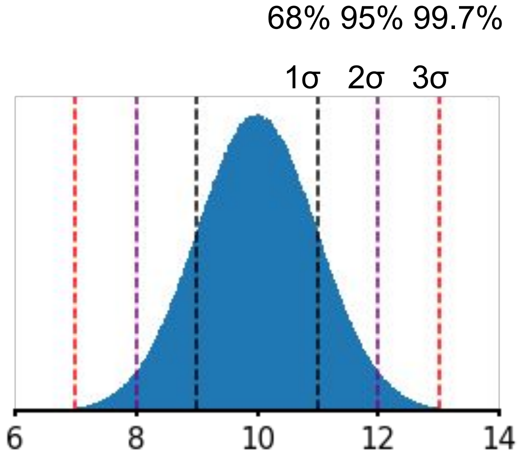






mean = 0.47
std = 0.48

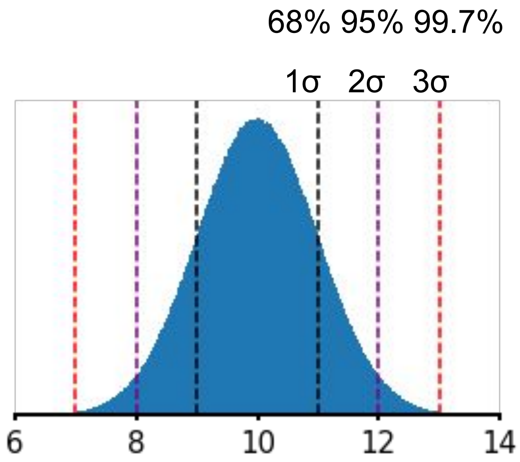
Normal distribution



ex: mean 10, std 1

Normal distribution

Where does it come from?



ex: mean 10, std 1

Normal distribution

Black box process

random

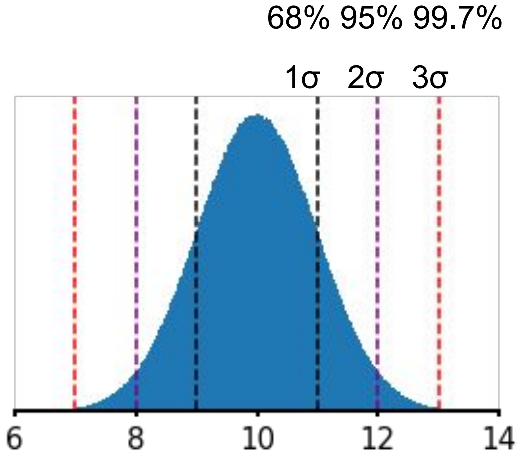
$X_1, X_2, X_3, \ldots X_n$

Law of nature:

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

will get more similar to a normal distribution with larger n

CLT



ex: mean 10, std 1

Normal distribution

Example in nature:

Human height

your height $\sim 1/2$(mother’s height + father’s height)

Normal distribution

Example in nature:

Human height

your height ~ 1/2(mother’s height + father’s height)

your height ~ 1/2(1/2(grandmother’s height + granfather’s height) +
1/2(grandmother’s height+ + granfather’s height))

...

your height ~ $\frac{1}{n} \sum_{i=1}^{n} X_i$

X_i - Individual height of your ancestor

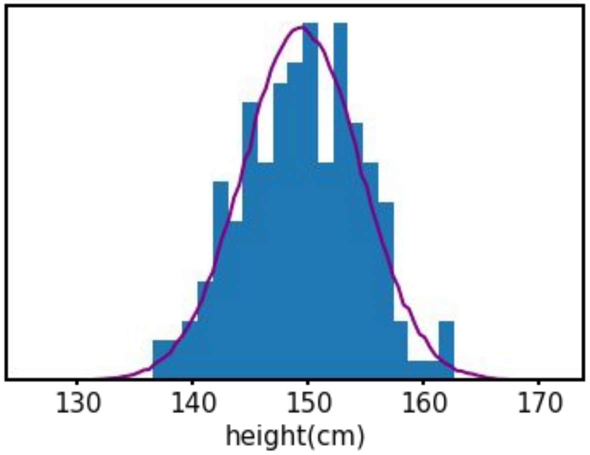
Could also reformulate this in terms of many independent genetic factors

Normal distribution

Example in nature:

Human height

174 samples,
> 20 years of age, female
(Howell1.csv)



Purple line:
generated normal distribution with estimated mean and std

Normal distribution

Distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$:

Black box process

random

$X_1, X_2, X_3, \ldots X_n$

Law of nature:

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

will get more similar to a normal distribution with larger n

CLT

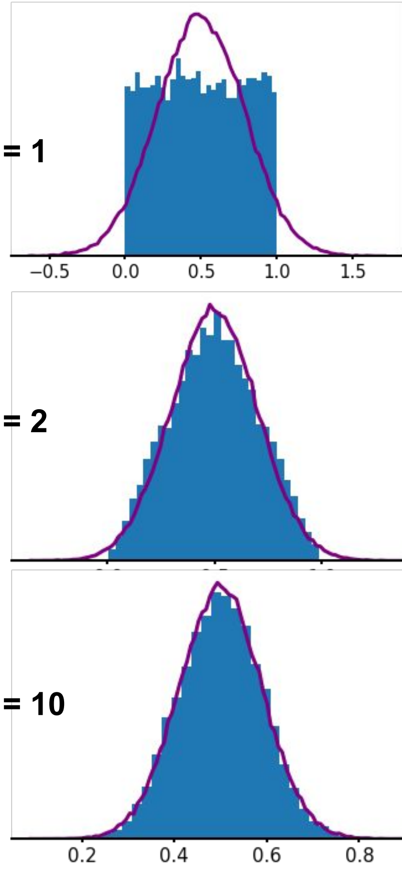
Normal distribution

Distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$:

n = 1

n = 2

n = 10



for $\mathbf{X_i}$ **- uniform distribution,**

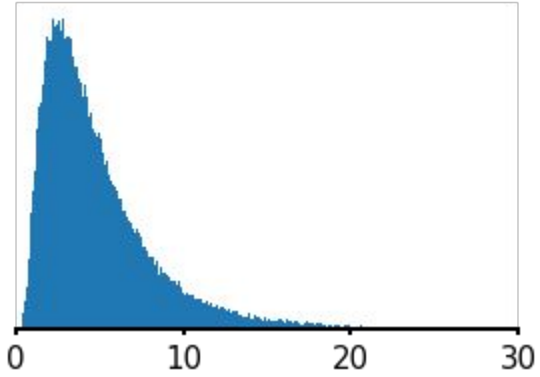
calculate $\frac{1}{n}\sum_{i=1}^{n} X_i$

and repeat 10000 times to generate the distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$

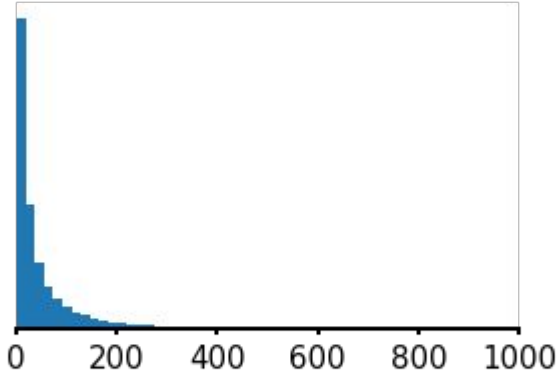
Lognormal distribution

Very common in nature, biology, and medicine as well

X_i is has lognormal distribution,
if the logarithm of X_i is normally distributed

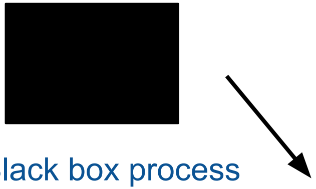


examples:



Log normal distribution

Where does it come from?



Black box process

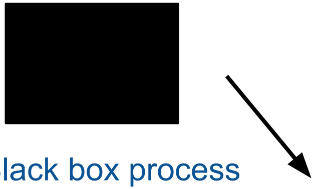
random

$$X_1, X_2, X_3, \ldots X_n$$

$$\prod_{i=1}^{n} X_i$$

will get more similar to a lognormal distribution with larger n

Log normal distribution



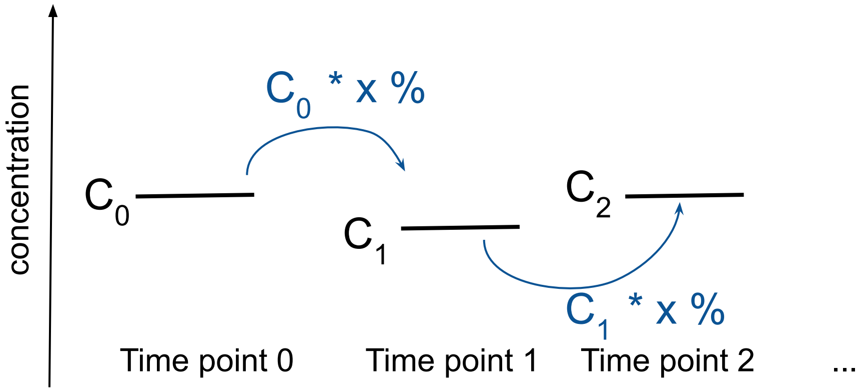
Black box process

C - constant

random

$X_1, X_2, X_3, \ldots X_n$

Where does it come from?



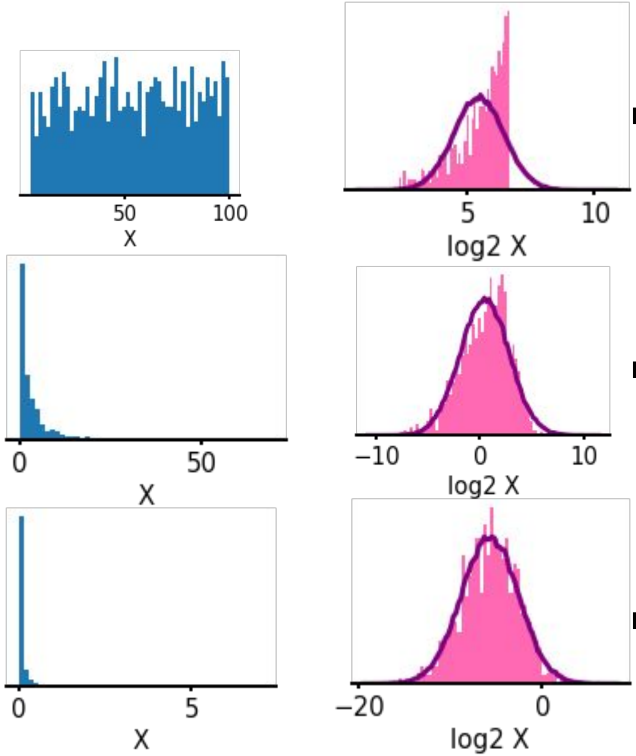
$$C^* \prod_{i=1}^{n} X_i$$

will get more similar to a lognormal distribution with larger n

Log normal distribution

$$\prod_{i=1}^{n} X_i$$

```
product_rand_sample=100
for i in np.arange(0,K):
        x=uniform(1,0.05,1000)
product_rand_sample=product_rand_sample*x
```



n = 1

n = 5

n = 10

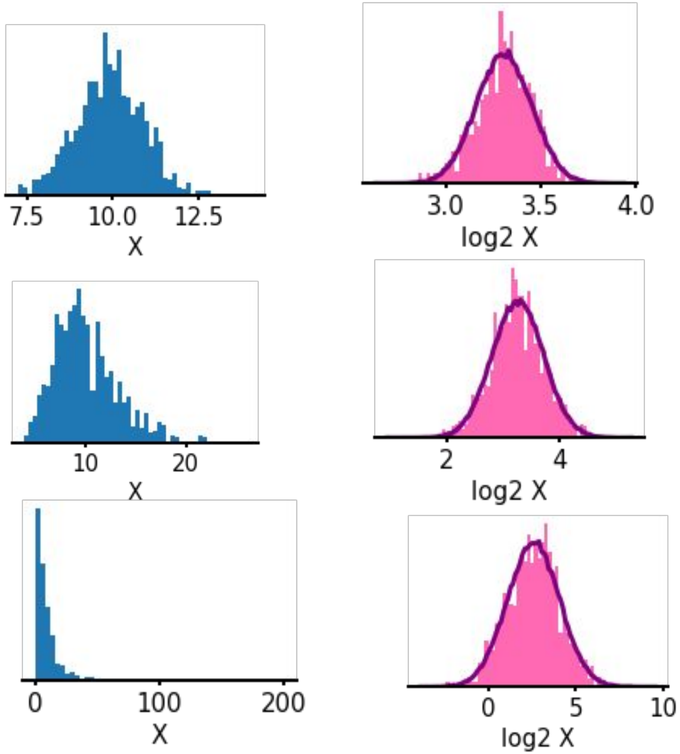
$$\prod_{i=1}^{n} X_i$$

```
product_rand_sample=10
for i in np.arange(0,K):
   x=normal(1,0.1,1000)
product_rand_sample=product_rand_sample*x
```

n = 1

n = 10

n = 100

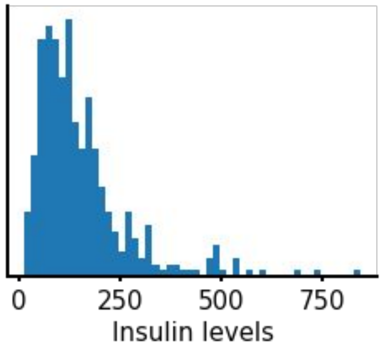


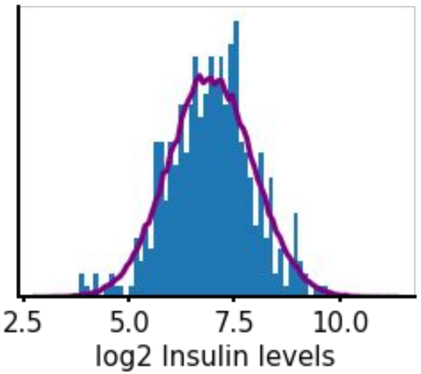
Log normal disitibution

Real world example:

n samples = 768

	CAR 10:1	CAR 12:1	CAR 18:2	CAR 18:1
PRECURSORINTENSITY:PI_GER_g1520_x5_pos_535.mzXML	13.9868355356746	13.7568182375003	13.8419376862341	14.5556110739153
PRECURSORINTENSITY:PI_GER_g1521_x5_pos_393.mzXML	13.7409130159943	12.9891351592992	14.0672933316656	13.9066482998761
PRECURSORINTENSITY:PI_GER_g1522_x5_pos_586.mzXML	13.5134309894261	13.2128612208716	13.7960614705597	14.2209741754035
PRECURSORINTENSITY:PI_GER_g1523_posx5_209.mzXML	14.3585864197179	13.753273432842	13.6868116704381	13.686974367965
PRECURSORINTENSITY:PI_GER_g1524_x5_pos_678.mzXML	14.0501662094891	13.4825236856777	13.9385369522722	14.1374823799142
PRECURSORINTENSITY:PI_GER_g1525_x5_pos_53.mzXML	14.3260462789677	13.0351690743192	13.6521219597924	13.9984763634375
PRECURSORINTENSITY:PI_GER_g1526_x5_pos_607.mzXML	13.5791747210104	13.4099018775543	13.6745992345387	14.136769413487
PRECURSORINTENSITY:PI_GER_g1527_x5_pos_149.mzXML	13.7328931321784	13.3035326766687	13.1183743337904	13.8391802206644
PRECURSORINTENSITY:PI_GER_g1528_x5_pos_620.mzXML	13.3172160543242	13.3865124321879	13.102865589948	14.1460553593803
PRECURSORINTENSITY:PI_GER_g1529_posx5_341.mzXML	13.4262432690065	12.7444423897913	13.8927467741453	13.6755301466585
PRECURSORINTENSITY:PI_GER_g1530_x5_pos_1094.mzXML	13.7197310034305	13.3832993321152	13.3434974789075	14.0896899836215
PRECURSORINTENSITY:PI_GER_g1531_x5_pos_83.mzXML	12.4364617880371	12.5416580011312	13.2692700285682	13.7421828206475
PRECURSORINTENSITY:PI_GER_g1532_x5_pos_567.mzXML	13.3551193273866	12.1406181976534	13.3700128857638	13.7912871544003
PRECURSORINTENSITY:PI_GER_g1533_posx5_281.mzXML	14.3520268785438	13.4683287650172	13.0853867945694	13.1196127474753
PRECURSORINTENSITY:PI_GER_g1534_x5_pos_110.mzXML	14.5072449140457	13.9698695366852	13.3307775372826	13.9274144221603
PRECURSORINTENSITY:PI_GER_g1535_x5_pos_544.mzXML	12.2498961710926	11.701271792652	12.5462480545009	13.2843472453556





Log normal disitibution

Real world example:

n features = 235

	CAR 10:1	CAR 12:1	CAR 18:2	CAR 18:1
PRECURSORINTENSITY:PI_GER_g1520_x5_pos_535.mzXML	13.9868355356746	13.7568182375003	13.8419376862341	14.5556110739153
PRECURSORINTENSITY:PI_GER_g1521_x5_pos_393.mzXML	13.7409130159943	12.9891351592992	14.0672933316656	13.9066482998761
PRECURSORINTENSITY:PI_GER_g1522_x5_pos_586.mzXML	13.5134309894261	13.2128612208716	13.7960614705597	14.2209741754035
PRECURSORINTENSITY:PI_GER_g1523_posx5_209.mzXML	14.3585864197179	13.753273432842	13.6868116704381	13.686974367965
PRECURSORINTENSITY:PI_GER_g1524_x5_pos_678.mzXML	14.0501662094891	13.4825236856777	13.9385369522722	14.1374823799142
PRECURSORINTENSITY:PI_GER_g1525_x5_pos_53.mzXML	14.3260462789677	13.0351690743192	13.6521219597924	13.9984763634375
PRECURSORINTENSITY:PI_GER_g1526_x5_pos_607.mzXML	13.5791747210104	13.4099018775543	13.6745992345387	14.136769413487
PRECURSORINTENSITY:PI_GER_g1527_x5_pos_149.mzXML	13.7328931321784	13.3035326766687	13.1183743337904	13.8391802206644
PRECURSORINTENSITY:PI_GER_g1528_x5_pos_620.mzXML	13.3172160543242	13.3865124321879	13.102865589948	14.1460553593803
PRECURSORINTENSITY:PI_GER_g1529_posx5_341.mzXML	13.4262432690065	12.7444423897913	13.8927467741453	13.6755301466585
PRECURSORINTENSITY:PI_GER_g1530_x5_pos_1094.mzXML	13.7197310034305	13.3832993321152	13.3434974789075	14.0896899836215
PRECURSORINTENSITY:PI_GER_g1531_x5_pos_83.mzXML	12.4364617880371	12.5416580011312	13.2692700285682	13.7421828206475
PRECURSORINTENSITY:PI_GER_g1532_x5_pos_567.mzXML	13.3551193273866	12.1406181976534	13.3700128857638	13.7912871544003
PRECURSORINTENSITY:PI_GER_g1533_posx5_281.mzXML	14.3520268785438	13.4683287650172	13.0853867945694	13.1196127474753
PRECURSORINTENSITY:PI_GER_g1534_x5_pos_110.mzXML	14.5072449140457	13.9698695366852	13.3307775372826	13.9274144221603
PRECURSORINTENSITY:PI_GER_g1535_x5_pos_544.mzXML	12.2498961710926	11.701271792652	12.5462480545009	13.2843472453556