For example:

- Toss a fair coin 10 times, count the number of heads
- Plant 20 seeds, let X record the number that germinate

These are all essentially the same problem and there are probability models to generalize this

<u>Discrete Uniform Distribution</u>

If a random variable X takes on some values x={a, a+1, a+2, …, b}, where *each value is equally likely*, then X is said to have a *discrete uniform distribution* on {a, a+1, a+2, …, b}

- Can also write X ~ DU(a,b)

range of values [a, b]

"~" : "is distributed"

Since probabilities are equally likely between a and b, f(x) is constant for all integers in (a,b)

Then:

$$f(x) = P(X = x) = \frac{1}{\underbrace{b-a}_{\text{difference between max and min}}+1}$$

to properly count # of stacks between [a, b]

Ex. Toss a fair die once.

x	1	2	3	4	5	6
$F(x)$	$1/6$	$2/6$	$3/6$			

...

So $F(x) = \frac{x}{6}$

$$f(x) = \frac{1}{6-1+1}$$

Example: Let Y be a discrete uniform r.v. with a=0 and b=6.

Calculate the probability that the roots of the equation $g(z) = 0$ are real when:

$$g(z) = \underbrace{0.75}_{a}z^2 + \underbrace{3Y}_{b}z + \underbrace{5Y + 8}_{c}.$$

By the quadratic formula, real roots -> b^2 - 4ac ≥ 0

$$\Rightarrow 9Y^2 - (4)(0.75)(5Y+8) \geq 0$$
$$\Rightarrow 9Y^2 - 15Y - 24 \geq 0$$

Doing trial and error (or factoring), the expression is greater than 0 when Y≥3

Discrete uniform: the probability that an x from (0,6) is chosen is

$$\frac{1}{6-0+1} = \frac{1}{7}$$

4 ways to choose a Y that yields real roots (Y=3,4,5,6)

$$\Rightarrow P(Y \geq 3) = \frac{4}{7}$$

Hypergeometric Distribution

- Population of N objects
- r "success"-type objects
- (N-r) "failure" type-objects
- Pick n objects (n ≤ N) without replacement : picking each n not independant
- If a random variable X represents the number of success-type objects, then X has a *hypergeometric distribution*.
    - X ~ HG(N,r,n)

Ex.

- 120 applicants competing for a job
- 80 qualified
- 5 selected for interview
- Let X = number of qualified applicants that are interviewed

Then:

- N = 120
- r = 80 (we are sampling from r)
- n = 5
- Possible values of X = {0,1,2,3,4,5}

$\binom{N}{n}$ arrangements in total

$\binom{r}{x}$ ways to pick x qualified applicants

$\binom{N-r}{n-x}$ ways to fill in the rest of the group n

$$\Rightarrow f(x) = \frac{\binom{N-r}{n-x}\binom{r}{x}}{\binom{N}{n}}$$

where $x \geq \max\{0, n-(N-r)\}$ and $x \leq \min\{r, n\}$.

Probability that two of the five selected are qualified:

$$f(x) = \frac{\binom{40}{5-x}\binom{80}{x}}{\binom{120}{5}} \rightarrow P(X=2) = \frac{\binom{40}{5-2}\binom{80}{2}}{\binom{120}{5}}$$

In the game of *Texas Hold'Em*, each player is dealt two private cards, and five community cards are dealt face up on the table. Each player is to make the best 5-card hand they can with their two private cards and the five community cards.

What is the probability that a particular player will have a spade flush? (i.e. 5 or more spades in this case)?

We want P(player has a spade flush) $= P(X=5) + P(X=6) + P(X=7)$

Let X = number of spades

Sampling without replacement; two categories (spade / not spade) -> *hypergeometric*

So X ~ HG(N=52, r=13, n=7)

$$f(x) = P(X=x) = \frac{\binom{39}{7-x}\binom{13}{x}}{\binom{52}{7}}$$

<u>Binomial Distribution</u>

Two possible outcomes: success (S) and failure (F)

- P(S) = p (0 < p < 1)
- P(F) = 1-p
- Repeat the experiment n times. (Trials are *independent events* – one doesn't affect the other)
    - Individual trials referred to as "Bernoulli trials" - either succeed or fail
- Let X = number of successes observed

Write X ~ Bin(n, p)

Underlying assumptions:

- **T**wo outcomes
- **I**ndependent trials
- **M**ultiple trials
- **S**ame probability of success in each trial

Example: coin flipping n times

- Two outcomes (heads/tails), independent trials, multiple trials, same probability each outcome

Ex. A fair coin is tossed 12 times. Let X represent the number of heads obtained. Then X ~ Bin(n,p)

Probability function:

- x successes
- n-x failures
- Total number of arrangements of successes and failures: $\binom{n}{x}$

f(x) is defined as the probability of getting x successes

$$f(x) = \underbrace{\binom{n}{x}}_{\text{\# ways to get } x \text{ successes}} \cdot \underbrace{p^x}_{x \text{ successes, } p \text{ chance each time}} \cdot \underbrace{(1-p)^{n-x}}_{n-x \text{ failures, } (1-p) \text{ chance each time}}$$

Ex. 75% of the students at a college with a large student population use Instagram. A sample of five students from this college is selected. What is probability that at least 3 students use Instagram?

Hypergeometric would work *if* we knew the population of the college

- X ~ (N=?, n = 5, r = 0.75N, X = number of students in sample that use Instagram)

So X ~ Bin(n=5, p=0.75, X = number of students in sample that use Instagram)

P(X ≥ 3) = P(X=3) + P(X=4) + P(X=5)

$$= \binom{5}{3}0.75^3 \cdot 0.25^2 + \binom{5}{4}0.75^4 \cdot 0.25 + \binom{5}{5}0.75^5 \cdot 0.25^0$$

$$= 0.8965$$

Binomial Distribution vs. Hypergeometric Distribution

Similarities:

1. Both have 2 types of outcomes; success and failure.
2. The experiment is repeated *n* times in each.
3. The r.v. *X* records the number of successes.

The Main Difference:

- The binomial distribution requires *n* INDEPENDENT trials, where the probability of success is the *same* in each trial (TIMS).
- In the hypergeometric setting, the *n* draws are made from a fixed number of objects (*N*) WITHOUT replacement. Hence, the trials are NOT independent.

Ex. Suppose we have 50 bottles of drinks placed in a big ice container such that the labels are not visible. It is known that 20 are energy drinks and 30 are soda pop. 5 cans are randomly selected. Find the probability that 3 are energy drinks.

Hypergeometric:

- Finite population (50)
- Sampling *without* replacement - remove 5 cans; depending on what was removed, the probability of getting an energy drink/soda pop changes. So each trial is *not* independent

Let X = number of energy drinks

So X ~ HG(N=50, r=20, n=5). We want P(X=3)

$$P(X=3) = \frac{\binom{30}{5-3}\binom{20}{3}}{\binom{50}{5}}$$

Now, if selections are done *with* replacement:

X ~ Bin(n=5, p=20/50)

We can use a binomial distribution to *approximate* a hypergeometric probability

- For example, if the population size is 1,000,000 and we are taking a sample of size 10, whether we sample with or without replacement won't matter much
- But if the population size is small, this probably won't work

Ex. In a recent shipment of 5000 tires to the ABC Tire company, 1000 of them are slightly marked. Ten tires from the shipment are randomly chosen and purchased by a consumer.

(a) What is the exact probability that 3 of the 10 tires are slightly marked?

Let X = the number of tires that are slightly marked

X ~ HG(N=5000, r=1000, n=10)

We are solving for P(X=3):

$$P(X=3) = \frac{\binom{4000}{7}\binom{1000}{3}}{\binom{5000}{10}} = 0.20148$$

(b) Using a suitable approximation, what is the *approximate probability* that 3 of the 10 tires purchased are slightly marked?

X ~ Bin(p=0.2, n=10) $\quad 0.2 : 1000/5000$ tires

$$P(X=3) = \binom{10}{3} \cdot (0.2)^3 \cdot (0.8)^7 = 0.20133$$

vs. 0.20148
(very close)

Negative Binomial Distribution

- Independent/Bernoulli trials
- Two outcomes
- p: probability of success in each trial (constant from trial to trial)
- Repeat the experiment until *k* successes are obtained

Let X = number of failures before the k-th success. The range of X is countably infinite.

X ~ NB(k, p)

$$\underbrace{\_\_\ \ \_\_\ \ \_\_\ \ \_\_\ \cdots\ \_\_}_{k-1 \text{ successes},\ x \text{ failures}}\quad \_\_$$

↑ success

$\Rightarrow x+k-1$ trials

$\Rightarrow$ # arrangements: $\binom{x+k-1}{x}$

So the probability function is

$$\binom{x+k-1}{x} p^k \underbrace{(1-p)^x}$$

probability of x failures before the kth success

↳ probability of k successes

Ex. Draw cards from a standard deck of 52 cards with replacement until you get 3 Aces. Let X represent the number of non-Aces that appear before the 3rd ace is obtained.

X ~ NB(k=3, p=1/13)

Another way to represent the negative binomial distribution is to use a different random variable Y, representing the total number of *trials* (not *failures*, like X) needed to get the k-th success

So Y = X + k where X ~ NB(k,p)

The probability function of Y is

$$f(y) = P(Y = y) = \binom{y-1}{k-1} p^k \cdot (1-p)^{y-k}$$

↳ $x = y - k$ failures

↳ k successes

↳ k − 1 successes in y − 1 trials

Ex. A start-up company is looking for 5 investors. Each investor will independently agree to invest in the company with probability 20%. The founder asks investors one at a time until 5 "yes" responses are obtained.

Let X = the total number of investors asked. (using the Y form)

$$f(x) = \binom{x-1}{4} (0.2)^5 \cdot (0.8)^{x-5}$$ → x trials, 5 successes, x − 5 failures

↳ x − 1 trials before final success

4 successes before final success

Geometric Distribution

- Two outcomes
- Independent trials, each with probability of success p
- Repeat the experiment until **1** success is obtained

Let X = number of failures before first success

This is a special case of the negative binomial distribution where k=1

Ex. Pascal is a 70% free throw shooter. The number of misses before he makes his first shot can be modeled by a geometric distribution, with X = the number of failures before his first success.

There is only one way to arrange x failures before the first success, so:

$$f(x) = (1-p)^x p$$

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} (1-p)^k p$$

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

Ex. Suppose that a company receives 60% of its orders over the internet.

(a) What is the probability that the fifth order received is the first internet order?

We want P(X=4) - 4 failures before the first success

$$\Rightarrow (0.4)^4 \cdot 0.6$$

(b) What is the probability that the eighth order received is the fourth internet order?

Here, we use a negative binomial distribution, where Y = non-internet orders received until the *fourth* internet order

Y ~ NB(p=0.6, k=4)

$$f(y) = \binom{7}{4}(0.4)^4(0.6)^4$$

7 trials before 4th success
4 failures
$\binom{7}{4}$ ways to arrange

(c) What is the probability that more than 3 total orders are required to get the first internet order?

At least 3 fails

We are looking for P(X ≥ 3) = 1 - P(X ≤ 2) = 1 - F(2)

$$= 1 - (0.4)^{x+1}$$
$$= 1 - (0.4)^3$$

<u>Poisson Distribution</u>

Let X = number of events of some type

The events occur according to some rate μ > 0

Write X ~ Poisson(μ)

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Examples of things that follow a Poisson distribution:

- Number of misprints on a page
- Number of people in a community who survive to age 100
- Number of lightning strikes in a region of Canada in a month

The Poisson distribution has connections to the binomial distribution:

- Binomial: number of successes in n trials, each with p probability
- Poisson: average rate μ = np — on average, μ successes in a sample of n

For example, if we flip a coin 50 times, the average rate of heads is (50)(0.5) = 25

As a limiting case of the binomial distribution

$$p = \frac{\mu}{n}$$

$$\Rightarrow f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)}}{x!}\left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \frac{\mu^x}{x!}\left(\frac{n^{(x)}}{n^x}\right)\left(1-\frac{\mu}{n}\right)^{n-x}$$

$$\lim_{n\to\infty}: \quad = \frac{\mu^x}{x!}\underbrace{\left[\frac{(n)(n-1)\cdots(n-x+1)}{n^x}\right]}_{\substack{\to 1 \text{ as } n\to\infty \\ \text{since numerator has} \\ x \text{ terms}}}\left(1-\frac{\mu}{n}\right)^n \underbrace{\left(1-\frac{\mu}{n}\right)^{-x}}_{\to 1 \text{ as } n\to\infty}$$

$$\lim_{n\to\infty}\left(1+\frac{k}{n}\right)^n = e^k$$

$$= \frac{\mu^x e^{-\mu}}{x!}$$

Ex. Let X = number of people born on January 1 in a group of 200

We want P(X=2)

X ~ Bin(n=200, p=1/365)

$$f(x) = P(X=2) = \binom{200}{2}\left(\frac{1}{365}\right)^2\left(\frac{364}{365}\right)^{198} = 0.08\overline{67}$$

Approximating using Poisson:

μ=np = 200/365

There are, on average, 200/365 people born on January 1 in some *continuous* population

(But the population is not continuous so Poisson sucks here)

Ex. A local restaurant is running a contest. A customer receives a ticket each time they purchase a combo. They claim that 1 in 9 tickets are winners. Say you buy 100 combos!

Assuming that the trials are independent, let's use the Poisson approximation to the binomial to solve for the probability that you get no more than 10 winning tickets.

In R, we can use pbinom(10, 100, 1/9) -> binomial distribution yields the exact answer

Approximation:

$X \sim \text{Poisson}(\mu=100/9)$

In R: ppois(10, 100/9)