YANDEX SCHOOL OF DATA SCIENCE

Probability and Statistics for Data Science

Lecture 2 – Random variables and probability distributions

Today

- Random variables
- Probability mass function (PMF)
- Expectation
- Variance
- Discrete distributions
- Continuous random variables
- Cumulative distribution function
- Convergence of random variables
- Law of large numbers
- Central limit theorem

Random variables (why?)

- Last week, we learnt how to calculate probabilities.
- We saw that it can get messy if the possible set of outcomes is large or if we are interested in different events.
- **Example:** Tossing a coin twice, we deal with the sample space

$$\Omega = \{HH, HT, TH, TT\}$$

- Number of times T appears
- Longest sequence of T
- Maximum number with the same outcome

Same sample space with different partitions

A way to deal with that is talking about random variables.

Random variables (how?)

- **Def:** A random variable is a mapping $X: \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

In our example: Considering the number of times that T appeared,

$$\{X = 0\} = \{HH\}$$
$$\{X = 1\} = \{HT, TH\}$$
$$\{X = 2\} = \{TT\}$$

What if we toss the coin n times?

$$\{X = i\} = \{i\ times\ T\ out\ of\ n\ tosses\}$$

Probability mass function (PMF)

- We can calculate the probability of each possible value x of X as follows:

$$P_X(x) = P(\{X = x\}) = \sum_{\omega : X(\omega) = x} P(\{\omega\})$$

- The function $\mathrm{P_X}(\cdot)$ is called the probability mass function of the random variable X.

Properties:

1. The events $\{\mathrm{X} = x\}$ and $\{\mathrm{X} = x'\}$ are disjoint for $x \neq x'$.
2. X forms a partition of the sample space: $\cup_x \{X = x\} = \Omega$.
3. $\sum_x P_X(x) = 1$.

The values that X can get is called the support of X, denoted $supp(X)$.

Example

- To calculate the PMF of X, we need, for each possible value x of X, to:

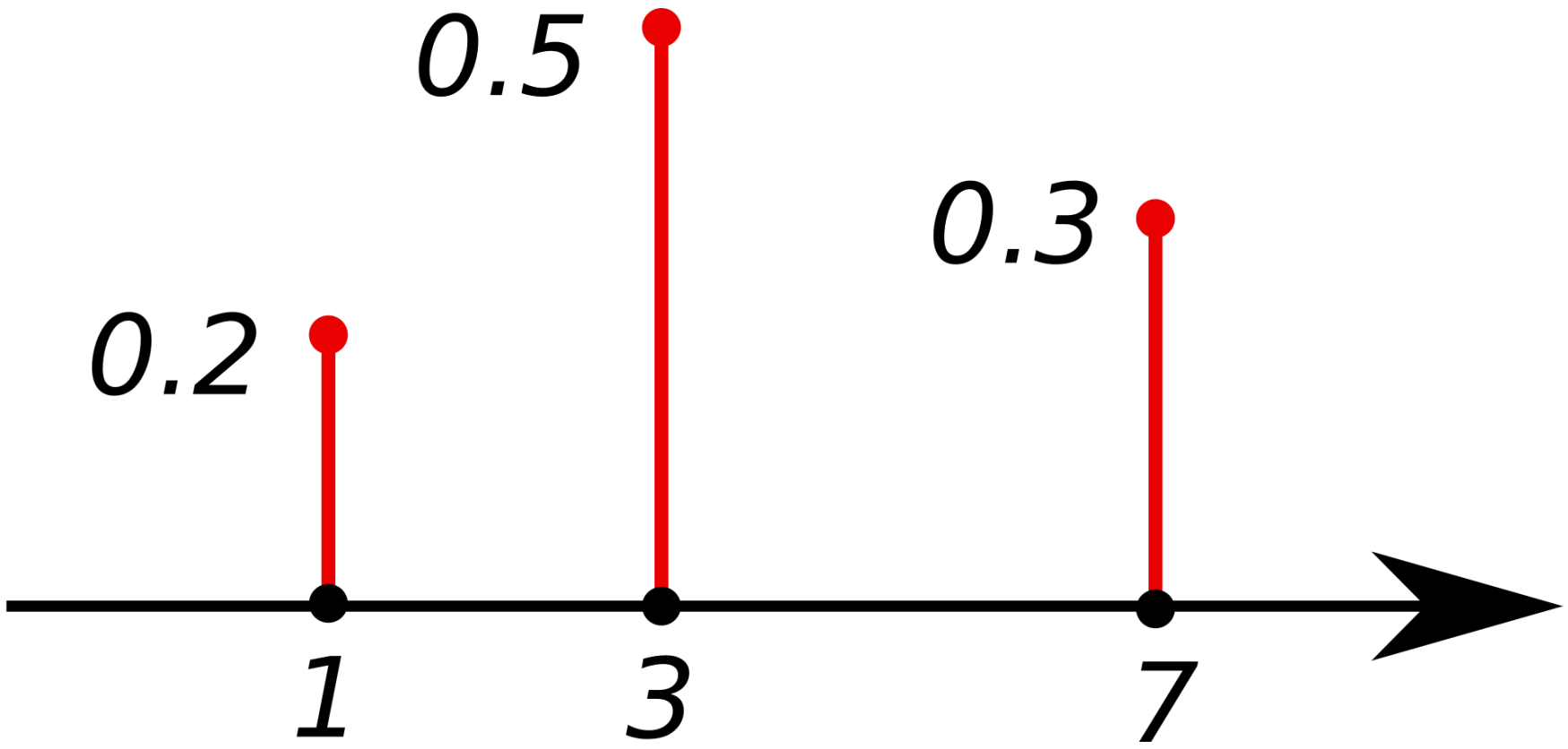
1. Collect the possible outcomes that form the event $\{\mathrm{X} = x\}$.
2. Add their probabilities to obtain $\mathrm{P_X}(x)$.

Example: Find the PMF of the random variable that counts the number of Tails in two tosses of a coin with probability 0.1 of Heads.

x	P(X=x)
0	0.1 x 0.1
1	2 x 0.1 x 0.9
2	0.9 x 0.9

Plot the PMF

What is the random variable in this case?



Expected value (expectation)

- **Def:** If X is a random variable with PMF $\mathrm{P}_X(\mathrm{x})$, then the expectation, or the expected value of X is defined by

$$E(X) = \sum_{x \in supp(X)} x \cdot P(X = x)$$

In words, the expected value of X is a weighted average of the possible values that X can take, each value being weighted by its corresponding probability.

It can also be interpreted as the average of the outcomes of infinetely many independent (but similar) experiments.

Expectation

- The support of X might be countably infitinite. This means that the infinite sum may not converge. So, the expectation exists only if one of the sums in the RHS is finite

$E(X) = \sum_{x \in supp(X)} xP(X = x) = \sum_{x \leq 0} xP(X = x) + \sum_{x > 0} xP(X = x)$

- The expected value does not necessarily belong to the support of X.
- $\min(X) \leq E(X) \leq \max(X)$
- The expectation can be computed through elementary events

$$E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega)$$

Example 1

In some casino game, a player can get:

- \$5.5 with probability 0.1
- \$3 with probability 0.5
- \$0

It costs \$2.5 to participate in the game.

Denote by X the gambler's revenue after two independent games.

Is it convenient to play this game?

$$supp(X) = \{6, 3.5, 1, 0.5, -2, -5\}$$

To see what happens in the long run, we need to compute $\mathrm{E}(X)$.

In this case, $\mathrm{E}(X) < 0$, so it is not convenient to participate.

Example 2

A contestant on a quiz show is presented with two questions, which he is to attempt to answer in some order he chooses. If he decides to try question i first, then he will be allowed to go on to question j only if the first answer is correct. Otherwise, he is not allowed to answer to the second question. We have the following information:

Question	Probability to answer correctly	Reward
Q1	0.6	$200
Q2	0.8	$100

Which question should the contestant answer first?

Example 2 – solution

Which question should the contestant answer first?

Question	Probability to answer correctly	Reward
Q1	0.6	$200
Q2	0.8	$100

Strategy 1: Answer Q1 first

$$E(X_1) = 0.6 \cdot 0.8 \cdot 300 + 0.6 \cdot 0.2 \cdot 200 + 0.4 \cdot 0 = 168$$

Strategy 2: Answer Q2 first

$$E(X_2) = 0.8 \cdot 0.6 \cdot 300 + 0.8 \cdot 0.4 \cdot 100 + 0.2 \cdot 0 = 176$$

Conclusion: Since $E(X_2) > E(X_1)$, it's better to attempt to answer Q2 first.

Expectation of a function

Proposition: Let X be a random variable with PMF P_X, and let $g(X)$ be a function of X. Then,

$$E\big(g(X)\big) = \sum_{x \in supp(X)} g(x) P_X(x)$$

- Another way to compute the expectation of $Y = g(X)$ is first to compute the distribution of Y and then use the regular formula for expectation.

$\boldsymbol{x}$	-1	0	1
$\boldsymbol{P_X(x)}$	0.3	0.6	0.1
$\boldsymbol{Y = x^2}$	1	0	1

Variance

- A dispersion measure for random variable is the <u>variance</u>.
- **Def:** The variance of the random variable X is defined to be

$$Var(X) = E\big(X - E(X)\big)^2$$

- The variance measures the "spread" of a distribution around its mean.
- $\sqrt{Var(X)}$ is called the <u>standard deviation</u>. It has the same measurement units as X.

- Remark: The variance is the expected value of the function

$$f(X) = \big(X - E(X)\big)^2$$

- A handy formula for the variance is

$$Var(X) = E(X^2) - E^2(X)$$

Families of distributions 1

Uniform distribution: $X \sim U(a, b)$

- A collection of equiprobable events
- $b \geq a$ are integers
- $supp(X) = \{a, a+1, \dots, b-1, b\}$
- PMF

$$P_X(x) = \frac{1}{b - a + 1}$$

- Example: A fair die.

Bernoulli distribution: $X \sim Ber(p)$

- An experiment with binary outcome 0/1 (Fail/Success)
- p is the probability of success
- $supp(X) = \{0,1\}$
- PMF

$$P_X(x) = p^x (1 - p)^{1-x}$$

Families of distributions 2

Binomial distribution: $X \sim Bin(n, p)$

- Number of successes out of n independent Bernoulli experiment.
- p is the prob. of success in each experiment.
- $supp(X) = \{0,1,2, \dots, n\}$
- PMF

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Example: Number of students that pass a test.

Geometric distribution: $X \sim Geo(p)$

- Number of Bernoulli trials until (and including) the first success.
- p is the prob. of success in each experiment.
- $supp(X) = \{1,2,3, \dots, \infty\}$
- PMF

$$P_X(x) = (1-p)^{x-1} p$$

- Special property: memoryless. Meaning:

$$P(\{X > s + t\} | \{X > t\}) = P(\{X > s\})$$

Families of distributions 3

Poisson distribution: $X \sim Pois(\lambda)$

- Number of events per unit of time with constant rate
- $\lambda > 0$ is the rate parameter
- $supp(X) = \{0,1,2,\dots,\infty\}$
- PMF

$$P_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- Example: The number of customers entering a post office on a given day.
- Special property: The Poisson distribution is an approximation to the binomial distribution as

$$p \to 0, n \to \infty, np \to \lambda$$

Continuous random variables

- Continuous RV's allow us to model continuous phenomena.
- Sometimes we can use a discrete model, but using a continuous one is possibly more accurate.
- Continuous models allow the use of powerful tools from calculus.
- In this part, we will see the concepts we have seen so far in their continuous version.
- In general, we move from $\sum$ to $\int$.

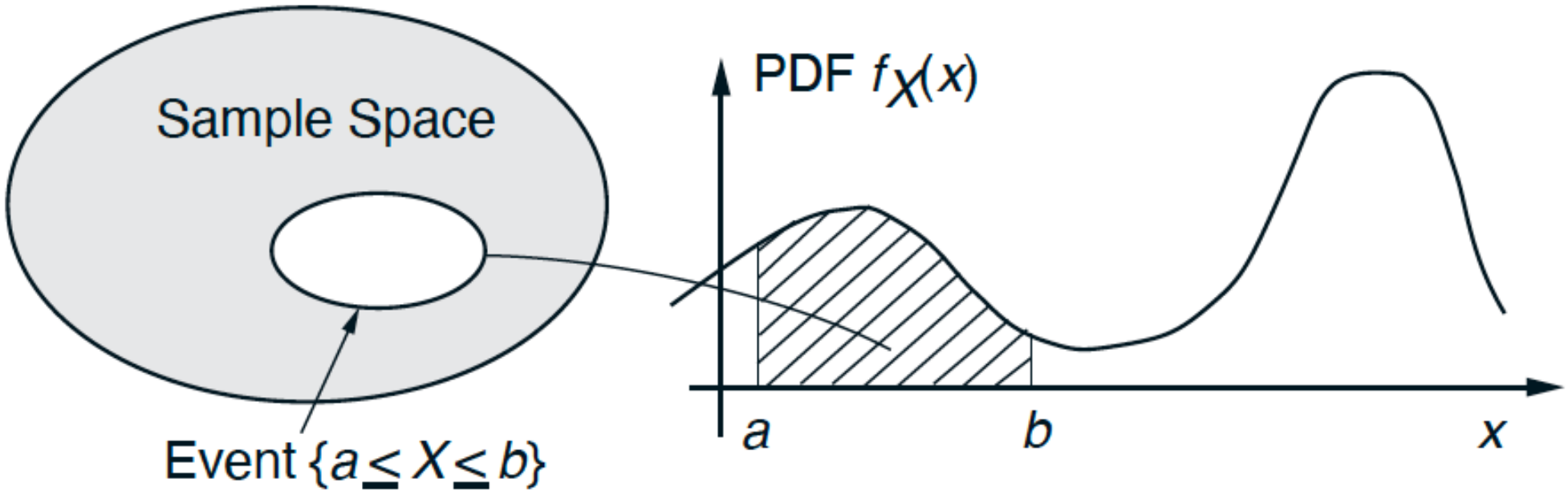
Continuous RV's and PDF's

- **Def:** A random variable X is called continuous if there is a nonnegative function f_X, called the probability density function of X such that

$$P(X \in B) = \int_B f_X(x)dx$$

  for every subset B of the real line.
- To qualify as a PDF, a function f_X must be nonnegative and must have the normalization property. That is, $f_X(x) \geq 0, \forall x$ and $\int_{\mathbb{R}} f_X(x)dx = 1$.

Interpretation of PDF's

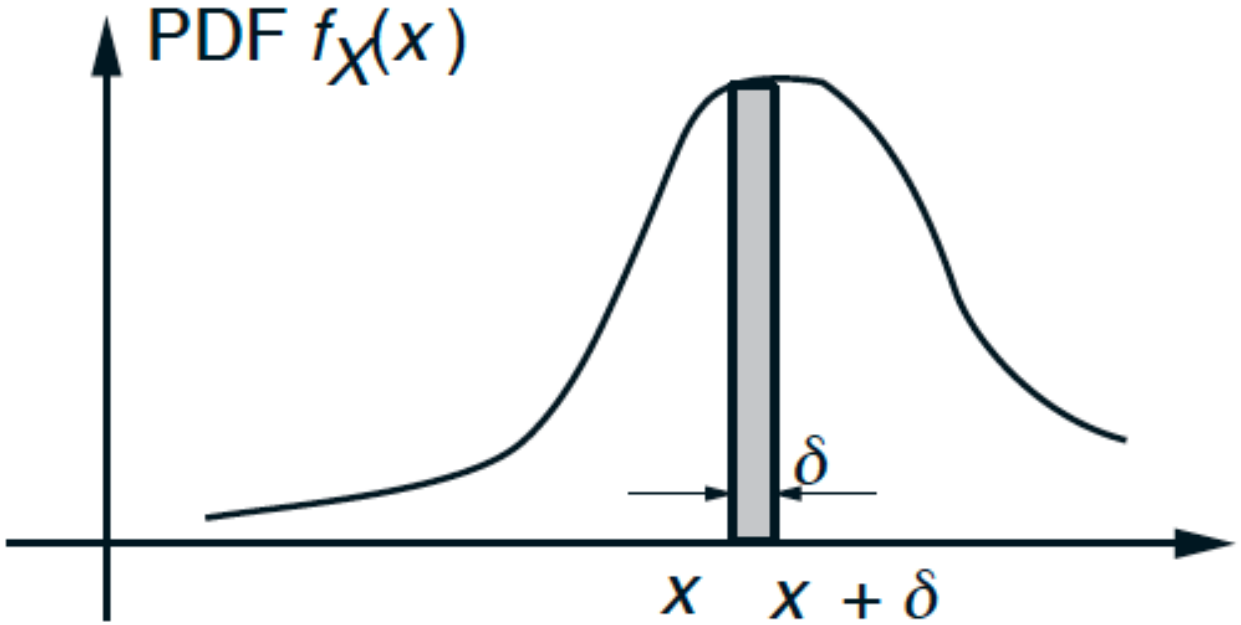


$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Interpretation of PDF's

$$P(x \leq X \leq x + \delta) = \int_{x}^{x+\delta} f_X(t)dt \approx f_X(x)\delta$$

We can view the PDF as "probability mass per unit of length". However, it is not a probability.



Example: Uniform distribution

Let X be a continuous random variable that takes values in an interval $[a, b]$, and assume that any two subintervals of the same length have the same probability.

What's the PDF of this random variable?

Solution: The PDF has the form

$$f_X(x) = \begin{cases} c & if\ a \leq x \leq b \\ 0 & o.w. \end{cases}$$

To find c, we integrate the function over $[a, b]$ and equate to one.

Example: piecewise constant PDF

Bob's driving time to work is between 15 and 20 minutes if the day is sunny, and betweem 20 and 25 minutes if the day is rainy, with all times being equally likely in each case. Assume that a day is sunny w.p. 2/3 and rainy w.p. 1/3. What is the PDF of the driving time, viewed as a random variable X?

Solution: "Equaly likely" means that the PDF is constant in each time interval. The PDF has the form

$$f_X(x) = \begin{cases} c_1 \; if \; x \in [15,20] \\ c_2 \; if \; x \in [20,25] \\ 0 \quad otherwise \end{cases}$$

We can determine these constants using the given probabilities:

$$\frac{2}{3} = P(sunny\ day) = \int_{15}^{20} c_1 dx = 5c_1$$

Expectation and more of a continuous RV

Let X be a continuous random variable with PDF f_X.

- Expectation:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

- Expectation of a function

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- Variance:

$$Var(X) = \int_{-\infty}^{\infty} \left(x - E(X)\right)^2 f_X(x) dx$$

Some famous continuous distributions

- **Uniform distribution**

When any two subintervals of the same length have the same probability.

- **Exponential distribution**

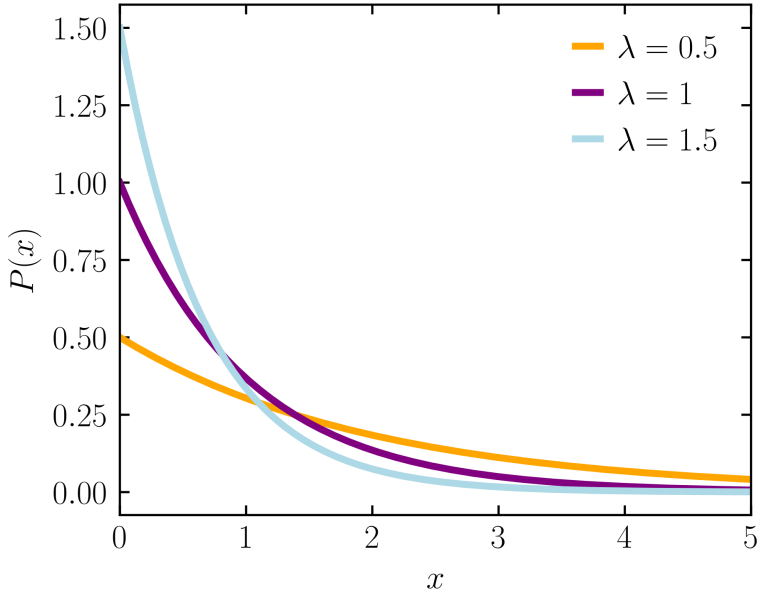
Time until some event happens. Memoryless.

- **Normal distribution**

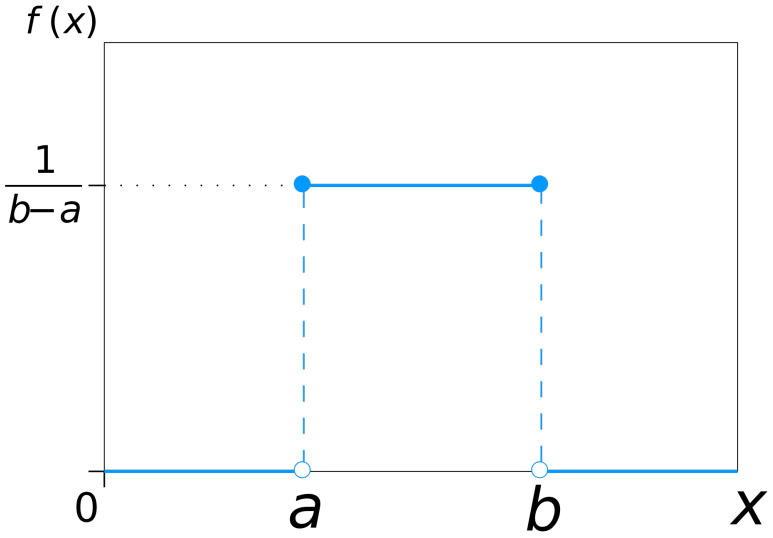
The average of many observations of a random variable with finite mean and variance is itself a random variable—whose distribution converges to a normal distribution.

Density plots

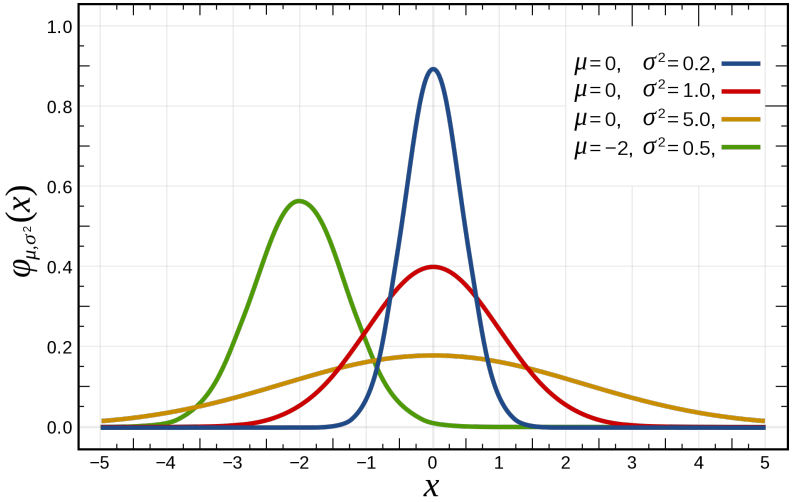
$X \sim Exp(\lambda)$



$X \sim U[a, b]$



$X \sim N(\mu, \sigma^2)$



Cumulative distribution function (CDF)

- **Def:** The cumulative distribution function (CDF) of a random variable X is denoted by F_X and provides the probability $P(X \leq x)$. In particular,

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} P(X = k) & X\ discrete \\ \int_{-\infty}^{x} f_X(t)dt & X\ cont. \end{cases}$$

Why CDF?

Theorem: Every probability distribution is uniquely defined by the CDF.

Properties of CDF $F_X(x) = P(X \leq x)$

- Monotonically nondecreasing: if $x \leq y$ then $F_X(x) \leq F_Y(y)$
- $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$

If X is discrete:

- $F_X(x)$ is piecewise constant as function of x
- We obtain the PMF by

$P_X(k) = P(X \leq k) - P(X \leq k-1) = F_X(k) - F_X(k-1)$

If X is continuous:

- $F_X(x)$ is continuous as function of x
- We obtain the PDF by

$$f_X(x) = \frac{dF_X}{dx}(x)$$

Normal distribution

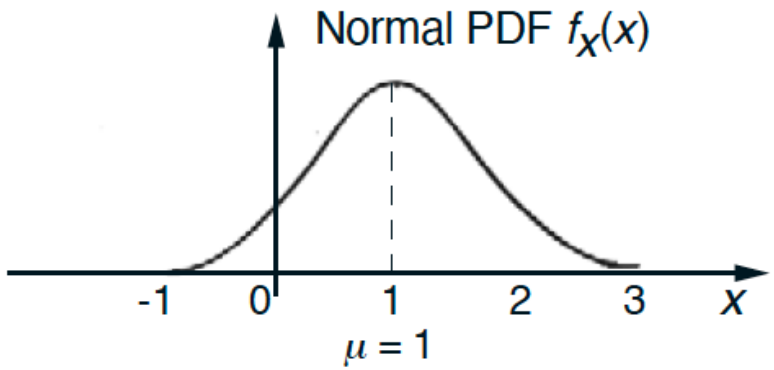
- **Def:** A continuous RV X is said to be normal if it has a PDF of the form

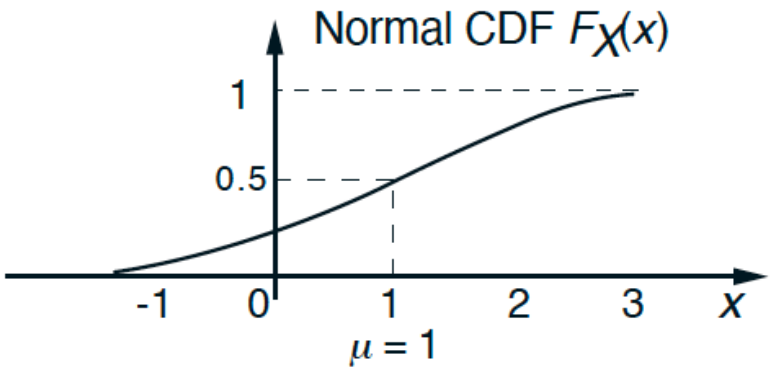
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma > 0$$

Properties:

- $E(X) = \mu, Var(X) = \sigma^2$
- Normality is preserved under linear transformations. That is, if $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
- A random variable $X \sim N(0,1)$ is called standard normal. We denote its CDF by Φ.
- By symmetry, $\Phi(-x) = 1 - \Phi(x)$
- We can standardize a normal random variable X by $Y = \frac{X-\mu}{\sigma}$

Example





The annual snowfall in Boston is modeled as a normal RV with a mean of $\mu = 60$ cm and a standard deviation of $\sigma = 20$. What is the probability that this year's snowfall will be at least 80 cm?

Solution: Let X be the snow accumulation and let $Y = \frac{X-\mu}{\sigma} = \frac{X-60}{20}$. Then,

$$P(X \geq 80) = P\left(\frac{X-60}{20} \geq \frac{80-60}{20}\right) = P\left(Y \geq \frac{80-60}{20}\right)$$
$$= P(Y \geq 1) = 1 - P(Y < 1) = 1 - \Phi(1) = 0.1587$$

Convergence of random variables

- In the deterministic world we are familiar with the notion of convergence of sequences

$$\lim_{n\to\infty} a_n = a \Leftrightarrow \forall \epsilon > 0 \exists N \in \mathbb{N}: \forall n > N, |a_n - a| < \epsilon$$

- The problem is that the requiremnt "for all" in a random world is too much to ask.
- **Def:** A sequence of random variables $X_1, X_2, \ldots$ converges to a if for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - a| \geq \epsilon) = 0$$

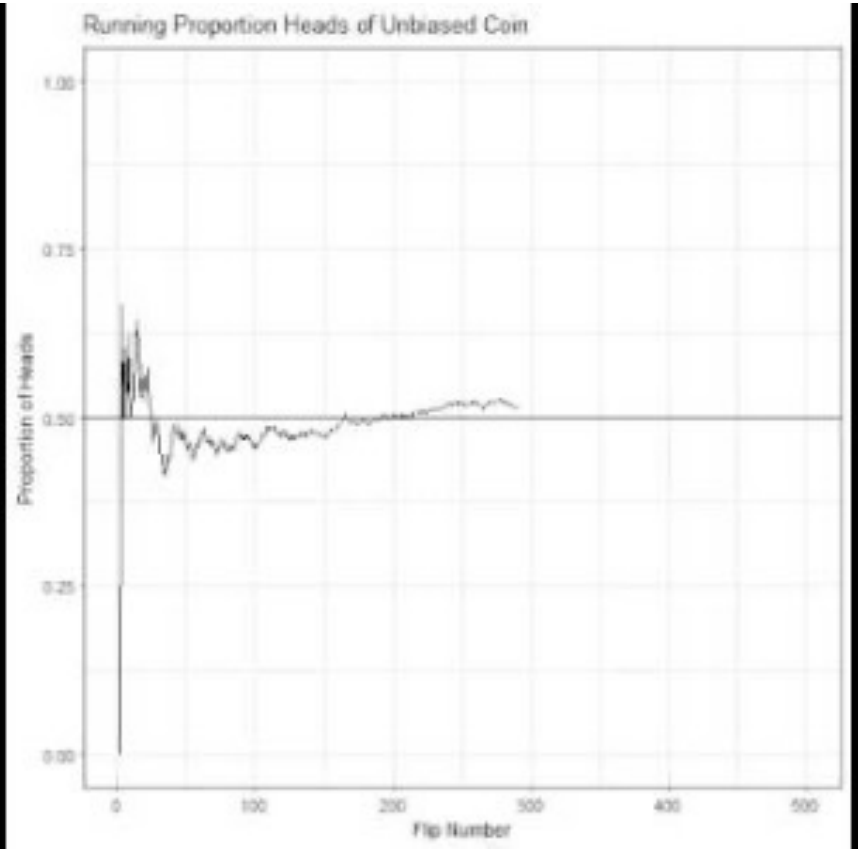
- There are more (weaker and stronger) notions of convergence of RV's that are beyond the scope of our course.

The (weak) Law of Large Numbers

Theorem: Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $\mathrm{E}(X) < \infty$. Then,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} E(X)$$

Where the convergence is in probability.



Central Limit Theorem

Theorem: Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ and finite variance σ^2. Then,

$$P\left(\frac{\overline{\boldsymbol{X}}_{\boldsymbol{n}} - \mu}{\sigma}\sqrt{n} \leq t\right) \xrightarrow[n\to\infty]{} \Phi(t)$$

Where Φ is the CDF of a $N(0,1)$ random variable.

Visualization of CLT here!

Play with the parameters. For example, alpha = 0.54, beta = 1.26, sample size=15, draws =1.

Remark: This kind of convergence is called convergence in distribution.

Example

Suppose that the number of errors per computer program has a Poisson distribution with mean 5. We get 125 programs. Let $X_1, \dots, X_{125}$ be the number of errors in the program. We want to approximate $P(\bar{X}_n < 5.5)$.

Solution: Let $\mu = E(X_1) = 5, \sigma^2 = Var(X_1) = 5$. Then,

$$P(\bar{X}_n < 5.5) = P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma}\right) \approx P(Z < 2.5)$$

$$= \Phi(2.5) = 0.9938$$

References

- Bertsekas, Dimitri P. and Tsitsiklis, John N.. "Introduction to Probability." 2008 .
- Ross, Sheldon M. A First Course in Probability / Sheldon Ross. Eighth edition, global edition. Harlow: Pearson Education Limited, 2010.
- Haviv, Moshe. Introduction to Descriptive Statistics and Probability, 2021.
- Wasserman, Larry. *All of statistics : a concise course in statistical inference*. New York: Springer, 2010.
- https://pradeepadhokshaja.wordpress.com/2017/04/06/visualizing-the-law-of-large-numbers-using-gganimate/