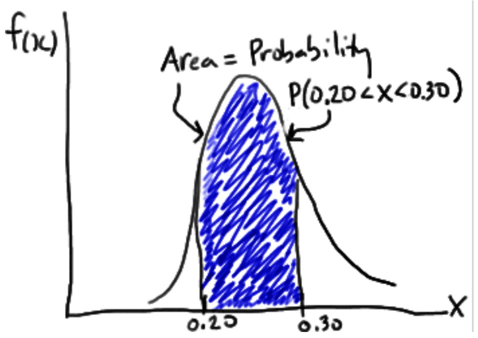
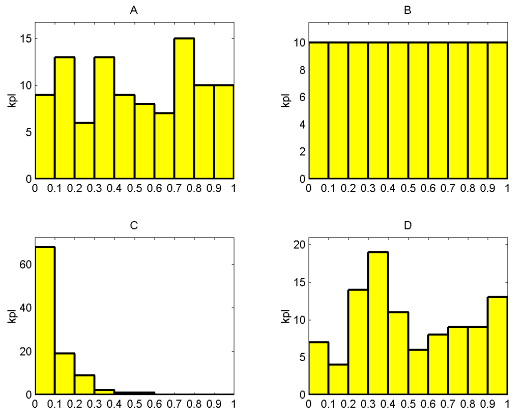
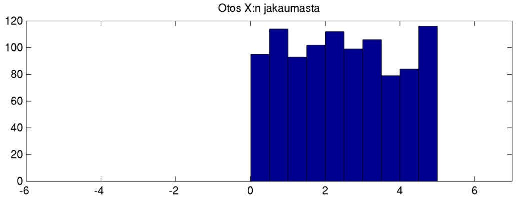
Lecture 6B -- miscellaneous



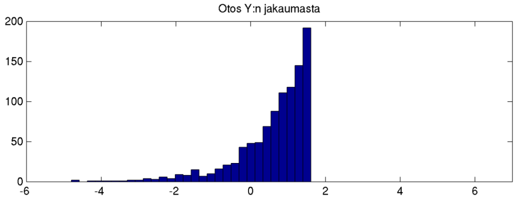
- Two approaches: Analytical and empirical.
  Case study: sums of fair dice



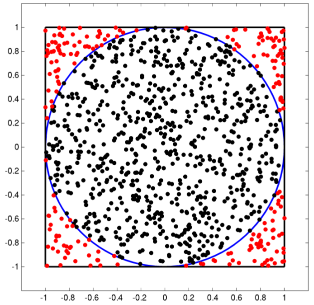
- A refresher of some continuous distributions
- Empirical histograms: How do they behave?



- How to understand transformations of R.V.



- Monte Carlo integration

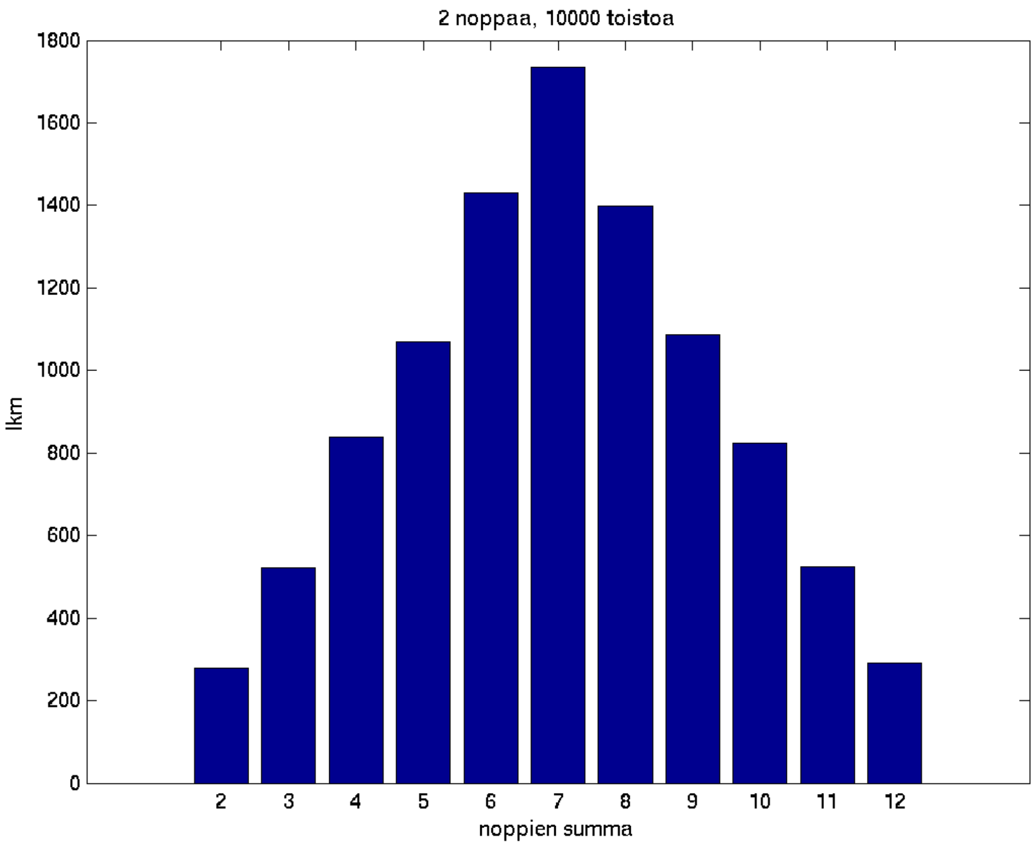


Rolling two fair dice: How is the sum distributed?

Value of the sum	Favorable outcomes (that have this particular sum)	Number of outcomes	Probability
2	”11”	1	1 / 36
3	”12”, ”21”	2	2 / 36
4	”13”, ”22”, ”31”	3	3 / 36
5	”14”, ”23”, ”32”, ”41”	4	4 / 36
6	”15”, ”24”, ”33”, ”42”, ”51”	5	5 / 36
7	”16”, ”25”, ”34”, ”43”, ”52”, ”61”	6	6 / 36
8	”26”, ”35”, ”44”, ”53”, ”62”	5	5 / 36
9	”36”, ”45”, ”54”, ”63”	4	4 / 36
10	”46”, ”55”, ”64”	3	3 / 36
11	”56”, ”65”	2	2 / 36
12	”66”	1	1 / 36
total		36	1

Rolling two fair dice: Empirical distribution of the sum

```
n = 10000;
a = dice(n);
b = dice(n);
s = a+b;
hist(s, 1:12)
```

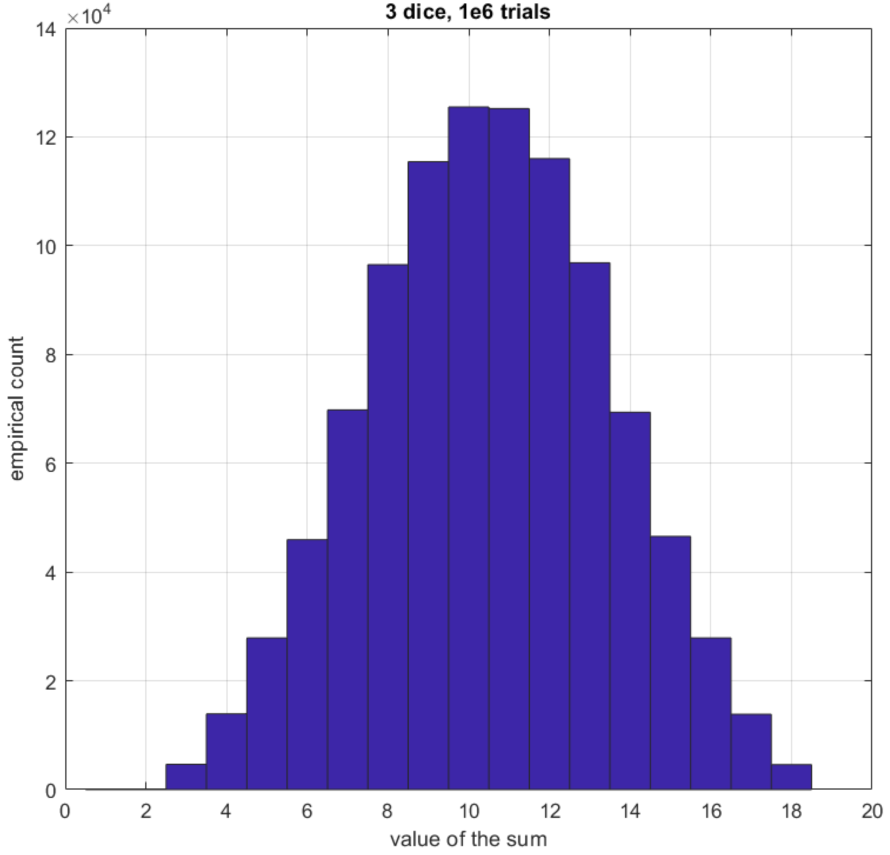


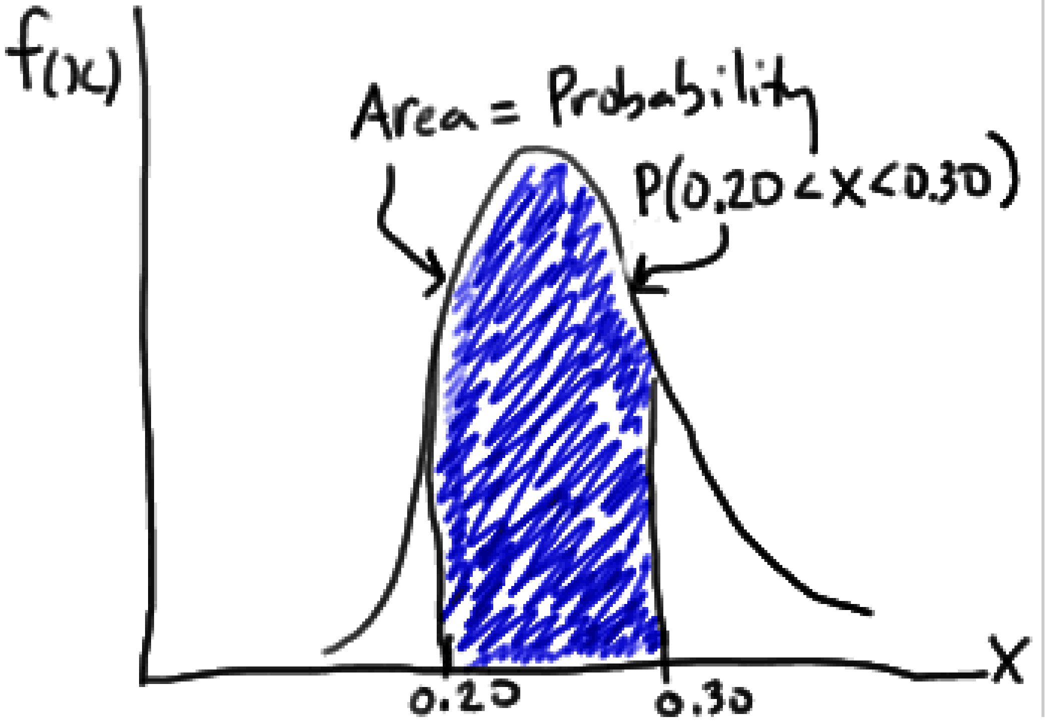
Rolling **three** fair dice: How is the sum distributed?

Value of the sum	**Favorable outcomes (that have this particular sum)**	**Number of outcomes**	**Probability**
3	”111”	1	1 / 216
4	”112”, ”121”, ”211”	3	3 / 216
5	”113”, ”122”, ”131”, ”212”, ”221”	5	5 / 216 correct?
6	”114”, …, ”231”, …, many …	…	…
7	Ouch! Getting complicated	…	…
…	…	…	…
17	”556”, ”565”, ”655”	3	3 / 216
18	”666”	1	1 / 216
total		216	216 / 216 = 1

Rolling three fair dice: Empirical distribution of the sum

```
n = 10000;
a = dice(n);
b = dice(n);
c = dice(n)
s = a+b+c;
hist(s, 1:18)
```





DEALING WITH CONTINUOUS RANDOM VARIABLES

Density function (of a <u>continuous</u> r.v.)

- If X has density function f,
  then f(x) **expresses a coefficient of proportionality**:
  prob. of X being on a short interval near x is f(x) * interval length

- Obs: f(x) **is not the probability** that X=x
- Eg. Predicted temperature, uniform distribution over [20, 30]
- Because f(x) is not a probability, it can easily be bigger than 1.
  (What does it then mean?)

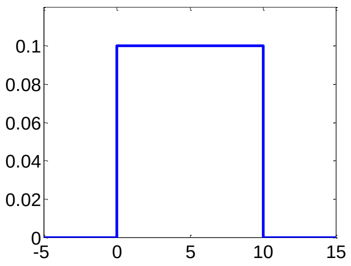
Cumulative distribution function (CDF)

- Although density is perhaps easier to grasp **intuitively and visually**, CDF is a nice tool for **calculations** with a given distribution.
- F(x) answers the question
  ”what is the probability that $X \leq x$”.
- E.g. for the temperature: F(25) = ½ means we have ½ probability for temperature at most 25.

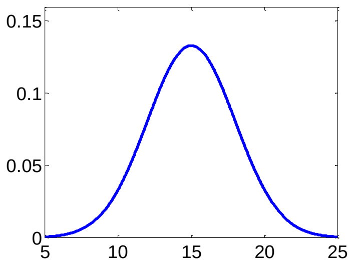
- CDF always monotonically increasing (when you move from left to right)
- CDF and density are related to each other (integral $\leftrightarrow$ derivative)

Density, CDF and mean E(X)

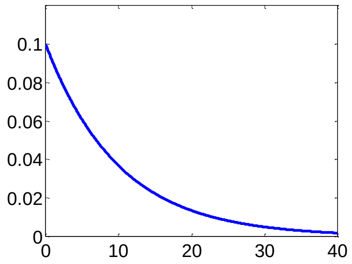
- Refresh four different continuous distributions



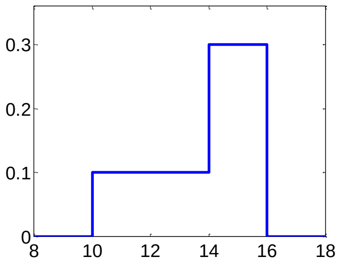
Unif(a, b)



N(μ, σ^2)



Exp(λ)

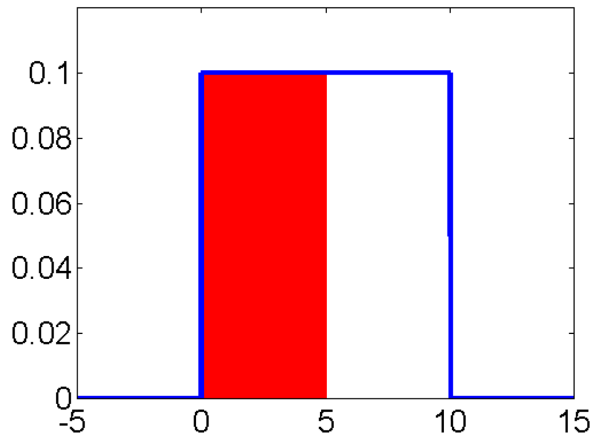


Arbitrary density

- **Probability for an interval**
  - From density: integral of density
  - From CDF: difference of values at endpoints

- **Expected value**
  - From density: integral of (density times x)

Uniform distribution



- Metro waiting time $X \sim$ **Unif(0, 10)**
- Density $f(x) = 1/10$ when $0<x<10$
- CDF $F(x) = x/10$ when $0<x<10$

Obs. F'(x)=f(x)

- Probability for an interval: integrate the density over that integral, **or just apply CDF:**

  $P(2 < X < 5) = F(5) - F(2) = 0.5 - 0.2 = 0.3$

  $P(X > 7) = 1 - F(7) = 1.0 - 0.7 = 0.3$

  $P(X < 5) = F(5) = 0.5$

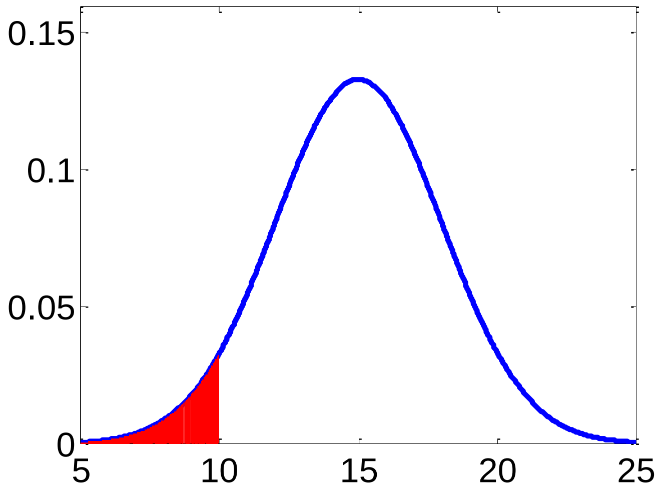
- Easy to show that the expected value equals the average of the endpoints

  $E(X) = (0+10) / 2 = 5.0$

  But surely you could do the integral:

  $$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{10} x \cdot 0.1 dx = 5.0$$

Normal distribution



- Bus travel time $X \sim N(15, 3^2)$

  Two parameters μ=15, σ=3

- Density $f(x) = c \cdot \exp(- \ldots)$
- CDF $F(x) = \Phi((x - 15) / 3)$

- Integrating the density is difficult, so use the CDF.
  Probability for an interval obtained as difference of CDF at endpoints

  $P(10 < X < 15) = F(15) - F(10) = 0.500 - 0.048 = 0.452$

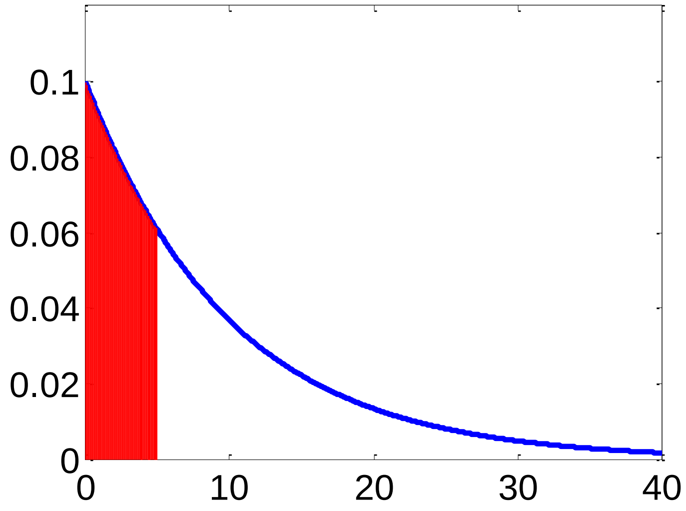
  $P(X > 20) = 1 - F(20) = 1.000 - 0.952 = 0.048$

  $P(X < 10) = F(10) = 0.048$

- Known that the expected value equals the parameter μ

  $E(X) = 15$

Exponential distribution



- Lamp lifetime $X \sim$ **Exp(0.1)**

  One parameter λ=0.1

- Density $f(x) = 0.1 \exp(-0.1x)$ when $x>0$
- CDF $F(x) = 1 - \exp(-0.1x)$ when $x>0$ Obs. $F'(x)=f(x)$

- Probabilities of intervals: either integrate the density,
  or just apply the CDF:

  $P(5 < X < 10) = F(10) - F(5) = 0.632 - 0.394 = 0.239$

  $P(X > 20) = 1 - F(20) = 1.000 - 0.865 = 0.135$

  $P(X < 5) = F(5) = 0.394$

- Known that the mean equals $1 / \lambda$

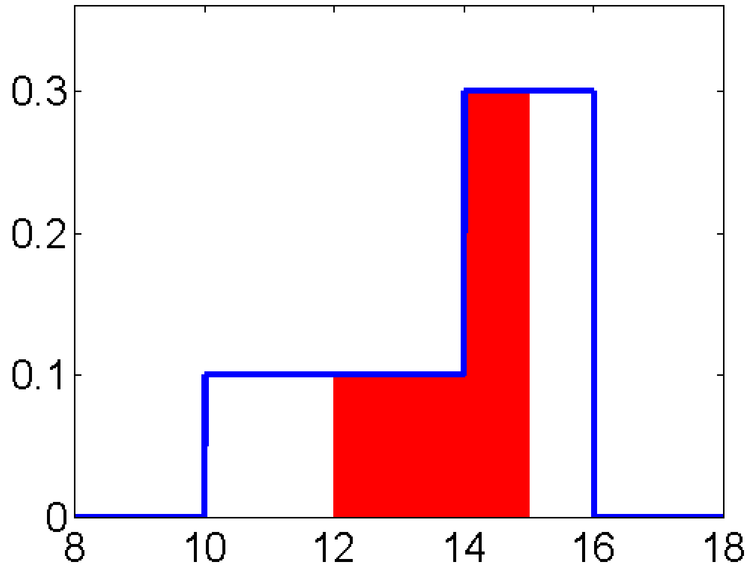
  $E(X) = 1 / 0.1 = 10$

  You could actually calculate that integral (need to know ”integration by parts”)

Arbitrary density

- Meteorologist tells you that tomorrow morrning temperature has density

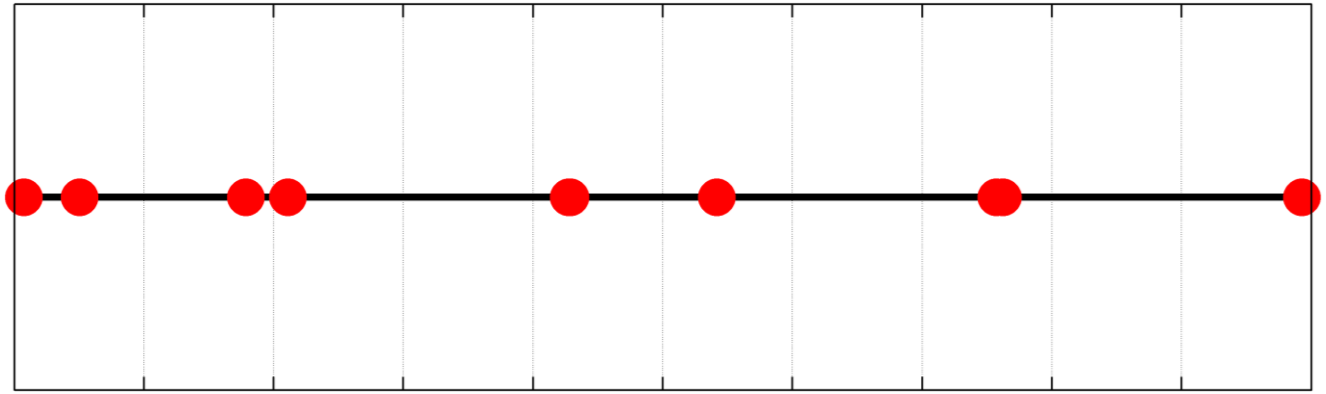
$$f(x) = \begin{cases} 0.1 & (10 \le x \le 14) \\ 0.3 & (14 \le x \le 16) \\ 0 & \text{otherwise} \end{cases}$$



- Probabilities and expected values by integrating

$$P(12 < X < 15) = \int_{12}^{15} f(x)dx = \int_{12}^{14} 0.1dx + \int_{14}^{15} 0.3dx = 0.2 + 0.3 = 0.5$$

$$E(X) = \int_{10}^{16} x \cdot f(x)dx = \int_{10}^{14} x0.1dx + \int_{14}^{16} x0.3dx = 4.8 + 9.0 = 13.8$$

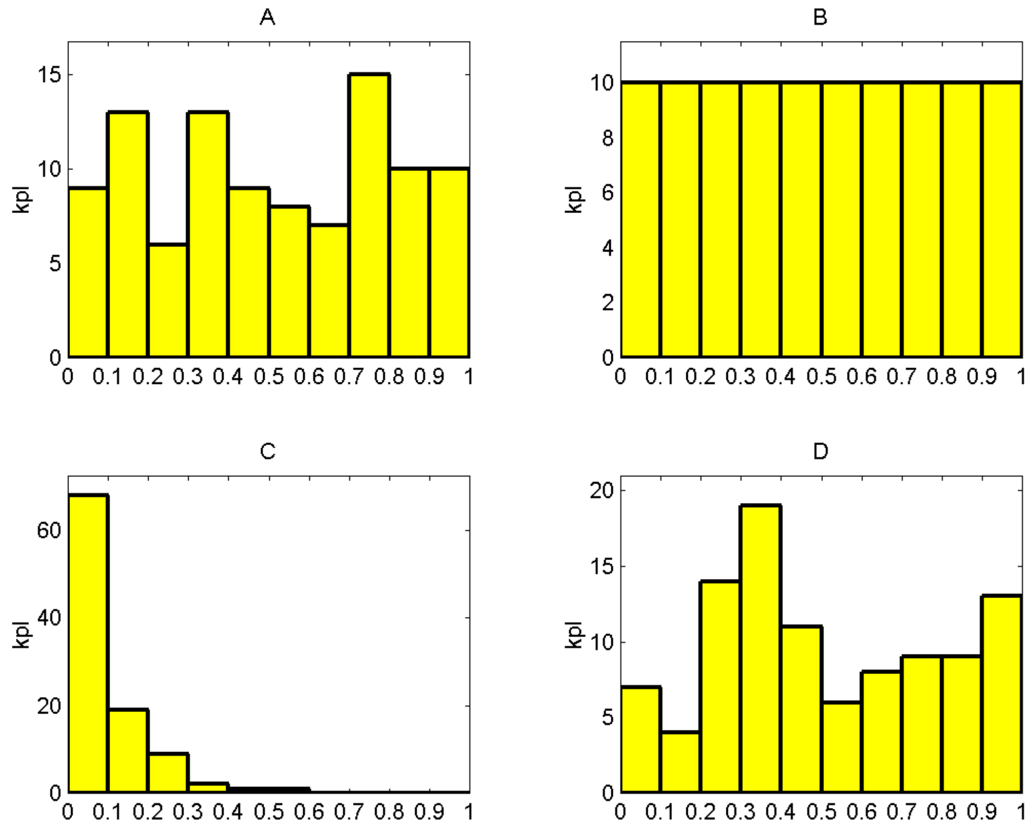


INDEPENDENT SAMPLING FROM UNIFORM DISTRIBUTION

Quiz

Each histogram here represents 100 numbers ”from some process”.

Which histogram(s) was/were made by taking
a **random sample** from the **uniform distribution** over [0, 1] ?

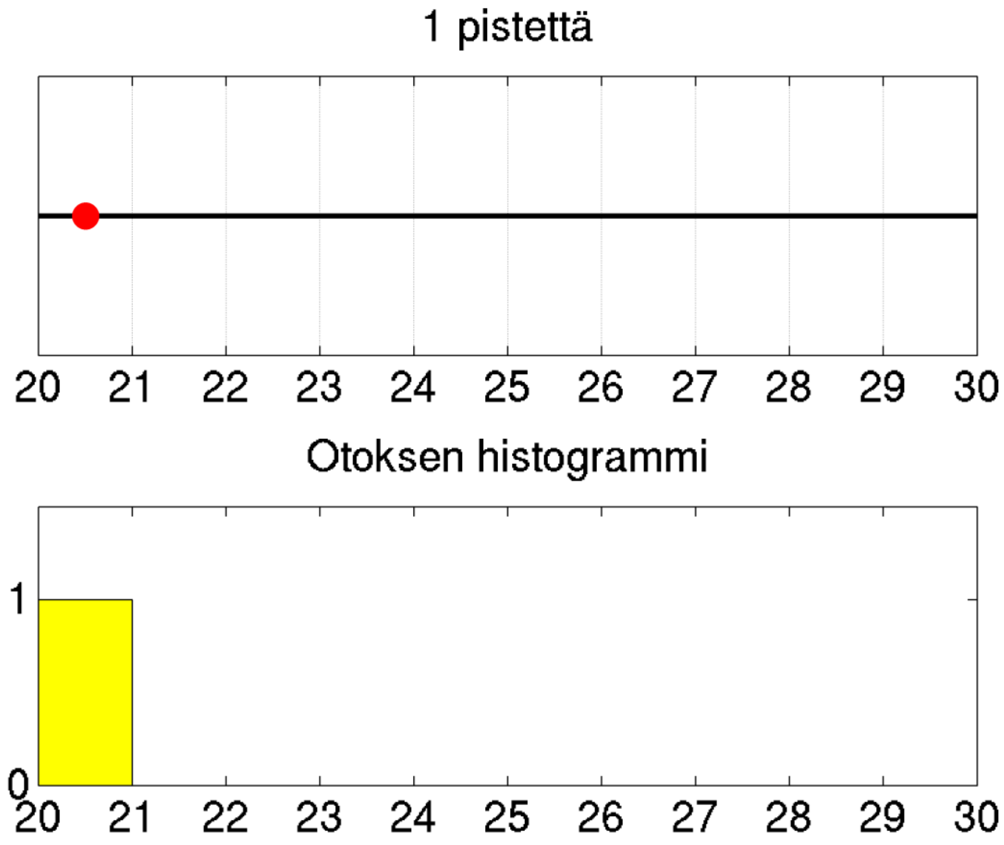


Sampling from uniform

Take independent random numbers
$X_1, X_2, X_3 \ldots \sim \text{Unif}(20, 30)$.

Where are they located?

Let us also plot a 10-bar empirical histogram.

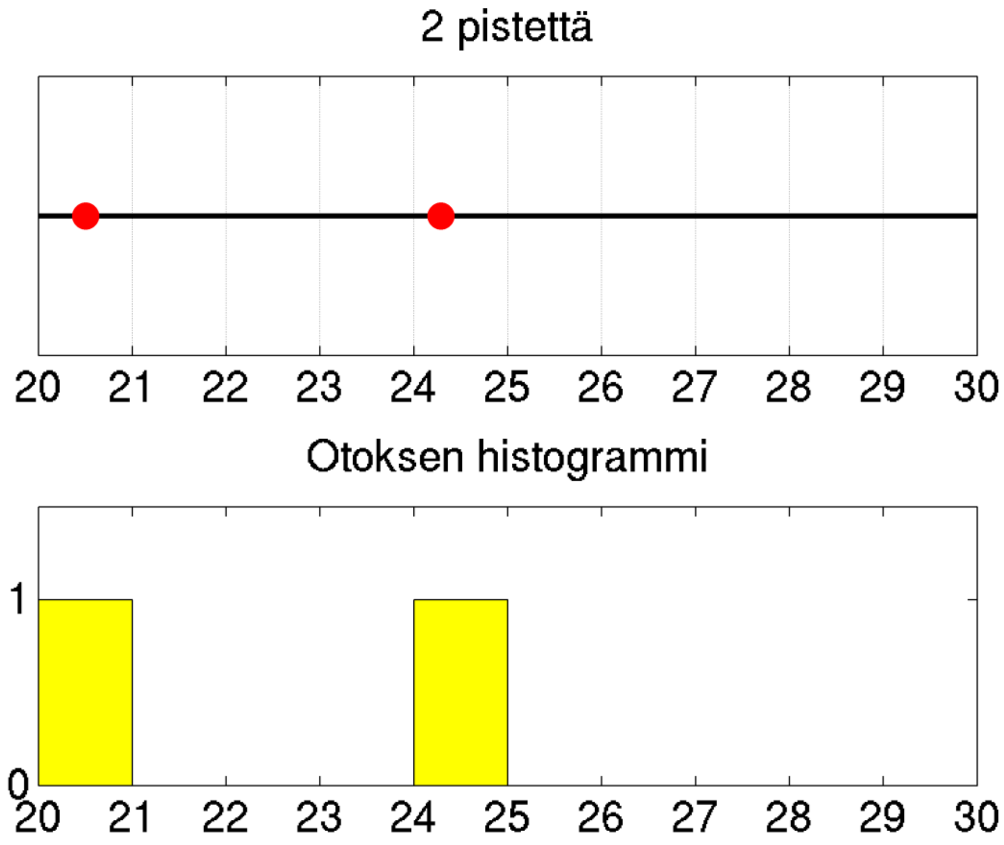


Sampling from uniform

Take independent random numbers
$X_1, X_2, X_3 \ldots \sim \mathrm{Unif}(20, 30)$.

Where are they located?

Let us also plot a 10-bar empirical histogram.

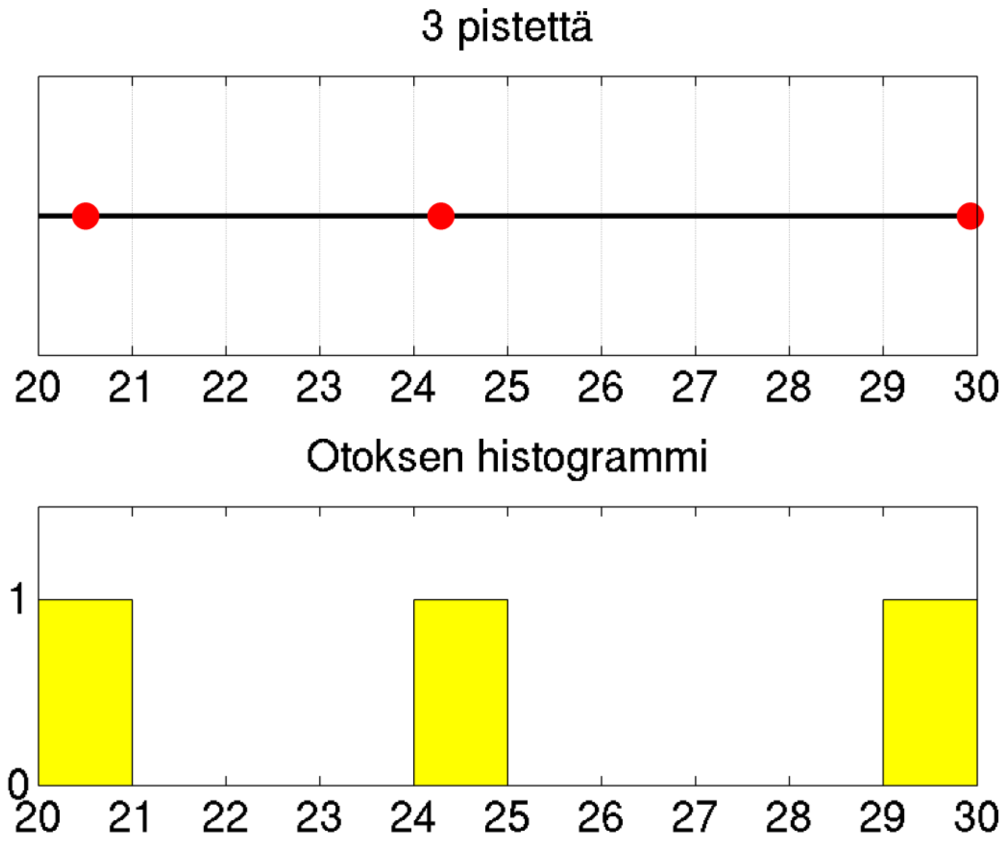


Sampling from uniform

Take independent random numbers
$X_1, X_2, X_3 \ldots \sim \text{Unif}(20, 30)$.

Where are they located?

Let us also plot a 10-bar empirical histogram.



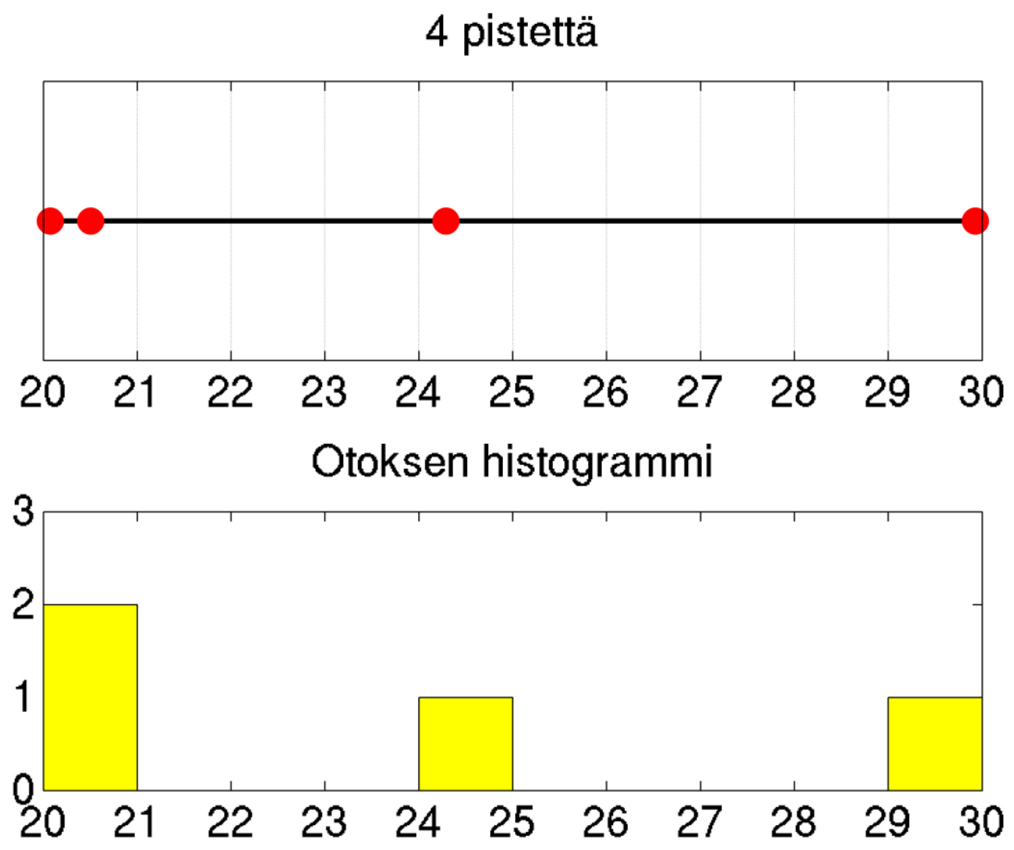
Sampling from uniform

Take independent random numbers
$X_1, X_2, X_3 \ldots \sim \text{Unif}(20, 30)$.

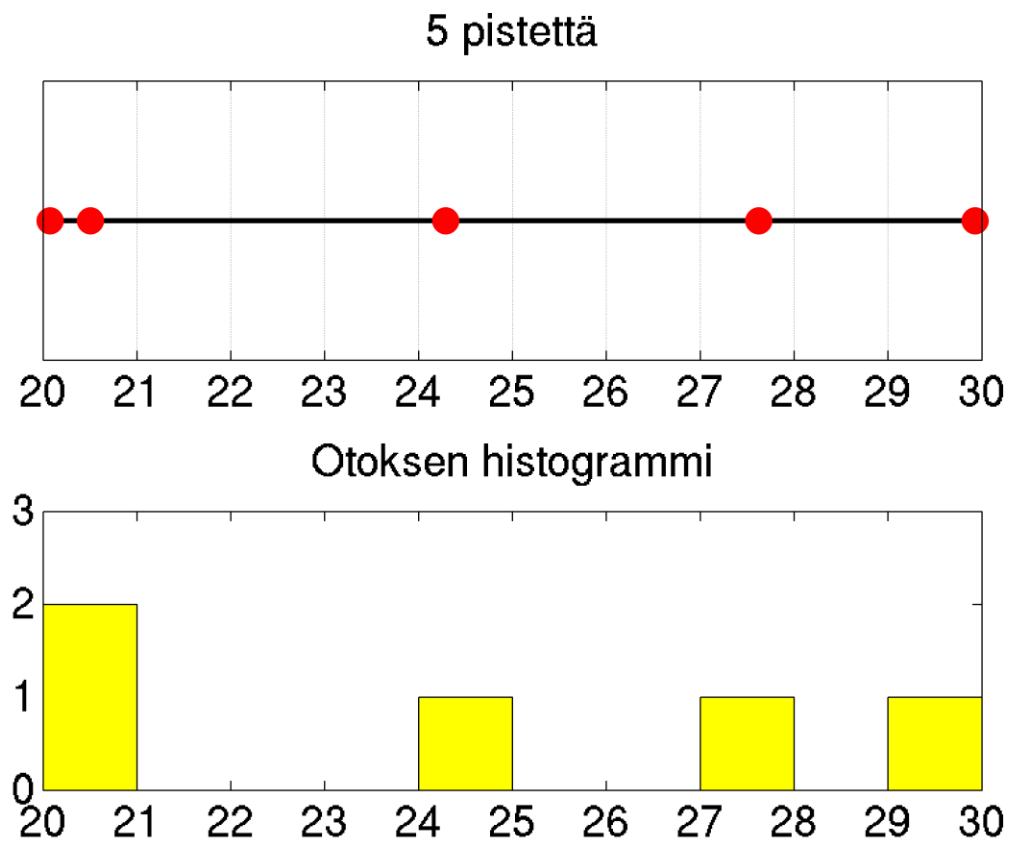
Observe independence:
Previous points do not affect later points.

1st point in (20,21)
does not prevent 4th point being there too

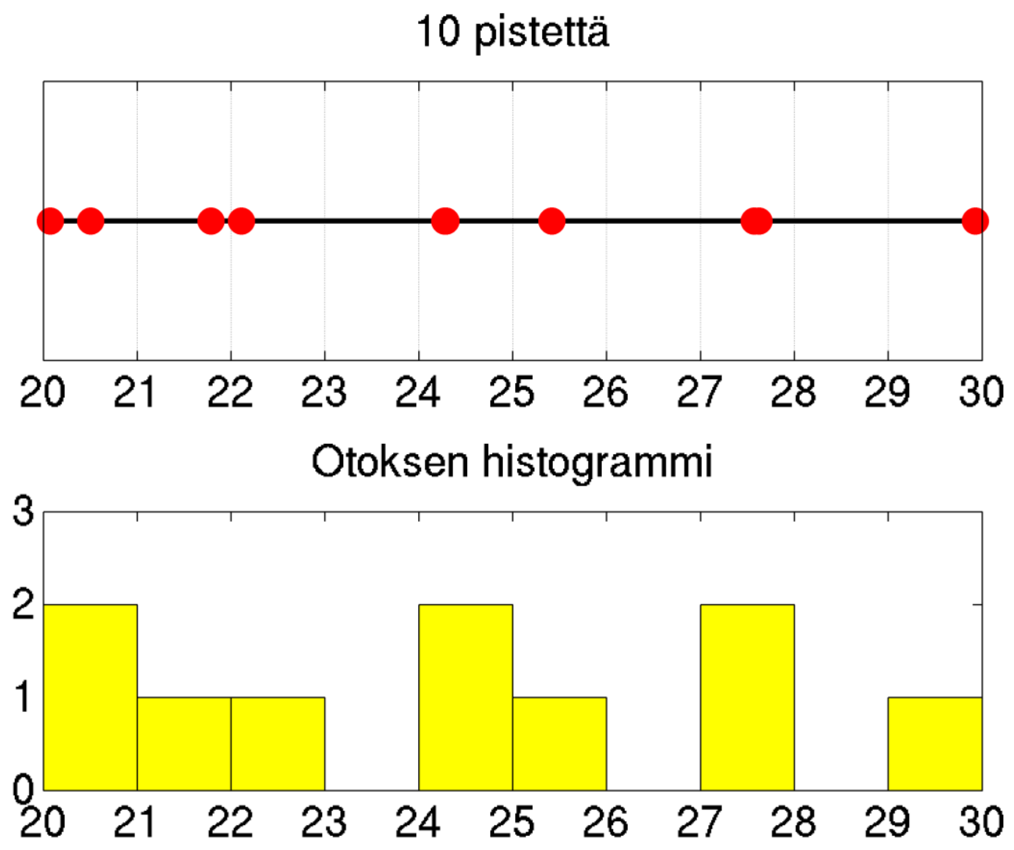
(does not even affect the probability of that event, which is still 1/10)



Sampling from uniform



Sampling from uniform

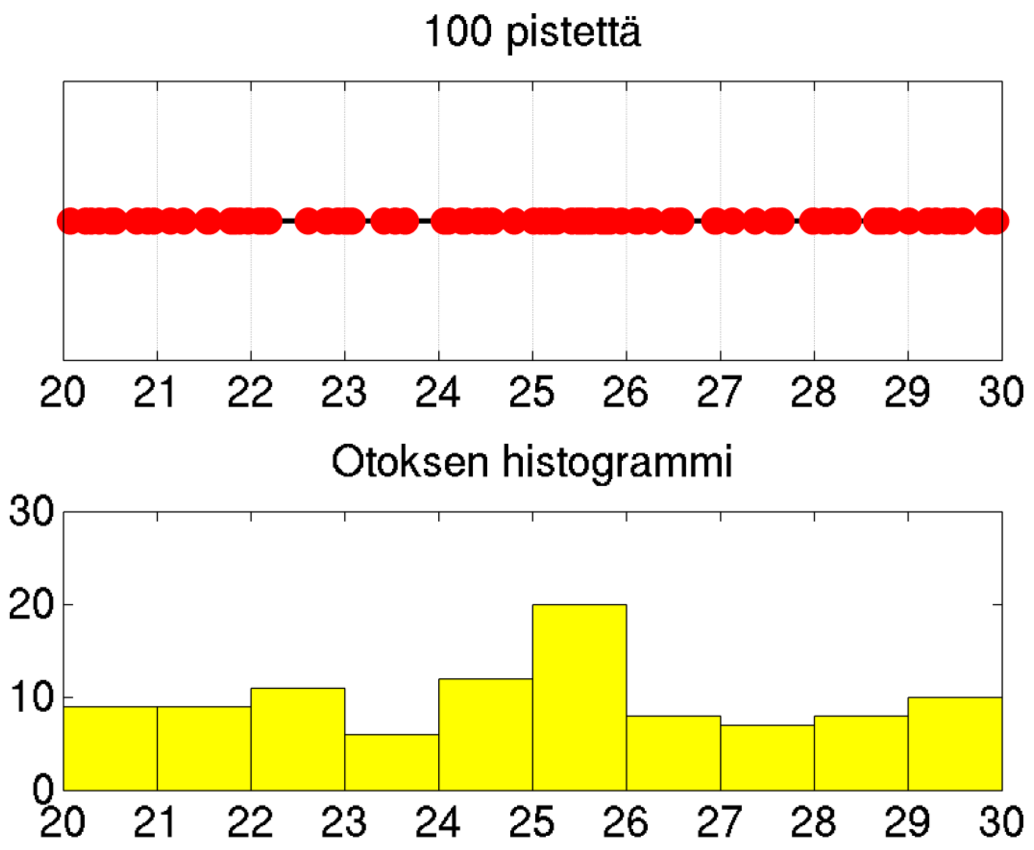


Sampling from uniform

100 pistettä:

For each bar, the
expected height
is 10,   (why?)

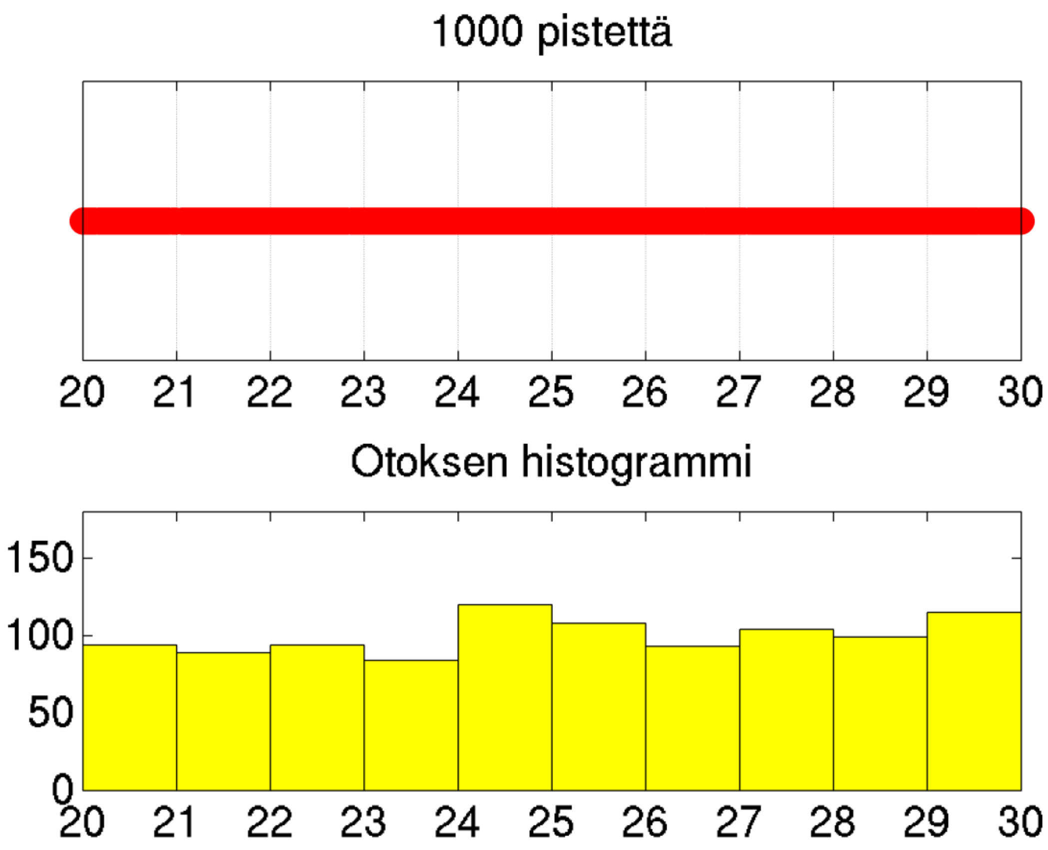
but the observed heights
have a lot of (random)
variation.



Sampling from uniform

1000 points: Difficult to even see the points.

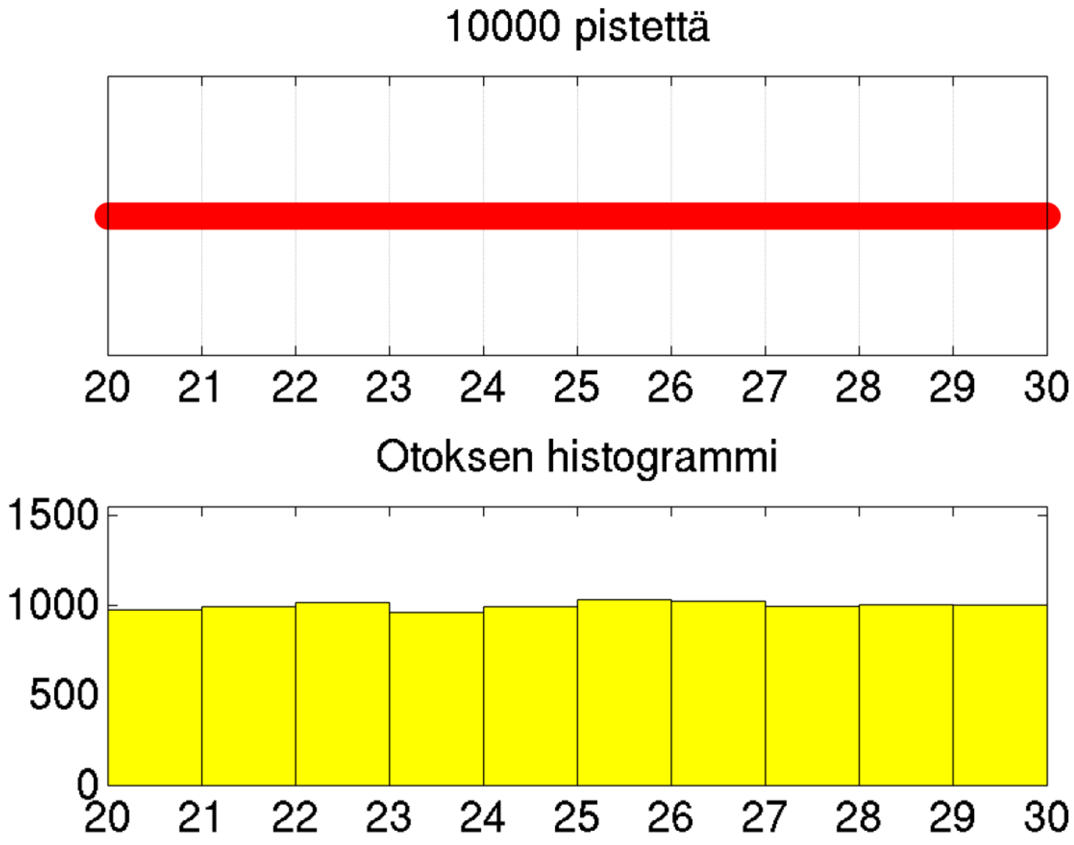
Histogram gives better rough idea of where the points are located.



Sampling from uniform

10 000 pistettä:

Fairly ”uniform”.



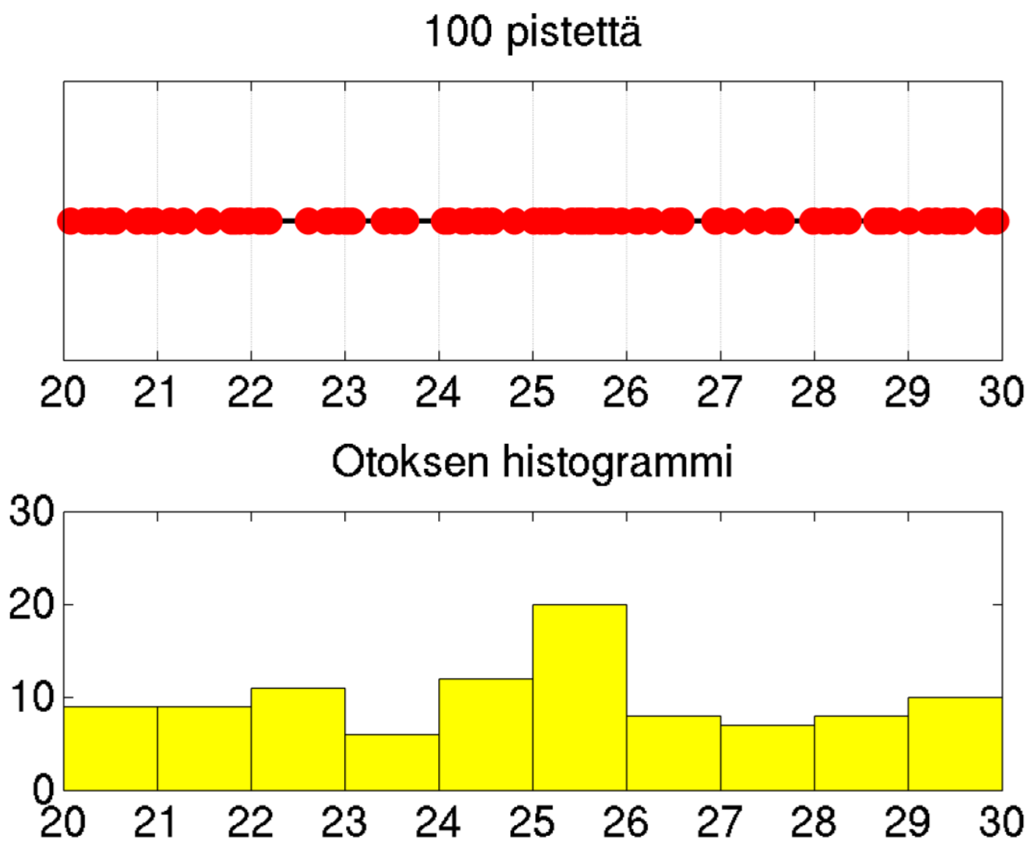
Distribution of a bar height

Let’s go back to 100 points.
Consider the **random** variable

Y_i = height of ith bar
= number of points that land in the ith interval

We can see that Y_i has a **binomial distribution. (why?)**

Next question: Probability of all bars equal?
Difficulty: Bar heights not independent.



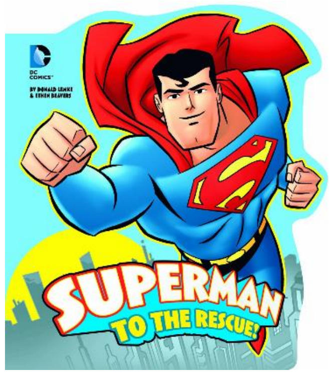
Distribution of a bar height

Let's go back to 100 points.
Consider the **random** variable

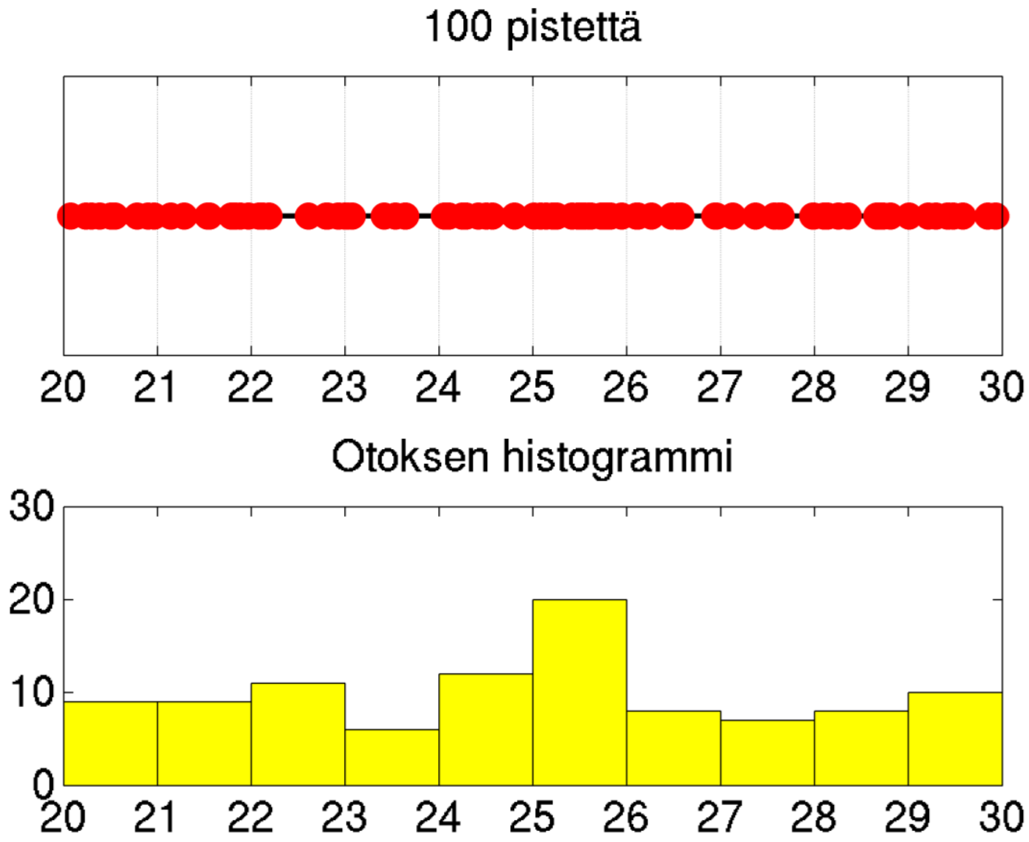
Y_i = height of ith bar
= number of points that land
in the ith interval

We can see that Y_i has a
binomial distribution. (why?)

Next question: Probability of all bars equal?
Difficulty: Bar heights not independent.



Multinomial distribution
to the rescue!



Multinomial distribution

- *n* independent trials, each has (e.g.) 3 exclusive possible outcomes.
- In each trial the outcomes have probabilities *p*, *q*, *r*.
- Probability that the outcomes have counts (*a*, *b*, *c*) [recall Lecture5B]

$$\binom{n}{a,b,c} \cdot p^a \cdot q^b \cdot r^c$$

- The count vector (*a*, *b*, *c*) has **multinomial distribution** with params *n* and (*p*, *q*, *r*).
- The three counts are random and **dependent from each other**
- For example, if *a*=*n*, then necessarily *b*=*c*=0. (why?)
- If more than 3 possibilities, it generalizes in the obvious way.
- If only 2 possibilities, you get back your familiar binomial distribution.

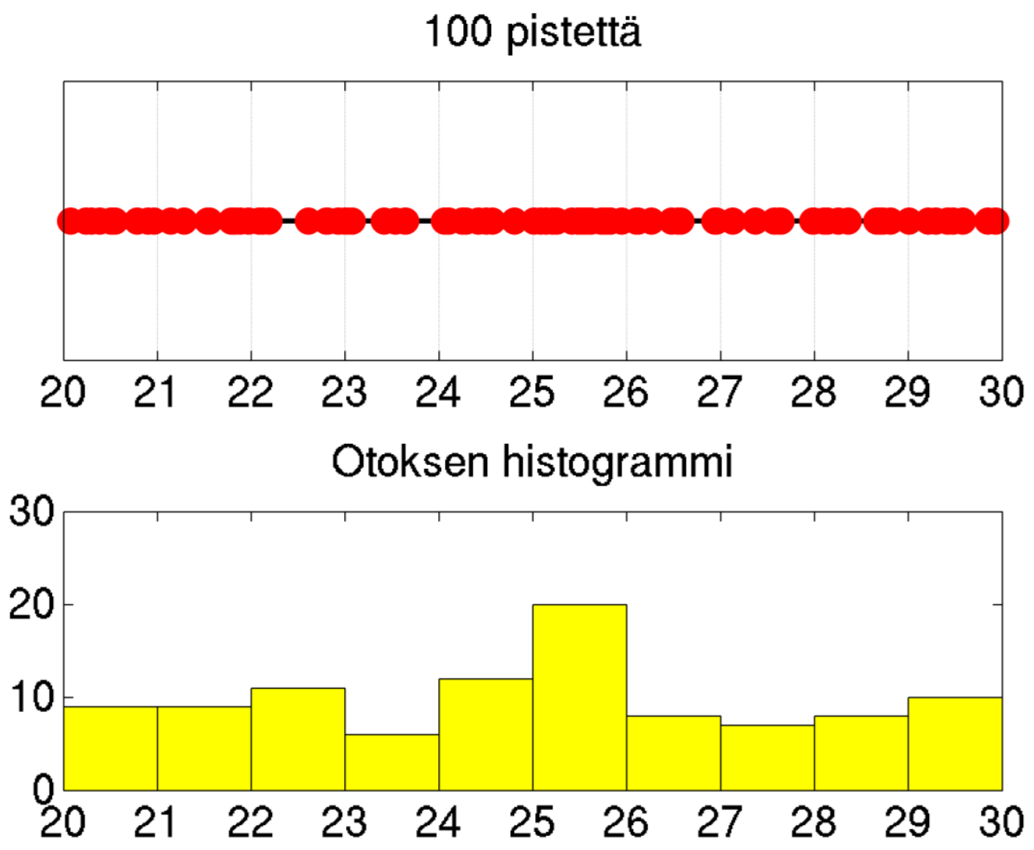
Joint distribution of the 10 bar heights

Let's go back to 100 points.
Consider the **random** variable

Y_i = height of ith bar
= number of points that land
in the ith interval

The random vector (Y_1 , Y_2 , Y_3 , …, Y_{10})
has the multinomial distribution
with parameters

n = 10
p_1 , p_2 , p_3 , …, p_{10} = 0.1



Calculate: 100 points, 10 bars

- To have ”all equal bars” we need counts exactly (10,10,10,…,10).
- That has probability =

$$\binom{100}{10,10,\ldots,10} \cdot 0.1^{10} \cdot 0.1^{10} \cdot \ \ldots \cdot 0.1^{10} \quad \approx \quad 2.4 \cdot 10^{-8}$$

- which is pretty small.

- Next question: What is the probability that every bar is ”near” the expected value, say, in interval [7, 13] ?
- Perhaps this has a much larger probability?

Calculating probability of ”nearly” equal bars

- Probability that each of the 10 bars has height in [7, 13] ?

- One method (EXACT): Consider each such count vector separately, for example (9,8,11,7,12,10,10,12,8,13) and many others.
- For each vector, compute the multinomial probability, and add up.

- Not nice manually. Can be done by computer (code will be on course page)
- We get that the ten bars of the histogram are
  - … all equal (10) with probability 2.3571e-08
  - … all in [9,11] with probability 1.5528e-04
  - … all in[8,12] with probability 0.0083
  - … all in [7,13] with probability 0.0747

- Second method (EMPIRICAL APPROXIMATION): Make many such (random) empirical histograms, and check HOW MANY of them have ”nearly equal bars”

Food for thought

- Sample n=100 points from uniform over [20,30]
- Make a 10-bar histogram

- Probability of heights (10,10,10,10,10,10,10,10,10,10) is 2.3571e-08
- Probability of heights (9,8,11,7,12,10,10,12,8,13) is 4.1796e-09

- All-equal bars has **bigger** probability (in the random process).
- Why does it still arouse **more suspicions** that it might not have come from such a random process?

Food for thought

- Sample n=100 points from uniform over [20,30]
- Make a 10-bar histogram

- Probability of heights (10,10,10,10,10,10,10,10,10,10) is 2.3571e-08
- Probability of heights (9,8,11,7,12,10,10,12,8,13) is 4.1796e-09

- All-equal bars has **bigger** probability (in the random process).
- Why does it still arouse **more suspicions** that it might not have come from such a random process?

- Hint:
  Are there any **other processes** how the result might have been obtained?
  Is it **possible** that such a process was being used?
  How probable is that? Could you apply **Bayesian inference** here?

Transformations of random variables

- There are many kinds of transformations (functions) you could apply to a RV.
- If X = tomorrows rainfall (in mm), perhaps we are interested in the water volume raining on a 1000 square-meter plot of land, in liters **$Y = 1000 \cdot X$**,
which is a new random variable.
  - (but not independent from X; it is **completely** dependent because value of X determines value of Y)

- The general way of finding out the distribution of any transformation:
  P(Y is something) = P(X is such that Y is something)
- Sometimes the CDF is very handy for this.

Cumulative distribution function (CDF)

- Although density is perhaps easier to grasp **intuitively and visually**, CDF is a nice tool for **calculations** with a given distribution.
- F(x) answers the question
  ”what is the probability that $X \leq x$”.
- E.g. for the temperature: F(25) = ½ means we have ½ probability for temperature at most 25.

- CDF always monotonically increasing (when you move from left to right)
- CDF and density are related to each other (integral $\leftrightarrow$ derivative)

Some typical transformation tasks

1. We know the **CDF** F_X and the transformation function $Y = g(X)$.
   Then what is the CDF of Y?
   → In principle this is **easy**: Solve $P(Y \leq y)$

2. We know the **density** f_X and the transformation function $Y = g(X)$.
   Then what is the density of Y?
   → **Not quite easy**.
   One method is to go through the CDF route:
   density of X → CDF of X → CDF of Y → density of Y

3. We have a method for **taking random numbers** $X_1 \ldots X_n$ iid from distribution of X,
   and we know the transformation function $Y = g(X)$.
   How do we take random numbers from the distribution of Y?

   → **Extremely easy**: apply the transformation: $Y_i = g(X_i)$

We illustrate some simple transformations with method 3.

Transformation: Adding a constant

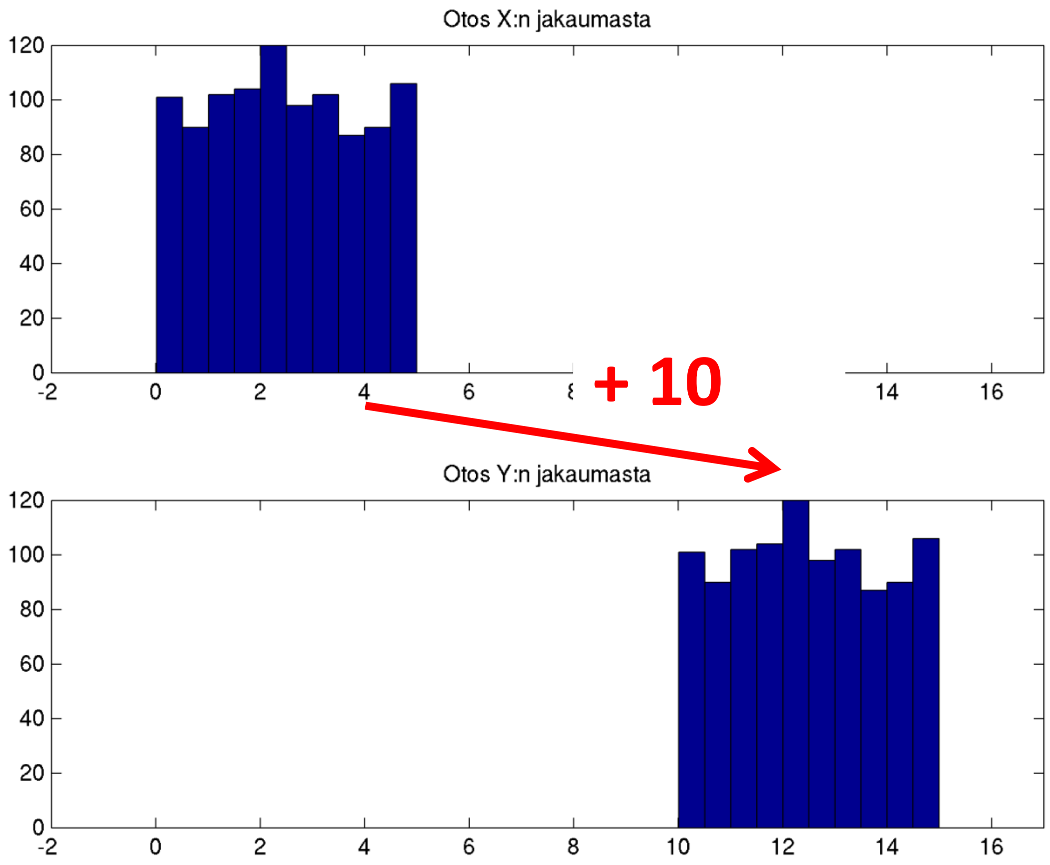
$X \sim \mathrm{Unif}(0, 5)$

$Y = X \mathbf{+ 10}$

$Y \sim$?

Apparently Y is uniform too.

With what parameters?



Transformation: Multiplying by constant

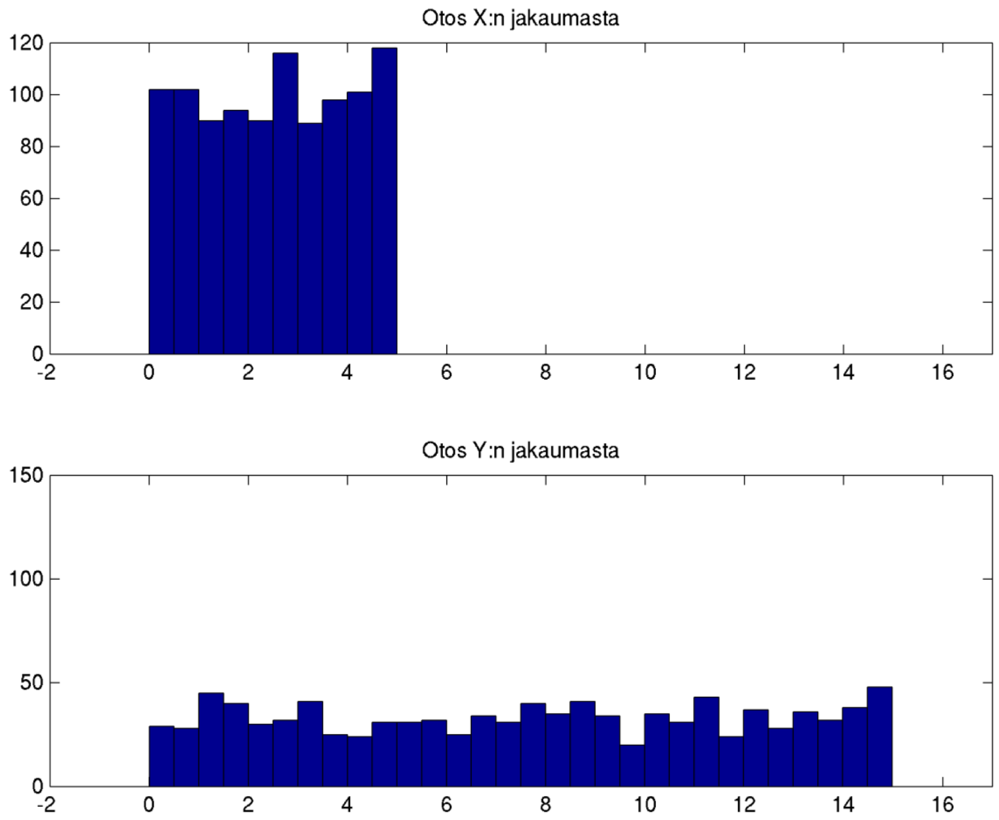
$X \sim \mathrm{Unif}(0, 5)$
$Y = \mathbf{3} \cdot X$

$Y \sim\ ?$

Apparently Y is
also uniform.

What parameters?

Observe the ”flattening”,
you can actually see this
in the formula of the density.

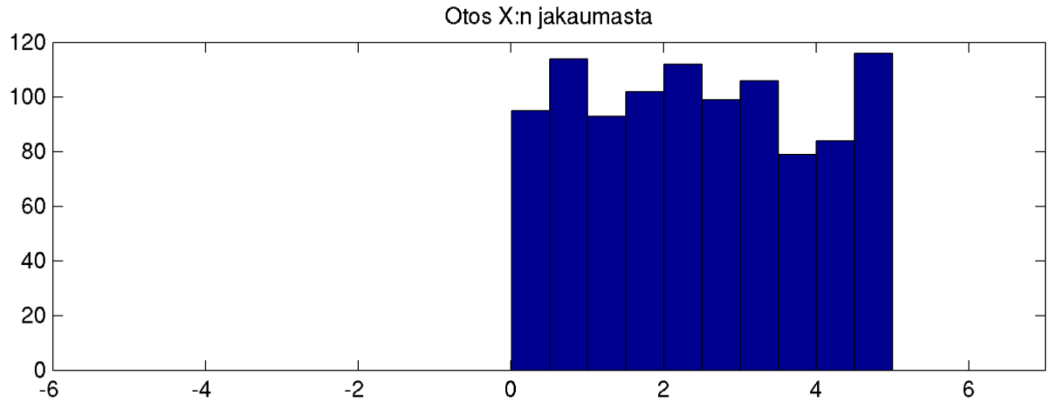


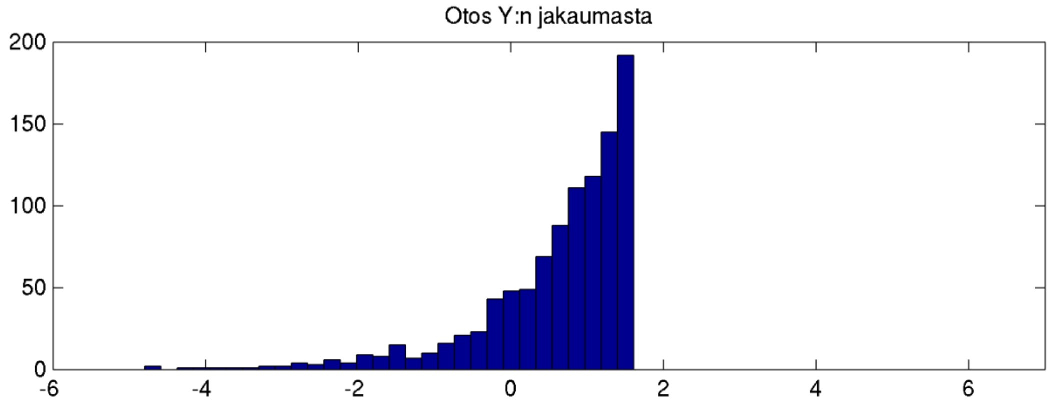
Tranformation: Natural logarithm

X ~ Unif(0, 5)
Y = **ln** X

Y ~ ?

It seems that Y has some more exotic distribution?





Solving the CDF for transformation

$X \sim$ Unif(0, 5)
$Y = \ln X$
$Y \sim$?

Solve the CDF of Y by elementary probability calculus.

$F_Y(a)$	$= P(Y \leq a)$	Definition of CDF.
	$= P(\ln X \leq a)$	Because $Y = \ln X$.
	$= P(X \leq \exp(a))$	exp is an order-preserving function.
	$= F_X(\exp(a))$	Definition of CDF
	$= \exp(a) / 5$	The CDF of a uniformly distributed X.

Then take the derivative to obtain the density function.
$f_Y(a) = (1/5) \exp(a)$

This is in the interval $(-\infty, \ln 5)$, because that's where the original interval (0, 5) goes when taking the logarithm!
→ Analytical expression that seems to match the empirical histogram!

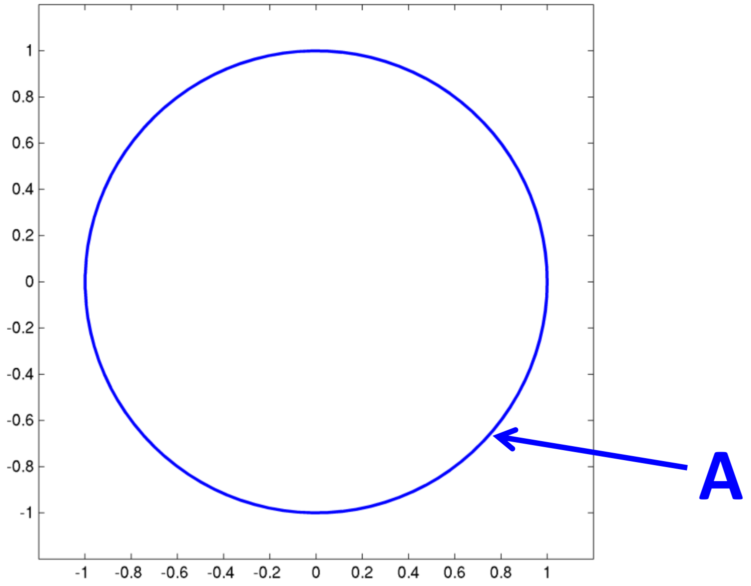
An application of the LLN:

Monte Carlo integration

Original problem, contains no probabilities

What is the area of this very complicated plane figure **A**?

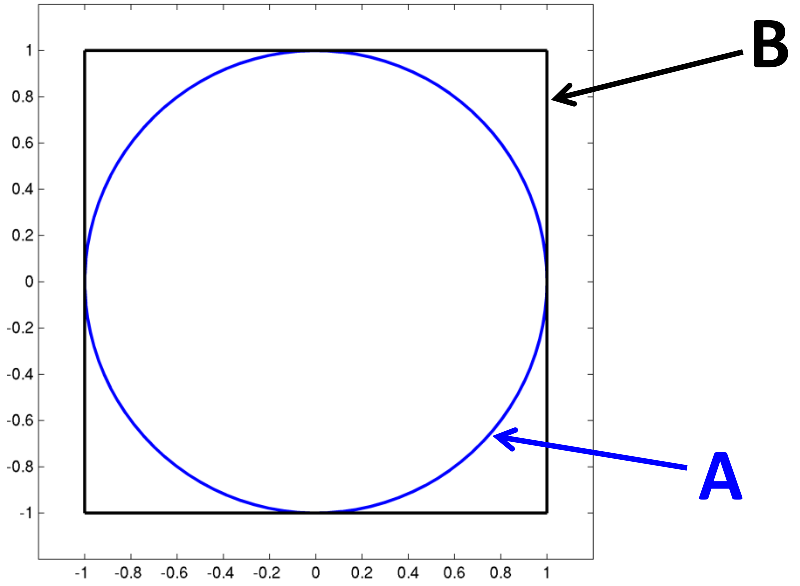
Suppose the only tool we have is a ”black box” way of **testing** if a point is **inside** A or not.



(inside the black box is a test whether $\mathbf{sqrt(}\boldsymbol{x}^{\mathbf{2}} \mathbf{+} \boldsymbol{y}^{\mathbf{2}}\mathbf{) < 1}$)

Change the problem, introduce probability

- What is the area of the complicated figure **A**? The only thing we know is whether a given point is in or out.
- Solution: We envelope A inside a bigger figure **B**,
  - whose area (= 4) we know

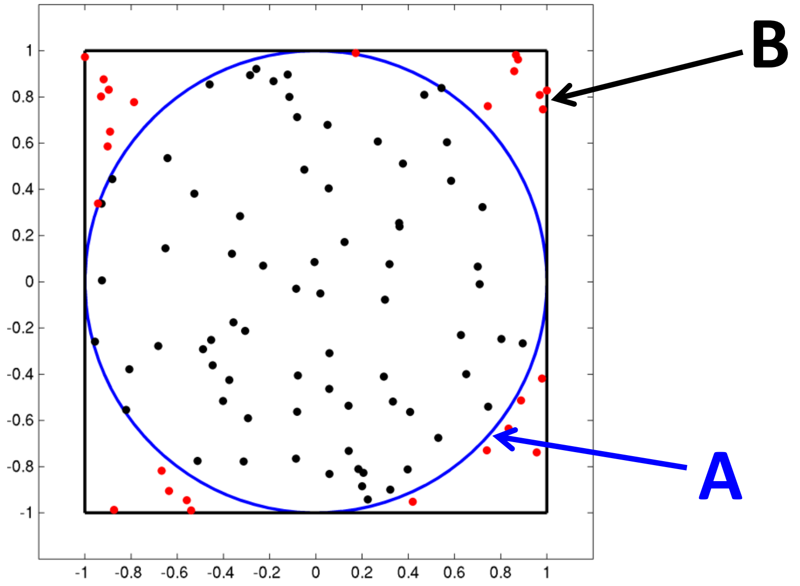


Change the problem, introduce probability

What is the area of the complicated figure **A**? The only thing we know is whether a given point is in or out.

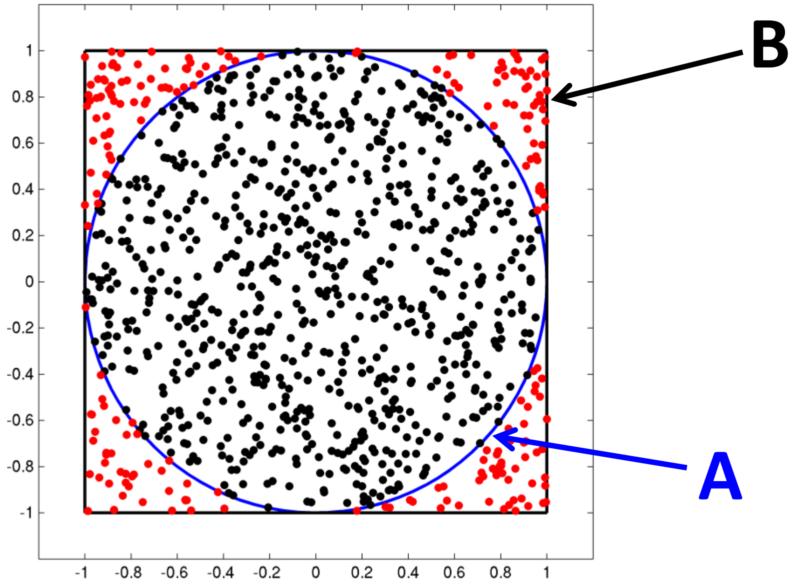
Solution: We envelope A inside a bigger figure **B**,

- whose area (= 4) we know
- and such that we can easily pick random points uniformly in B



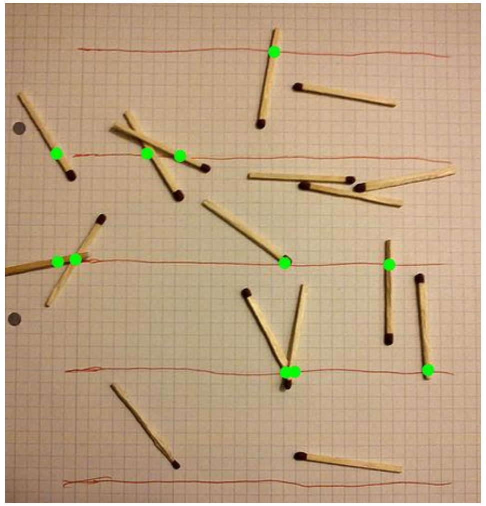
Monte Carlo integration

- Random point inside with probability $p = m(A) / m(B)$, m = area
- Repeat n times
- Law of large numbers says relative freq $f_n \approx p$
- Thus estimate $\boldsymbol{m(A)} = p\, m(B) \approx \boldsymbol{f_n\, m(B)}$



Monte Carlo integration

Similar method in ”Buffon’s needle”
https://en.wikipedia.org/wiki/Buffon%27s_needle_problem



$\pi \approx 3.1$

n	points in *A*	*m*(*A*) ≈
100	80	3.200000
1 000	783	3.132000
10 000	7 849	3.139600
100 000	78 544	3.141760
1 000 000	785 132	3.140528

Same method can be applied to **more complicated shapes**, also in higher-dimensional spaces.
E.g. what is the ”volume” of a *d*-dimensional hyperball?

Such methods are commonly applied in modern science when you cannot
(or do not want to) solve a complicated integral analytically.

Main pitfall is accuracy: obtaining 1 more decimal requires 100 * more repetitions.