Section 9: Uniform distribution, CDFs and sampling

STOR 155.02, Spring '21

updated 2021-03-21

What you will learn

- continuous distributions: uniform, normal
- CDFs and quantile functions
- sampling any r.v. using the quantile function and uniform r.v.

Resources

- Textbook: ch 1.7, 3.5, 4.1

Continuous distributions

So far we've seen discrete probability distributions

'discrete' because

$$P(X = x_i) > 0$$

for distinct outcomes $x_1 \ldots$

which are often whole numbers

examples

$$X \text{ is Bernoulli(p)}, \quad x_1 = 1, x_2 = 0$$

$$X \text{ is Binomial(N, p)},$$

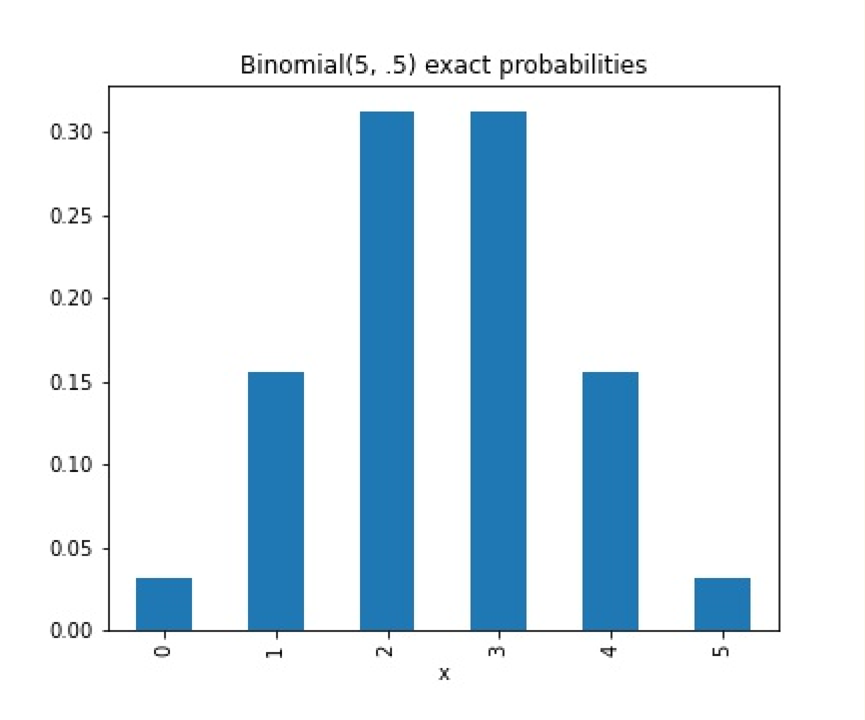
$$x_1 = 0, \ldots x_{N+1} = N$$

$$X \text{ is Geometric(p)},$$

$$x_1 = 1, x_2 = 2, \ldots$$

Imagine this on a bar chart, with height as probability

$$P(X = k) = \binom{5}{k} p^k (1-p)^{5-k}, \quad k = 0 \ldots 5$$



Continuous random variables

key facts

takes a *range* of possible values

$$\text{e.g.,} \quad 0 \leq X \leq 1$$

$$\text{e.g.,} \quad -\infty < X < \infty$$

no gaps between values

$$P(X = x) = 0$$

for any particular value x

instead have a **probability density function** $p(x)$

$$\frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \approx p(x)$$

for a small value Δx.

$$P(X \leq x)$$

$$= \text{area under the curve p(y) for } y \leq x$$

$$P(a \leq X \leq b) = \int_a^b p(x) dx$$

if you remember calculus from high school.

Uniform distribution

mother of all

Uniform(0, 1) distribution

$p(x) = 1 \quad \text{for } 0 \leq x \leq 1$ and 0 for all other x.

Expectation, variance

$$E(X) = 1/2$$

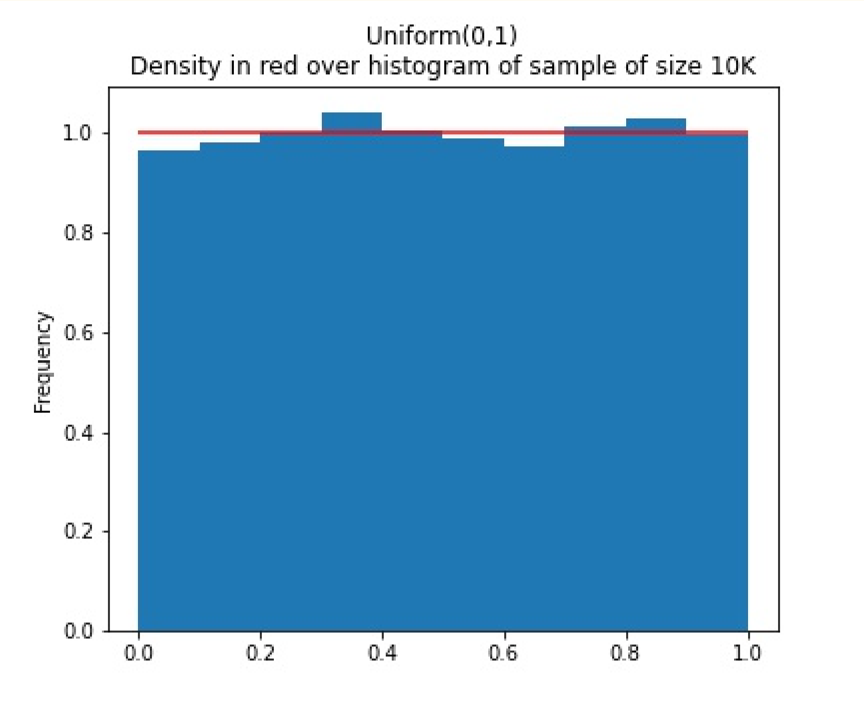
$$Var(X) = 1/12$$

for those who have taken calculus

$$E(X) = \int_0^1 x dx, \quad Var(X) = \int_0^1 x^2 dx - \left(\int_0^1 x dx\right)^2$$

Exact density vs. histogram

Q: Why does it make sense that $E(X) = 1/2$?



Uniform(a, b) for $a < b$

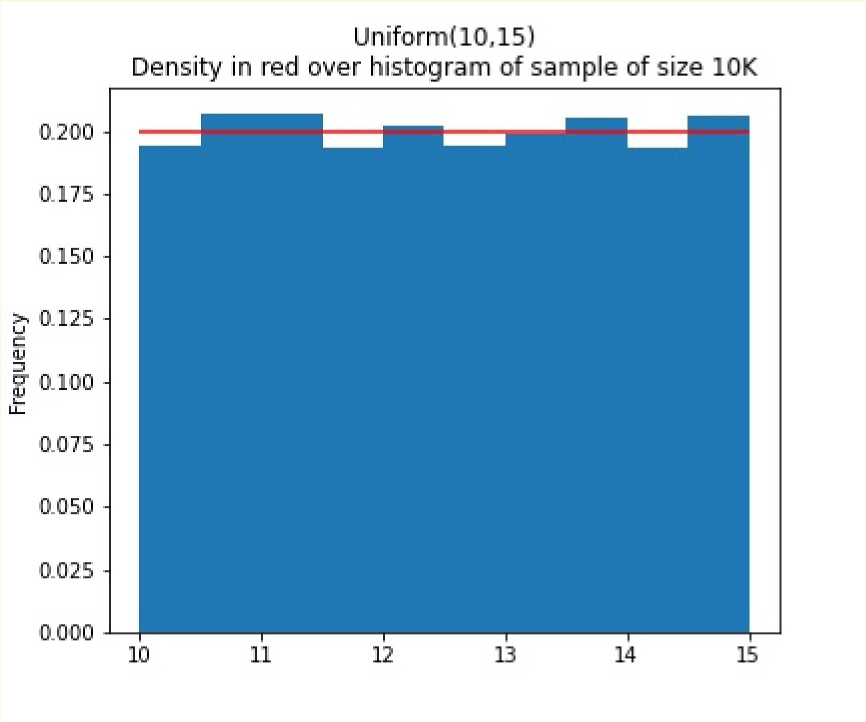
Density is a horizontal line between a and b.

$p(x) = \frac{1}{b-a}$ for $a \leq x \leq b$ and 0 for all other x.

Expectation, variance

$$E(X) = (a+b)/2 = \text{ midpoint between a, b}$$

$$Var(X) = (b-a)^2/12$$



Uniform(10,15)
Density in red over histogram of sample of size 10K
Frequency
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
10
11
12
13
14
15

Definition: Cumulative distribution function (CDF)

$$F_X(x) = P(X \leq x) = \text{CDF of X at } x$$

Discrete distributions

$$F_X(x_K) = \sum_{i=1}^{K} P(X = x_i)$$

sum of probabilities for possible outcomes $x_1 \leq x_2 \ldots \leq x_K$ where x_K is one of the possible outcomes of X.

To calculate $F_X(x)$ for x **between** possible values, use the next possible value below x.

Continuous distributions

Area under the curve given by density function $p(\cdot)$ to the left of x

CDF of Uniform(0, 1)

$$F_X(x) = P(X \leq x) = x$$

since this is the area of the rectangle of height 1, width x.

CDF of Uniform(a, b)

$$F_X(x) = \frac{x-a}{b-a}$$

CDF of Binomial(10, 1/2)

$$F_X(4) = \sum_{i=1}^{4} \binom{10}{i} p^i (1-p)^{10-i}$$

Use a computer!

Calculating CDFs by hand is either tedious or not really possible

Definition: Inverse CDF, aka quantile function

$$F_X^{-1}(u) = \text{uth theoretical quantile of distribution for X}$$

Remember

When looking at data, the uth quantile is the

- value of a variable
- so that u proportion of observations
- are less than or equal to that value.

Same idea here

Sample from any distribution!

X is some random variable with CDF F_X.

$$U \sim \text{Uniform}(0, 1)$$

then

$$F_X^{-1}(U) \text{ has the same distribution as } X$$

how to use this

Sample independent Uniform(0, 1) random variables $U_1 \ldots U_N$, then

$$F_X^{-1}(U_1), \ldots F_X^{-1}(U_N) \text{ gives N independent samples of } X$$

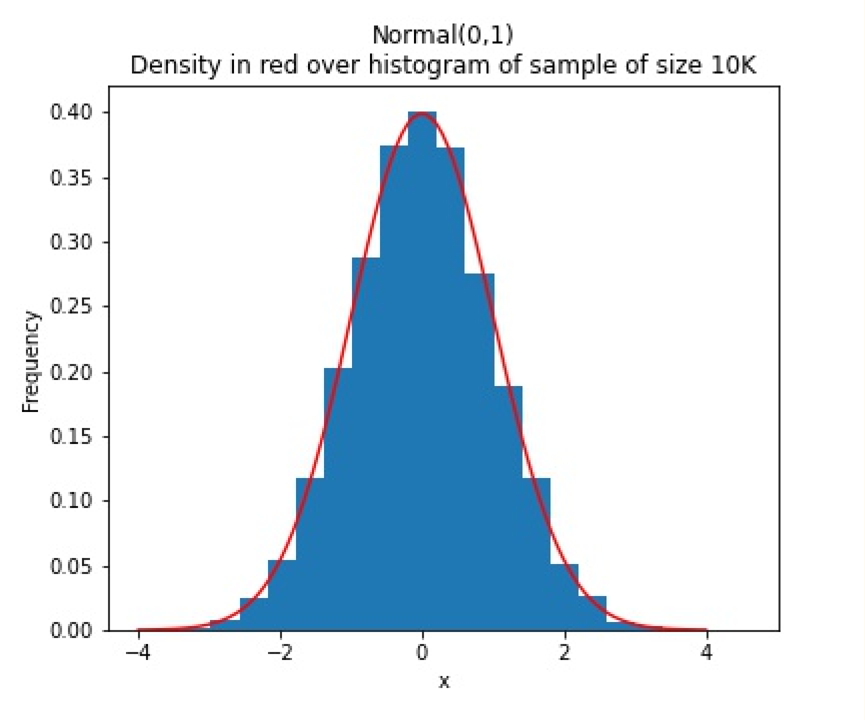
This is the best (only?) way to sample random variables in Excel

Normal distribution

Normal(0, 1) distribution

The famous **bell curve**

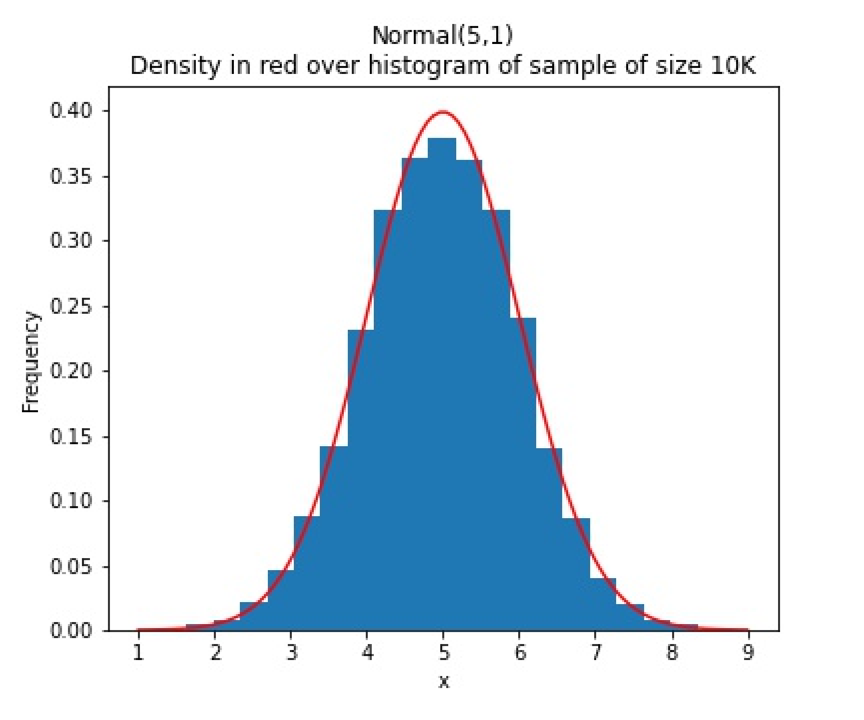
$$E(X) = 0, \qquad \sqrt{Var(X)} = SD(X) = 1, \qquad -\infty < X < \infty$$

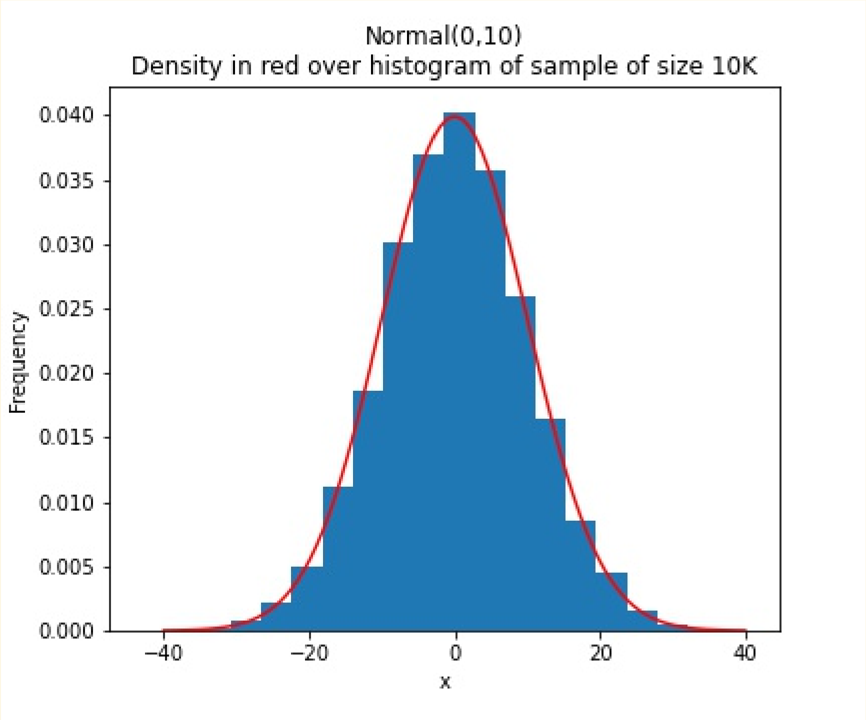


Normal(m, s)

The **mean shifts the middle of the bell** and the **standard deviation determines how wide the bell is**.

$$E(X) = m, \qquad SD(X) = s, \qquad -\infty < X < \infty$$




Normal(0,10)
Density in red over histogram of sample of size 10K
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Frequency
−40
−20
0
20
40
x

Convert Normal(m, s) to Normal(0, 1)

If X is Normal(m, s) then

$$Z = \frac{X - m}{s} \text{ is Normal}(0, 1)$$

This is called a 'z-score'

these will come back later in hypothesis testing

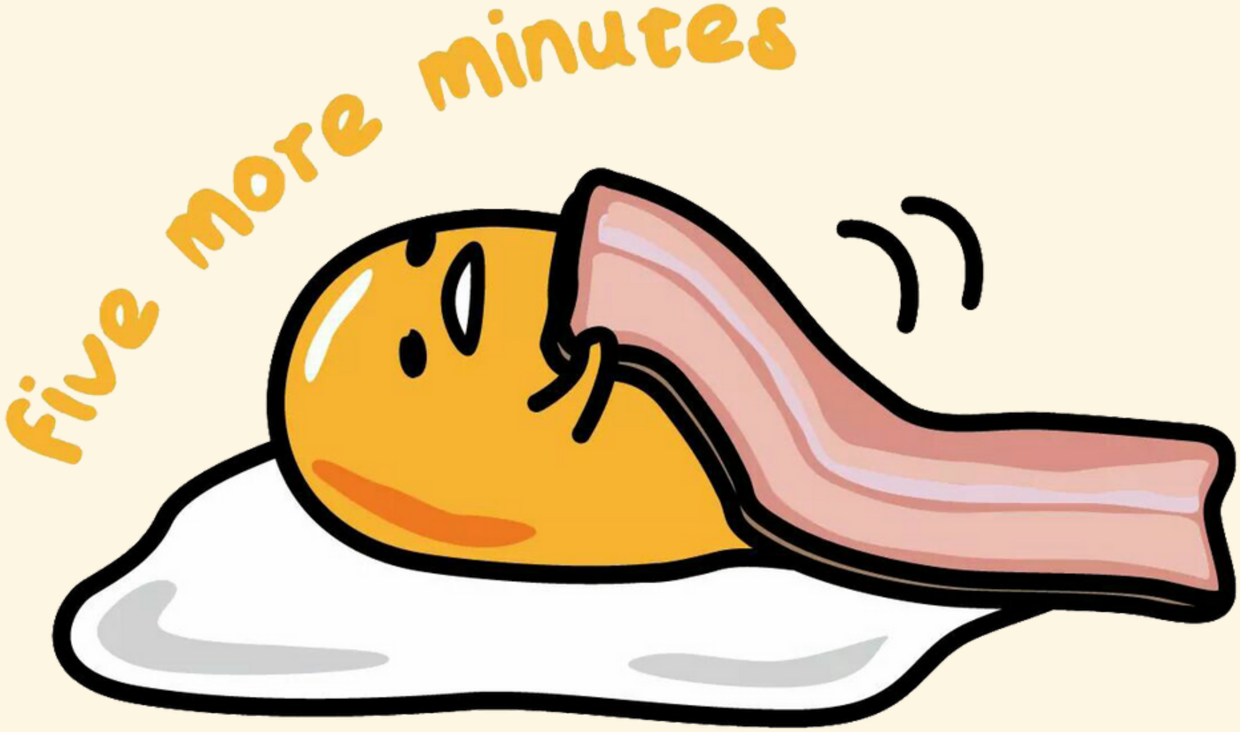
often used in statistics classes to calculate CDFs of Normal(m, s)

using a 'z-table'

But we won't do that in this class because ... mumble mumble ... computers exist

PollEv.com/brendanbrown849

poll closes at __________


five more minutes