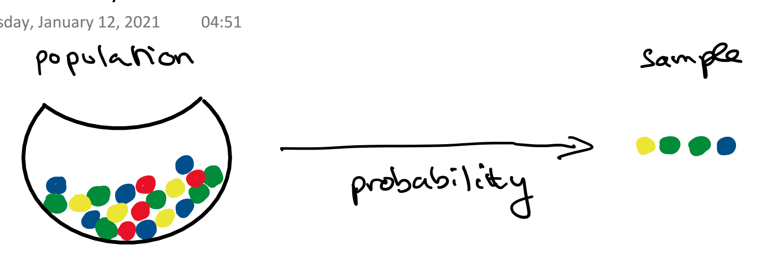
Probability

Tuesday, January 12, 2021 04:51



Probability reasons about a sample, knowing characteristics of the population

Statistics reasons about characteristics of a population, knowing a sample.

Probability:

A probability of an event is the expected rate the event will appear were you to sample a large number of times.

Distribution:

A population is best characterized by its distribution: the probability associated with each possible subset. Typically expressed as a function of outcomes/events that returns probabilities.

Example: Flipping a fair coin

$\Omega = \{H, T\}$ → sample space (all possible outcomes)

$\mathcal{F} = 2^{\Omega} = \{\phi, \{H\}, \{T\}, \{H,T\}\}$ → event space

$P : \mathcal{F} \to [0, 1]$ → probability measure

where
$$P(\Omega) = 1$$
$$P(A) \geq 0 \quad \forall A \in \mathcal{F}$$
axioms
$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ for } A_i \in \mathcal{F} \text{ s.t. } A_i \cap A_j = \emptyset \text{ for } i \neq j$$
disjoint

$(\Omega, \mathcal{F}, P)$ prob. space.

From these we can deduce:

$$P(A^c) = 1 - P(A) \quad \text{since } P(A \cup A^c) = P(\Omega)$$
$$P(A) \leq 1 \quad \text{since } P(A^c) \geq 0$$
$$P(\phi) = 0 \quad \text{since } P(\Omega) = 1$$

And a few other properties:

1) $P(A \cup B) = P(A) + P(B) - P(A \cap B) \underbrace{\leq}_{\text{union bound}} P(A) + P(B)$

2) $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \underset{\text{Bayes Rule}}{=} \frac{P(B \mid A) P(A)}{P(B)}$

Definition: Events A & B are independent iff

$$P(A \cap B) = P(A) P(B)$$

or, equivalently, iff

$$P(A \mid B) = P(A)$$

Going back to our coin flip example where

$$\Omega = \{H, T\} \text{ and } \mathcal{F} = \{\phi, \{H\}, \{T\}, \{H,T\}\}$$

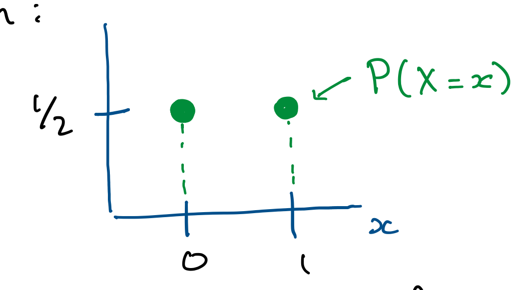
We define a random variable $X : \Omega \to \mathbb{R}$

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \{H\} \\ 0 & \text{if } \omega = \{T\} \end{cases}$$

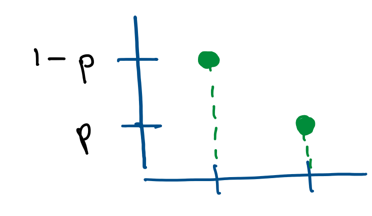
So $P_X(1) = P(X = 1) = P(\{\omega \mid X(\omega) = 1\}) = P(\{H\})$

Note: If Ω is a group of people, $X : \Omega \to \mathbb{R}$ might return, for each person, their height, their age, how much they like a product etc.

So the distribution of X can be plotted as such:



Extending this to an unfair coin where $P(\{H\}) = p \Rightarrow P(\{T\}) = 1 - p$



This is usually called <u>Bernoulli Trial</u>.

Now suppose we flip the coin multiple times. We'll call X_1 the RV describing the first flip, X_2 the second etc.

$X_1, X_2, \ldots$ are identically distributed Bernoulli Trials. These are also independent RV since the result of one does not affect the result of another. This means

$$P_{X_1, X_2}(x_1, x_2) = P_{X_1}(x_1) P_{X_2}(x_2)$$

What is the probability that the first r flips are all Tails?

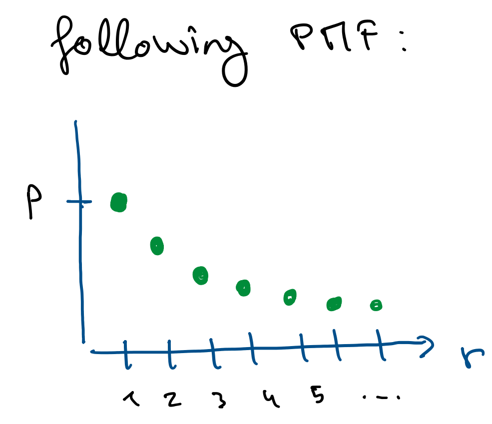
$$P_{X_1}(0) P_{X_2}(0) \ldots P_{X_r}(0) = (1-p)^r$$

What is the distribution of the number r of flips it will take to get the first H?

$$P_R(r) = (1-p)^{r-1} p$$

This is called the <u>Geometric Distribution</u>.

It has the following PMF:

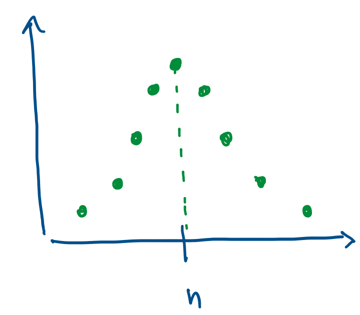


Suppose we now flip this coin n times. What is the distribution of the number k of Hs?

$$P_{K_n}(k) = \underbrace{\binom{n}{k}}_{\text{ways to get } k \text{ H's}} \underbrace{p^k (1-p)^{n-k}}_{k \text{ H's AND } (n-k) \text{ T's}}$$

This is called the <u>Binomial Distribution</u>.

It has the following PMF:

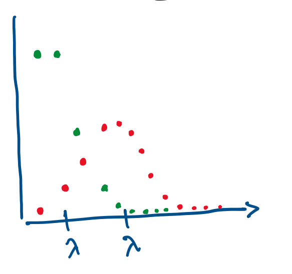


Now suppose, out of quarantine boredom, you start regularly flipping a coin over a long period of time. And, because you're so bored, you notice that on average there are about $\mu = 60$ H's per hour. What is the distribution of the number k of H's at any given minute?

$$P_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where } \lambda = \mu \cdot 1 \text{ min} = \frac{1}{\text{min}} \cdot \text{min} = 1$$

This is called the <u>Poisson Distribution</u>.

It has the following PMF:



Q: What is the probability that the interval between occurances is 2min?

Notice that $P(T = t)$ is based on a continuum of values t. Assuming that all intervals of time t are possible, what happens if $P(T = 2) = p > 0$? By continuity, $P(T = 2 + \varepsilon) \approx p > 0$. Recall that $P(\Omega) = 1$, but summing over a continuum of values means we necessarily will get $P(\Omega) = \infty$. The only way for $P(\Omega) < \infty$ is if $P(T = t) = 0$ but clearly a given t will occur – so what's going on?

The question asked as is doesn't make sense. Instead we must ask for the prob that $T > t$, or the probability that $T \leq t$ to get the <u>cummulative distribution function</u>

$$P(T > t) = e^{-\mu t}$$

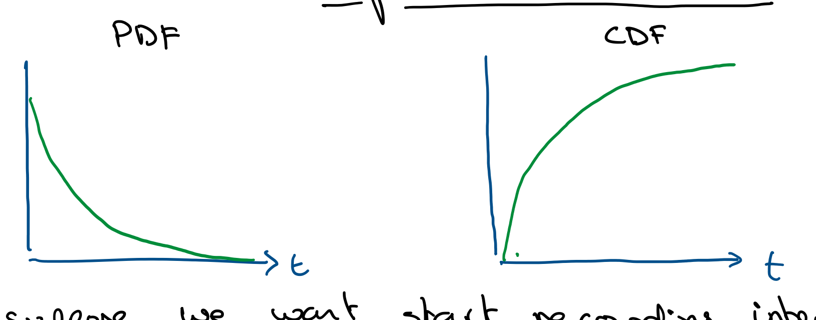
We can also ask for the probability that the time interval T will fall in infinitesimal range dt centered at t

$$f(t) = \mu e^{-\mu t}$$

This gives us the <u>probability density function</u>. Because the PDF gives us the instantaneous change in probability, we can get the probability that $t \in [t_1, t_2]$ by integrating the above pdf:

$$P(t \in [t_1, t_2]) = \int_{t_1}^{t_2} \mu e^{-\mu t} dt$$
$$= P(t \geq t_1, t \leq t_2)$$
← so we can also get this from the CDF.

This is called the <u>Exponential Distribution</u>



Now suppose we want start recording interval lengths to get a sense of the average interval length. Call t_1 the first interval, t_2 the second etc. Let

$$\bar{t}_n = \frac{t_1 + \cdots + t_n}{n}$$

be the mean after n intervals. What is the distribution of $\bar{t}_n$?

$$f(x) = \frac{1}{\sigma \sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $\mu = 1/\lambda$ → mean of the times
$\sigma = 1/\lambda$ → standard deviation of the times

This is called the <u>Normal Distribution</u> or the <u>Gaussian Distribution</u>. The phenomenon that the sample average follows a normal distribution is called the <u>Central Limit Theorem</u>.

Characteristics of Distributions:

<u>Expectation</u> ("mean")

$$\mathbb{E}[X] = \begin{cases} \sum_x x p(x) \\ \int_{-\infty}^{\infty} x f(x) dx \end{cases}$$

<u>Properties:</u>

1) $\mathbb{E}[aX] = a\mathbb{E}[X]$
2) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3) $\mathbb{E}[c] = c$

<u>Variance:</u>

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

what is the average distance to the mean? Gives an idea of how spread out the distribution is.

<u>Properties:</u>

1) $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2) $Var(aX) = a^2 Var(X)$
3) $Var(X + c) = Var(X)$