📖

Probability Unit 4

Topic: Continuous Distributions

- Date: @February 22, 2024
- Lane: *Probability*, *Statistics*

What it covers

- Events on the real number line.
- Formal definition of probability density & probability distribution.
- Expectation of continuous distributions. Moments. $\int xP(X = x)$
- The normal distribution $\boldsymbol{N}(\mu, \sigma^2)$
- Simulating distributions.

Requires

Random variables, discrete distributions.

Required by

Linear regression

Density estimation

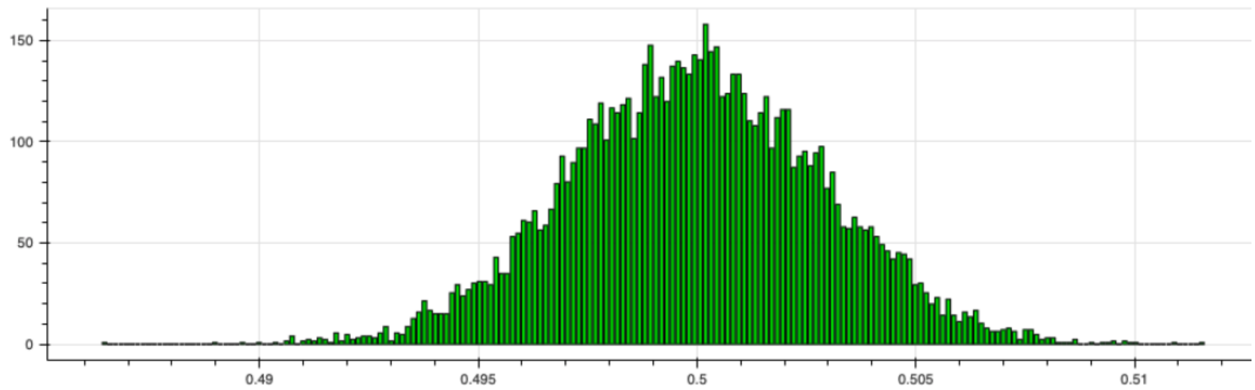
How to assign probabilities to continuous variables.

Conundrum: For $x \in \mathbb{R}, P(X = x) = 0!$ The definition of an *event* for continuous r.v.s are *intervals* on the real line.

> 💡 Probability is fundamentally about discrete events, and the entire theory carries over to continuous spaces.

Continuous distributions as the limit of discrete ones. Where there are summations for discrete distributions there are integrals for continuous distributions.

This is the histogram of 10000 runs of the sum of 10000 random numbers between 0 and 1. (A total of 1E8 draws from a random variable.



Continuous distributions

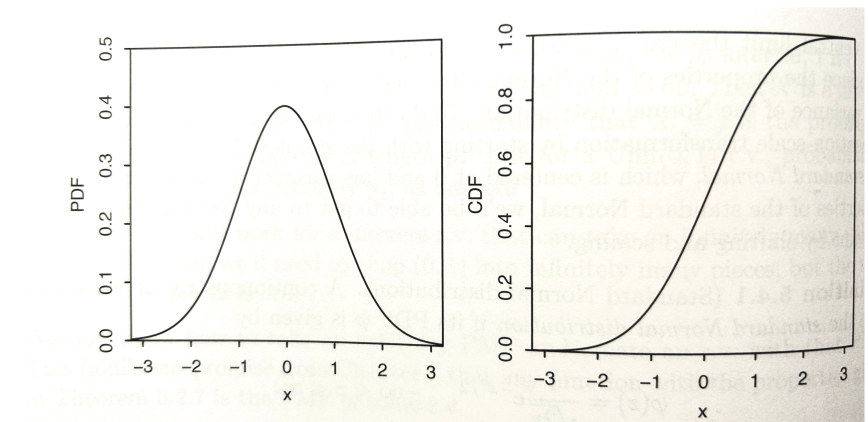


FIGURE 5.9
Standard Normal PDF φ (left) and CDF Φ (right).

> Definition. **Probability Density Function (PDF), Cumulative Distribution Function (CDF).** A continuous function of a random variable. The area over a (possibly unbounded) interval defines the probability of an event. Conventionally we use $f(x)$ for the PDF, and $F(x)$ for the CDF

$$P(a < X < b) = \int_a^b f(x)dx = F(x)|_a^b$$

The CDF is the integral of the PDF. Note the left hand side is a probability, the right hand side is a function.

For discrete r.v.s we speak of a "probability mass function" (PMF) rather than a PDF. It's possible to have a combined PMF - PDF. Then the CDF will have discontinuities.

Comparing a discrete CDF with a continuous one.

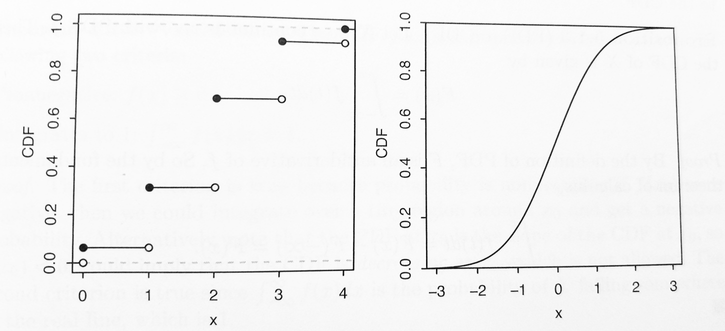


FIGURE 5.1
Discrete vs. continuous r.v.s. Left: The CDF of a discrete r.v. has jumps at each point in the support. Right: The CDF of a continuous r.v. increases smoothly.

Simulation Theorem

A uniform distribution assigns the same probability to every random variable in its domain. So it's PDF on the unit interval is just the 1 by 1 square. It is useful as a universal probability distribution simulation method. Starting with a uniform distribution any distribution can be simulated using the inverse of the distribution's CDF function. If ***U*** is a uniformly distributed random variable on the interval (0,1), then the random variable ***X*** with CDF *F(x)* is distributed $X \sim F^{-1}(U)$.

The "normal" distribution

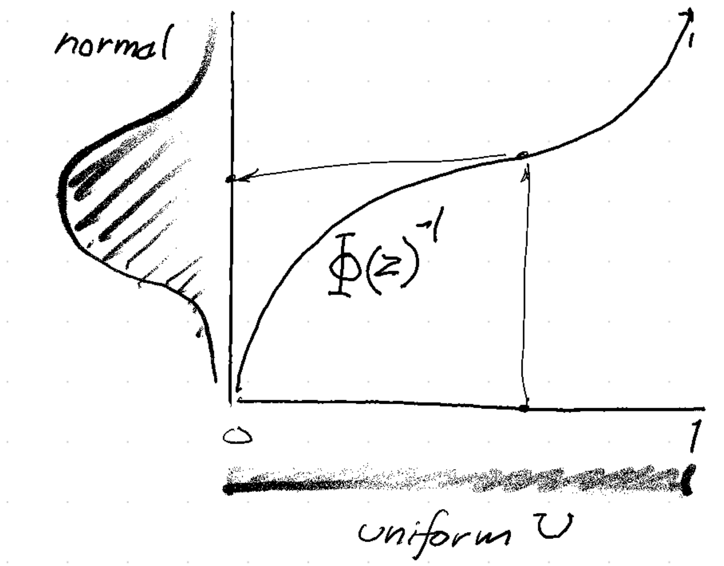
Also called a "Gaussian" distribution, specifically when speaking of e.g. "Gaussian errors".

💡 The normal distribution $\boldsymbol{N}(\mu, \sigma^2)$, with mean μ and variance, σ^2. A *standardized normal* random variable is written as $\boldsymbol{Z} \sim \boldsymbol{N}(0,1)$.

💡 The normal PDF is $P(Z = z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.
The normal CDF is
$P(Z < z) = \Phi(z) = \int_{-\infty}^{z} \varphi(x)\, dx$

There is no closed form for the CDF, however packages such as scipy include the "error function" erf(), such that $\Phi(z) = 1/2(1 + \text{erf}(z/\sqrt(2))$.

Using the inverse CDF to generate normal variate samples:



The normal distribution is useful as the limit of other distributions as the sample size *n* gets large.

Assignment

Read: Evans, Chapter 2.4, 2.51 - 2.53, 2.10

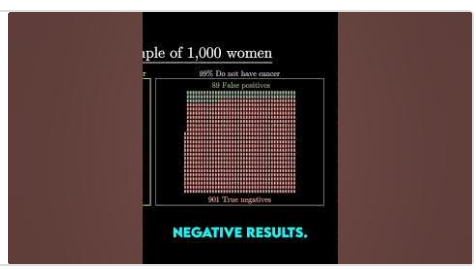
For a future class:

Watch this video on the Base Rate Fallacy, an application of Bayes rule.

The medical test paradox (well "paradox")

A link to the full video about Bayesian thinking is at the bottom of the screen.Or, for reference: https://youtu.be/lG4VkPoG3koLong-to-short editing by

https://www.youtube.com/shorts/xIMlJUwB1m8



References

Comprehensive libraries for families of distributions are available in the python package scipy.stats, for the PDF, CDF, inverse CDF, among others, and for generating r.v.s from the distribution. The python package numpy as has functions to generate random variables for numerous distributions.