# Probability distributions

CS 206: Discrete Structures II

# Probability mass function

The probability mass function[1] gives the probability of a particular value:

$$\text{pmf}_R(x) = \mathbb{P}\left(R = x\right)$$

If we roll a die and $R_1 =$ the value rolled, $R_2 =$ the value squared:

- $\text{pmf}_{R_1}(5) = \mathbb{P}\left(R_1 = 5\right) = 1/6$
- $\text{pmf}_{R_2}(16) = \mathbb{P}\left(R_2 = 16\right) = 1/6$
- $\text{pmf}_{R_2}(100) = \mathbb{P}\left(R_2 = 100\right) = 0$

[1]The book uses "probability density function" instead

# Cumulative distribution function

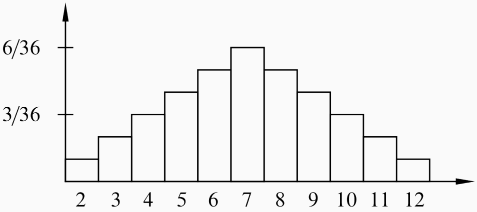
We can look at a cumulative version of probability as long as the co-domain is ordered, called the cumulative distribution function:

$$\begin{aligned} \mathsf{cdf}_R(x) &= \mathbb{P}\left(R \leq x\right) \\ &= \sum_{x_i \leq x} \mathbb{P}\left(R = x_i\right) \end{aligned}$$

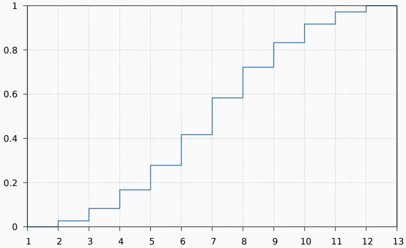
# Example: PMF and CDF

Let's roll a die twice and let $R$ be the sum of the two rolls.

$\mathrm{pmf}_R$




$\mathrm{cdf}_R$




- CDF is monotonically non-decreasing
- CDF goes from 0 at the left to 1 at the right

# Indicator variables

An indicator random variable $I_E$ (aka a Bernoulli variable):

- has value 1 if event $E$ occurs
- and value 0 otherwise.

And it has expectation:

$$\begin{aligned}\mathbb{E}\left(I_E\right) &= 0 \cdot \mathbb{P}\left(I_E = 0\right) + 1 \cdot \mathbb{P}\left(I_E = 1\right) \\ &= \mathbb{P}\left(E\right)\end{aligned}$$

# Indicator variables

For example, let's flip three coins.

Let $M$ be the event that they are all the same (all tails or all heads).

Then its expected value is

$$\begin{aligned} \mathbb{E}\left(I_M\right) &= 0 \cdot \mathbb{P}\left(\bar{M}\right) + 1 \cdot \mathbb{P}\left(M\right) \\ &= 1/4 \end{aligned}$$

since 2 of the 8 possible outcomes for three coin flips are in $M$.

# Common distributions

A few distributions come up often:

- Bernoulli distribution
- Uniform distribution
- Binomial distribution
- Geometric distribution

Let $R$ be a Bernoulli random variable where $E$ occurs with probability $p$. $R$ can only take on values 0 and 1, so the pmf is a map

$$f_p : \{0, 1\} \to [0, 1]$$

where

$$f_p(0) = P\{R = 0\} = 1 - p$$
$$f_p(1) = P\{R = 1\} = p$$

# Uniform distribution

If $R$ is a random variable where all values in the codomain are equally likely, then if there are $n$ values in the codomain, each must have probability $1/n$ (since the probabilities must sum to 1). The pmf is then

$$f_R(x) = \frac{1}{n}$$

and the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < a_1 \\ k/n & \text{if } a_k \leq x < a_{k+1} \\ 1 & \text{if } a_n \leq x \end{cases}$$

# Uniform distribution

For example, if we roll a die and let $R$ be the number rolled, then each of the values from 1 to 6 are equally likely.

So the probability of rolling a 4 is

$$f_R(4) = 1/6$$

The cdf tells us things like the probability of rolling a 4 or less:

$$F_R(4) = 4/6 = 2/3$$

(the sum of the probabilities of all values $\leq 4$)

# Binomial distribution

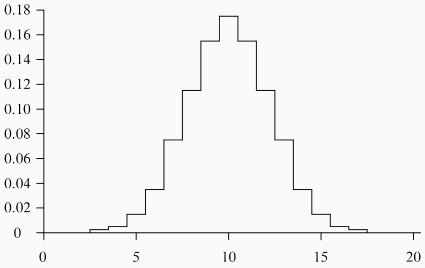
Suppose that we do $n$ independent trials, each of which results in one of two outcomes (e.g., success and failure) with equal probability.

Then if $R$ represents the number of successes, it follows a binomial distribution with a pmf of

$$f_n(k) = \binom{n}{k} \frac{1}{2^n}$$

# Binomial distribution

Suppose we flip a coin 20 times and let $R$ be the number of heads. The probability that $R$ will take on particular values from 0 to 20 yields this graph:




where the probability of, say, 8 heads is $\binom{20}{8}\frac{1}{2^{20}}$.

(There are $2^{20}$ length 20 sequences of Hs and Ts, each equally likely, and the number with 8 Hs is $\binom{20}{8}$)

# Binomial distribution

We can generalize this a bit. Suppose the coin isn't fair:

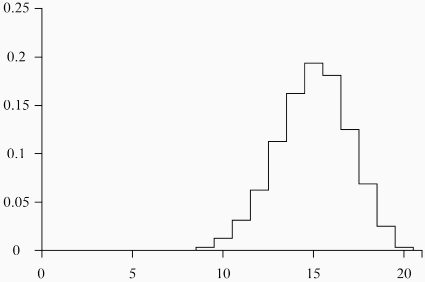
- it has probability $p$ of yielding heads and $q = 1 - p$ of tails.

Then the pmf becomes

$$f_{n,p}(k) = \binom{n}{k} p^k q^{n-k}$$

# Binomial distribution

Suppose we flip a *biased* coin 20 times, where $\mathbb{P}(H) = 3/4$ and let $R$ be the number of heads. Now we get this graph:




The most likely outcome is that 3/4 of the flips result in heads.

What is the probability of getting 8 heads?

# Binomial distribution

If we flip 5 fair coins and count heads, what is the pmf?

$$\mathbb{P}(0) = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$\mathbb{P}(1) = \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 = \frac{5}{32}$$

$$\mathbb{P}(2) = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}$$

$$\mathbb{P}(3) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}$$

$$\mathbb{P}(4) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = \frac{5}{32}$$

$$\mathbb{P}(5) = \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

# Geometric distributions

Suppose a system runs a program at the end of each hour.

It crashes with probability $p$.

What is the expected time till the first crash?

If we let $C$ be the time till the first crash, then we want to find $\mathbb{E}\left(C\right)$.

Let $A$ indicate whether it crashes in the first hour. Then

$$\mathbb{E}\left(C\right) = \mathbb{E}\left(C|A\right)\mathbb{P}\left(A\right) + \mathbb{E}\left(C|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)$$

# Geometric distributions

$$\mathbb{E}\left(C\right) = \mathbb{E}\left(C|A\right)\mathbb{P}\left(A\right) + \mathbb{E}\left(C|\bar{A}\right)\mathbb{P}\left(\bar{A}\right)$$

We have that

- $\mathbb{E}\left(C|A\right) = 1$
- $\mathbb{P}\left(A\right) = p$
- $\mathbb{P}\left(\bar{A}\right) = 1 - p$

# Geometric distributions

$$\mathbb{E}\left(C\right) = p + \mathbb{E}\left(C|\bar{A}\right)(1-p)$$

For $\mathbb{E}\left(C|\bar{A}\right)$:

- it will not crash for one hour
- then we are back to the original question!
- so this is equal to $1 + \mathbb{E}\left(C\right)$

# Geometric distributions

So, all together:

$$\begin{aligned}
\mathbb{E}\left(C\right) &= \mathbb{E}\left(C|A\right)\mathbb{P}\left(A\right) + \mathbb{E}\left(C|\bar{A}\right)\mathbb{P}\left(\bar{A}\right) \\
&= p + \mathbb{E}\left(C|\bar{A}\right)(1-p) \\
&= p + (1 + \mathbb{E}\left(C\right))(1-p) \\
&= p + (1-p) + \mathbb{E}\left(C\right) - \mathbb{E}\left(C\right)p \\
&= 1 + \mathbb{E}\left(C\right) - \mathbb{E}\left(C\right)p
\end{aligned}$$

And therefore:

$$\begin{aligned}
\mathbb{E}\left(C\right)p &= 1 \\
\mathbb{E}\left(C\right) &= 1/p
\end{aligned}$$

# Geometric distributions

A random variable $R$ has a geometric distribution with parameter $p$ if

- its codomain is $\mathbb{Z}^+$ (nonnegative integers)
- $\mathbb{P}\left(R = i\right) = (1-p)^{i-1}p$

If the probability of some event is (independently) reached with probability $p$ at each step, then the expected number of steps until we see that event is $1/p$.

Example: let's roll a six-sided die.

How many rolls should we expect it takes to get a 5?

Let $C$ be the number of rolls till we see a 5.

Let $A$ indicate whether we get a 5 on the first roll.

$$
\begin{aligned}
\mathbb{E}(C) &= \mathbb{E}(C|A)\,\mathbb{P}(A) + \mathbb{E}\left(C|\bar{A}\right)\mathbb{P}\left(\bar{A}\right) \\
&= p + (1 + \mathbb{E}(C))(1-p) \\
&= \frac{1}{6} + (1 + \mathbb{E}(C))\frac{5}{6} \\
&= \frac{1}{6} + \frac{5}{6} + \frac{5}{6}\mathbb{E}(C) \\
&= 1 + \frac{5}{6}\mathbb{E}(C)
\end{aligned}
$$

# Geometric distributions

$$\mathbb{E}\left(C\right) = 1 + \frac{5}{6}\mathbb{E}\left(C\right)$$

So we have

$$\begin{aligned} \frac{1}{6}\mathbb{E}\left(C\right) &= 1 \\ \mathbb{E}\left(C\right) &= 6 \end{aligned}$$

# Sums of indicator random variables

Suppose $n$ people throw their hats in the air and then catch one. What's the expected number of people who get their own hat back?

Let $G$ be the number of people who get their own hat back.

Let $G_i$ be an indicator random variable that the $i$th person gets their hat back (which happens with probability $1/n$).

So $G = G_1 + G_2 + \cdots + G_n$.

# Sums of indicator random variables

Suppose $n$ people throw their hats in the air and then catch one. What's the expected number of people who get their own hat back?

Then

$$\begin{aligned}\mathbb{E}\left(G\right) &= \mathbb{E}\left(G_1 + G_2 + \cdots + G_n\right) \\ &= \mathbb{E}\left(G_1\right) + \mathbb{E}\left(G_2\right) + \cdots + \mathbb{E}\left(G_n\right) \\ &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} \\ &= 1\end{aligned}$$

## Sums of indicator random variables

Given events $A_1$, $A_2$, ..., $A_n$, how many do we expect to occur?

Let $R$ be the number that occur, and $R_i$ the indicator random variable for event $A_i$. Then

$$\begin{aligned}
\mathbb{E}\left(R\right) &= \sum_{i=1}^{n} \mathbb{E}\left(R_i\right) \\
&= \sum_{i=1}^{n} \mathbb{P}\left(R_i = 1\right) \\
&= \sum_{i=1}^{n} \mathbb{P}\left(A_i\right)
\end{aligned}$$

# Expectation of a binomial distribution

Recall the pmf for a binomial distribution. For example, let $J$ be the number of heads if we flip a coin $n$ times, where $p$ is the probability of heads and $q$ that of tails:

$$\mathbb{P}\left(J = k\right) = \binom{n}{k} p^k q^{n-k}$$

What is $\mathbb{E}\left(J\right)$?

# Expectation of a binomial distribution

By definition,

$$\begin{aligned}\mathbb{E}\left(J\right) &= \sum_{k=0}^{n} k\mathbb{P}\left(J = k\right) \\ &= \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}\end{aligned}$$

But we can simplify this...

# Expectation of a binomial distribution

Let $J_i$ be the indicator random variable that the $i$th coin flip is heads.

Then $J = J_1 + J_2 + \cdots + J_n$. And

$$
\begin{aligned}
\mathbb{E}\left(J\right) &= \sum_{k=1}^{n} \mathbb{P}\left(J_i\right) \\
&= pn
\end{aligned}
$$

Suppose every time you go to office hours, you get a ticket of some color.

If you collect all $n$ colors, you get a 4.0!

For example, this sequence would be sufficient ($n = 5$):

blue green green red blue orange blue orange gray

In general, how many office hours would you expect to have to go to?

# Coupon collector problem

Let's divide this into segments where the last coupon of each segment is a new color:

$$\underbrace{\text{blue}}_{X_0}\ \underbrace{\text{green}}_{X_1}\ \underbrace{\text{green red}}_{X_2}\ \underbrace{\text{blue orange}}_{X_3}\ \underbrace{\text{blue orange gray}}_{X_4}$$

Suppose there are $n$ colors, and let $X_k$ be the length of the $k$th segment.

# Coupon collector problem

Then the total number of coupons we collect is $T = X_0 + X_1 + \cdots + X_{n-1}$.

By linearity of expectation:

$$\begin{aligned}\mathbb{E}\left(T\right) &= \mathbb{E}\left(X_0 + X_1 + \cdots + X_{n-1}\right) \\ &= \mathbb{E}\left(X_0\right) + \mathbb{E}\left(X_1\right) + \cdots + \mathbb{E}\left(X_{n-1}\right)\end{aligned}$$

# Coupon collector problem

What is $\mathbb{E}(X_k)$?

At that point, we have collected $k$ colors.

- probability that the next coupon is a duplicate color is $k/n$

Thus the chance it's new is $1 - \frac{k}{n} = \frac{n-k}{n}$.

So $\mathbb{E}(X_k) = \frac{n}{n-k}$ (geometric distribution)

## Coupon collector problem

Finally,

$$
\begin{aligned}
\mathbb{E}\left(T\right) &= \mathbb{E}\left(X_0\right) + \mathbb{E}\left(X_1\right) + \cdots + \mathbb{E}\left(X_{n-1}\right) \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \cdots + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + \frac{1}{1}\right) \\
&= n\left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right)
\end{aligned}
$$

# Coupon collector problem

For $n = 5$, we have

$$5\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\right) \approx 11.42$$

Another example: how many times would you expect to have to roll a die before you get all 6 possible values?

$$6\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}\right) = 14.7$$