Probability Review

15.075 Cynthia Rudin

A probability space, defined by Kolmogorov (1903-1987) consists of:

- A set of *outcomes* S, e.g.,

  for the roll of a die, $S = \{1, 2, 3, 4, 5, 6\}$,

  for the roll of two dice, $S = \left\{ \binom{1}{1}, \binom{1}{2}, \binom{2}{1}, \binom{1}{3}, \ldots, \binom{6}{6} \right\}$

  temperature on Monday, $S = [-50, 50]$.

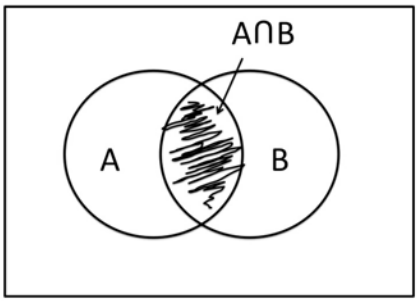
- A set of *events*, where an event is a subset of S, e.g.,

  $$\text{roll at least one 3} \quad : \quad \left\{ \binom{1}{3}\binom{3}{1}, \ldots, \binom{3}{3} \right\}$$
  $$\text{temperature above } 80^\circ \quad : \quad (80, 150]$$

  The union, intersection and complement of events are also events (an algebra).

- A *probability measure*, which gives a number between 0 and 1 to each event, where $P(S) = 1$ and

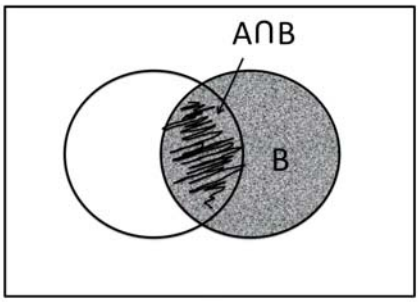
  $$A \cap B = \emptyset \qquad \Rightarrow \qquad P(A \cup B) = P(A) + P(B).$$



Think of P as measuring the size of a set, or an area of the Venn diagram.

The conditional probability of event A given event B is:

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$



Event A is independent of B if $P(A|B) = P(A)$. That is, knowing B occurred doesn't impact whether A occurred (e.g. A and B each represent an event where a coin returned heads). In that case,

$$P(A) = P(A|B) := \frac{P(A \cap B)}{P(B)} \qquad \text{so} \qquad P(A \cap B) = P(A)P(B).$$

Do not confuse independence with *disjointness*. Disjoint events A and B have $P(A \cap B) = 0$.

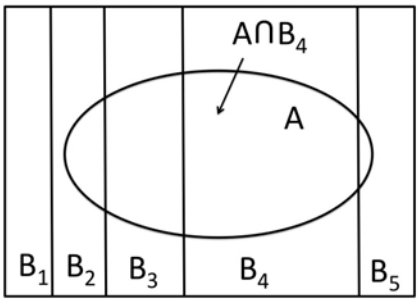
For a *partition* $B_1, \ldots, B_n$, where $B_i \cap B_j = \emptyset$ for $i \neq j$ and $B_1 \cup B_2 \cdots B_n = S$ then

$$A = (A \cap B_1) \cup (A \cap B_2), \ldots, \cup (A \cap B_n)$$

and thus

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i} P(A|B_i)P(B_i)$$

from the definition of conditional probability.



Bayes Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

Derivation of Bayes Rule

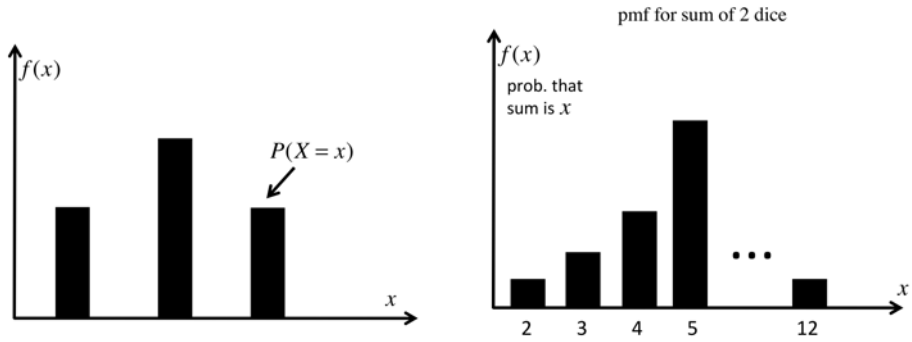
Monty Hall Problem

A random variable (r.v.) assigns a number to each outcome in S.

Example 1: toss 2 dice: random var. X is the sum of the numbers on the dice.
Example 2: collections of transistors: random var X is the min of the lifetimes of the transistors.

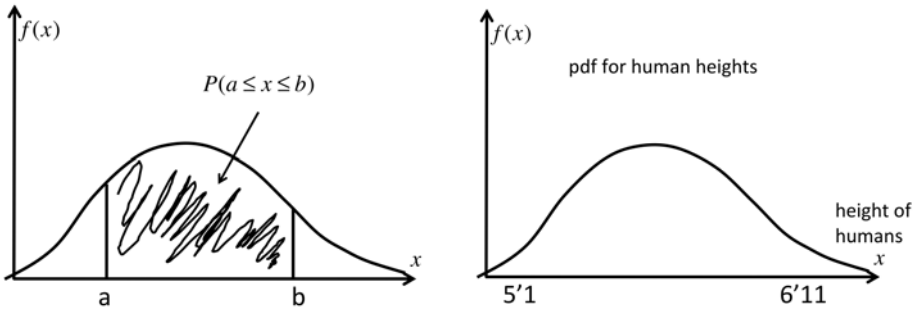
A probability density function (pdf... or probability mass function pmf) for random variable X is defined via:

- $f(x) = P(X = x)$ for discrete distributions



(Note $f(x) \geq 0$, $\sum_x f(x) = 1$.)

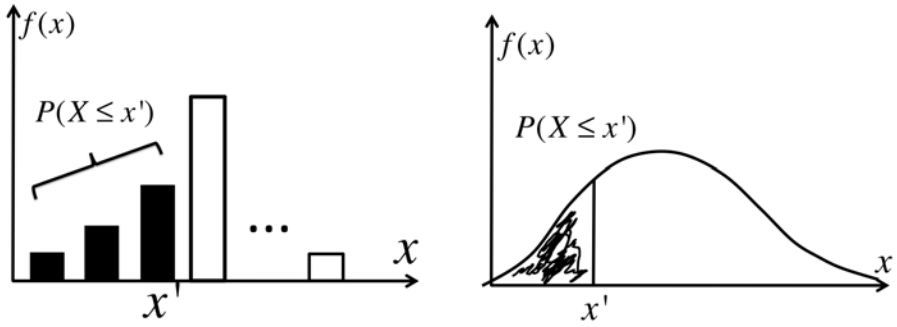
- $P(a \leq X \leq b) = \int_a^b f(x)dx$ for continuous distributions.



(Note $f(x) \geq 0$, $\int f(x) = 1$.)

The cumulative distribution function (cdf) for r.v. X is:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \sum_{k \leq x} f(k) \ \text{(discrete)} \\ &= \int_{-\infty}^{x} f(y)dy \ \text{(continuous)} \end{aligned}$$



The expected value (mean) of an r.v. X is:

$$\mathbf{E}(X) = \mu = \sum_x xf(x) \quad \text{(discrete)}$$

$$\mathbf{E}(X) = \mu = \int_x xf(x)dx \quad \text{(continuous)}.$$

Expectation is *linear*, meaning

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y).$$

Roulette

The variance of an r.v. X is:

$$Var(X) = \sigma^2 = \mathbf{E}(X - \mu)^2.$$

Variance measures dispersion around the mean. Variance is not linear:

$$Var(aX + b) = a^2 Var(X).$$

The standard deviation (SD) is:

$$SD(x) = \sigma = \sqrt{Var(X)}.$$

Note that people sometimes use another definition for variance that is equivalent:

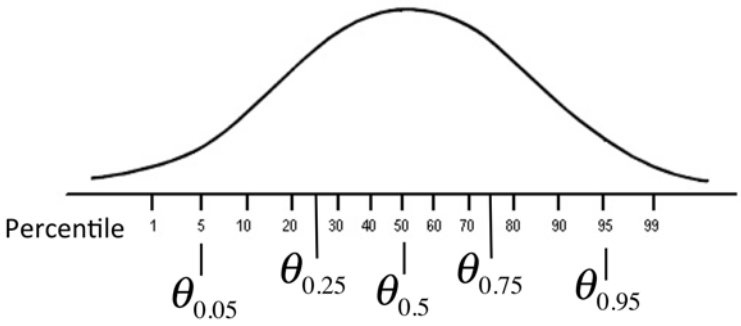
$$Var(X) = \sigma^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Sum of Two Dice

Quantiles/Percentiles The p^{th} quantile (or $100p^{\text{th}}$ percentile), denoted θ_p, of r.v. X obeys:

$$P(X \leq \theta_p) = p.$$

Note: 50^{th} percentile, or $.5^{\text{th}}$ quantile, is the median $\theta_{.5}$.



Exponential Distribution

Section 2.5. Jointly Distributed R.V.'s

Jointly distributed r.v.'s have joint pdf's: $f(x, y) = P(X = x, Y = y)$.

Their covariance is defined by

$$Cov(X, Y) := \sigma_{x,y} := \mathbf{E}[(X - \mu_x)(Y - \mu_y)].$$

The first term considers dispersion from the mean of X, the second term is dispersion from the mean of Y.

If X and Y are independent, then $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y) = \mu_x\mu_y$ (see "expected value" on Wikipedia). So, multiplying the terms out and passing the expectation through, we get:

$$Cov(X, Y) = \mathbf{E}[(X - \mu_x)(Y - \mu_y)] = \mathbf{E}(XY) - \mu_x\mathbf{E}(Y) - \mu_y\mathbf{E}(X) + \mu_x\mu_y = 0.$$

Useful relationships:

1. $Cov(X, X) = \mathbf{E}[(X - \mu_x)^2] = Var(X)$
2. $Cov(aX + c, bY + d) = ab\ Cov(X, Y)$
3. $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$ where $Cov(X, Y)$ is 0 if X and Y are indep.

The correlation coefficient is a normalized version of covariance so that its range is [-1,1]:

$$\rho_{XY} := Corr(X, Y) := \frac{Cov(X, Y)}{\overline{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Section 2.6.

Chebyshev's Inequality (we don't use it much in this class)

$$\text{For } c > 0, P(|X - \mathbf{E}X| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Chebyshev's inequality is useful when you want a *upper bound* on how often an r.v. is far from its mean. You can't use it for directly calculating $P(|X - \mathbf{E}X| \geq c)$, only for bounding it.

Weak Law of Large Numbers

Let's say we want to know the average number of pizza slices we will sell in a day. We'll measure pizza sales over the next couple weeks. Each day gets a random variable X_i which represents the amount of pizza we sell on day i. Each X_i has the same distribution, because each day basically has the same kind of randomness as the others. Let's say we take the average of over the couple weeks we measured. There's a random variable for that, it's $\bar{X} = \frac{1}{n}\sum_i X_i$.

- Does $\bar{X}$ have anything to do with the average pizza sales per day? In other words, does measuring $\bar{X}$ tell us anything about the X_i's? For instance, (on average) is $\bar{X}$ close to the average sales per day, $E(X_i)$?

- Does it matter what the distribution of pizza sales is? For instance, what if we usually have 1-4 customers but sometimes we have a conference where we have 45-50 people; that is kind of a weird distribution. Does that mean that the $\bar{X}_i$ needs to be adjusted in some way to help us understand the X_i's?

It turns out the first answer is yes, in fact pretty often, the average $\bar{X}$ is very similar to the average value of X_i. Especially when n is large. The second answer is also yes, but as long as n is large, $\bar{X}$ is very similar to the average value of X_i. This means that no matter what the distribution is, as long as we measure enough days, we can get a pretty good sense of the average pizza sales.

Weak LLN: Let $\bar{X}$ be a r.v. for the sample mean $\bar{X} = \frac{1}{n}\sum_i X_i$ of n iid (independent and identically distributed) r.v.'s. The distribution for each X_i has finite mean μ and variance σ^2. Then for any $c > 0$,

$$P(|\bar{X} - \mu| \geq c) \to 0 \text{ as } n \to \infty.$$

Weak LLN says that $\bar{X}$ approaches μ when n is large. Weak LLN is nice because it tells us that no matter what the distribution of the X_i's is, the sample mean approaches the true mean.

Proof Weak LLN using Chebyshev

Section 2.7. Selected Discrete Distributions

Bernoulli $X \sim Bernoulli(p)$ "coin flipping"

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \quad \text{"heads"} \\ 1 - p & \text{if } x = 0 \quad \text{"tails"} \end{cases}$$

Binomial $X \sim Bin(n, p)$ "n coins flipping," "balls in a bag drawn with replacement," "balls in an infinite bag"

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n,$$

where $\binom{n}{x}$ is the number of ways to distribute x successes and $n - x$ failures,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

(If you have n coins flipping, $f(x)$ is the probability to get x heads, p is the probability of heads.) $X \sim Bin(n, p)$ has

$$\mathbf{E}(X) = np, \quad Var(X) = np(1-p).$$

You'll end up needing these facts in Chapter 9.

Hypergeometric $X \sim \text{HyGE}(N, M, n)$ "balls in a bag drawn without replacement," where:
N is the size of the total population (the number of balls in the bag),
M is the number of items that have a specific attribute (perhaps M balls are red),
n is our sample size,
$f(x)$ is the probability that x items from the sample have the attribute.

$$f(x) = \frac{\begin{array}{cc}\text{ways to draw } x \text{ balls} & \text{ways to draw } n-x \text{ balls} \\ \text{with attribute} & \text{without attribute}\end{array}}{\text{ways to draw } n \text{ balls}} = \frac{\left(\begin{array}{c} M \\ x \end{array}\right)\left(\begin{array}{c} N-M \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}.$$

Multinomial Distribution "generalization of binomial"
Think of customers choosing backpacks of different colors. A random group of n customers each choose their favorite color backpack. There is a multinomial distribution governing how many backpacks of each color were chosen by the group. Below, x_k is the number of people who ordered the k^{th} color backpack. Also, $f(x)$ is the probability that x_1 customers chose color 1, x_2 customers chose color 2, and so on.

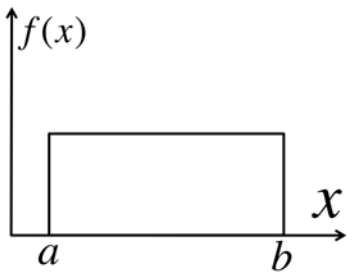
$$\begin{aligned} f(x_1, x_2, x_3, \ldots, x_k) &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_k = x_k) \\ &= \frac{n!}{x_1! x_2! x_3! \ldots x_k!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \ldots p_k^{x_k} \end{aligned}$$

where $x_i \geq 0$ for all i and $\quad_i x_i = n$ (there are n total customers), and p_i is the probability that outcome i occurs. (p_i is the probability to choose the i^{th} color backpack).

2.8 Selected Continuous Distributions

Uniform Distribution $X \sim U[a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



Poisson Distribution $X \sim Pois(\lambda)$ "binomial when $np \to \lambda$," "rare events"

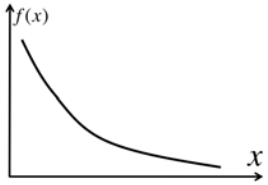
$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{x = 0,1,2,...}$$

$X \sim Pois(\lambda)$ has

$$\mathbf{E}(X) = \lambda, \quad Var(X) = \lambda.$$

Exponential Distribution $X \sim Exp(\lambda)$ "waiting times for Poisson events"

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$



Gamma Distribution $X \sim Gamma(\lambda, r)$ "sums of r iid exponential r.v.'s," "sums of waiting times for Poisson events"

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} \text{for } x \geq 0,$$

where $\Gamma(r)$ is the "Gamma" function, which is a generalization of factorial.

Customers on line

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

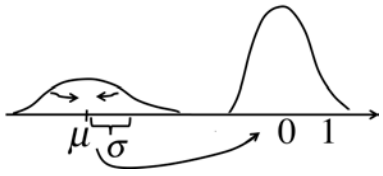
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \text{ for } \infty < x < \infty.$$

We have that:

$$\mathbf{E}(X) = \mu, \quad Var(X) = \sigma^2.$$

We often *standardize* a normal distribution by shifting its mean to 0 and scaling its variance to 1:

$$\text{If } \quad X \sim N(\mu, \sigma^2) \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \quad \text{"standard normal"}$$



Since we can standardize any gaussian, let's work mostly with the standard normal $Z \sim N(0, 1)$:

$$\text{pdf:} \qquad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$\text{cdf:} \qquad \Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(y) dy$$

We can't integrate the cdf in closed form. This means that in order to get from z to $\Phi(z)$ (or the reverse) we need to get the answer from a book or computer where they integrated

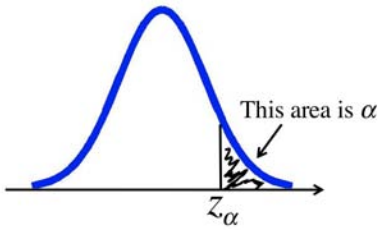
it numerically, by computing an approximation to the integral. MATLAB uses the command `normcdf`. Table A.3 in your book has values for $\Phi(z)$ for many possible z's. Read the table left to right, and up to down. So if the entries in the table look like this:

z	0.03
-2.4	0.0075

This means that for $z = -2.43$, then $\Phi(z) = P(Z \leq z) = 0.0075$. So the table relates z to $\Phi(z)$. You can either be given z and need $\Phi(z)$ or vice versa.

75^{th} percentile calculation

Denote z_α as the solution to $1 - \Phi(z) = \alpha$.



z_α is called the upper α critical point or the $100(1-\alpha)^{\text{th}}$ percentile.

Linear Combinations of Normal r.v.'s are also normal.

For n iid observations from $N(\mu, \sigma^2)$, that is.

$$X_i \sim N(\mu, \sigma^2) \text{ for } i = 1, \ldots, n,$$

the sample mean $\bar{X}$ obeys:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hmm, where have you seen σ^2/n before? Of course! You saw it here:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\sum_i Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

This is the variance of the mean of n iid r.v.'s who each have $Var(X_i) = \sigma^2$.

15.075J / ESD.07J Statistical Thinking and Data Analysis

Fall 2011