Definitions and terms to know - part 3

Notation for random variables

Often it is useful to suppose that some real world phenomena can be modeled as a random process. So we might say, for example, let X be a random variable that represents the height of all people ages 18-25. Then we'd have to decide what distribution makes the most sense for this random variable. E.g. Is a Normally distributed random variable ok? Are there some numbers that are not possible to observe?

Typically we denote random variables with capital letters from the end of the English alphabet: X, Y, W, Z. If the distribution of a random variable is one of the more common ones (like those listed below) we'll use the short-hand $\sim$ following the random variable which means "is distributed as".

Sometimes we talk about two random variables with the same distribution. In this case, we generally use the same letter but index the random variables to be able to differentiate among them: X_1, X_2, X_3.

The **sample space** of a random variable is the set of all possible values that is could possibly take. Once the random variable takes on a particular value, we denote that value with the lowercase version of the same letter. If we have multiple observations of the same random variable, we index these observations to distinguish among them. The last observation has the index n, where n corresponds to the total number of observations, i.e. the sample size.

Often we may need to consider a (simple) **random sample** from the same population (modeled by some random variable). A random sample is always drawn so that each observation is independent of another and so that all observations are drawn from the same population. We call this an "independent and identically distributed" (often abbreviated **IID**) sample or just a "random" sample.

Common random variables: discrete

Bernoulli

- **Intitutive definition:** For binary categorical variables. Represents a single success or failure. (Leave all preconceived notions of success at the door, it has no positive connotation in statistics.) E.g. a single flip of a coin can either result in heads or tails. (Whichever outcome we consider a "success" is up to us to decide.)
- **Notation:** $X \sim Bern(p)$
- **R functions:** `rbinom(n, size=1, prob)` Note a Binomial random variable (below) is just the sum of the number of success of n (independent) Bernouli random variable.
- **Distribution parameters:** p the probability of a "success"
- **Sample space:** $S = \{0, 1\}$ or $S = \{True,\ False\}$ or $S = \{Positive,\ Negative\}$

Binomial

- **Intuitive definition:** For binary categorical variables. Represents the number of successes or failures out of some fixed number of different trials or (possibly biased) coin flips. E.g. If we consider "heads" as a "success" and flip a coin several times, then the total number of heads follows a binomial distribution. It's important to keep in mind that the coin we're flipping n times is the same coin and that the outcome of the first flip is *independent* of the outcome of the second flip, and so on.
- **Notation:** $X \sim Bin(n, p)$ Here n is the number of trials, or, flips of "a coin" so n is the largest possible number of successes we could observe.
- **R functions:** `rbinom(n, size, prob)` In this R function, n actually represents the number of *size*-length coin flips performed in total and *prob* is the probability of a single success in a single flip/trial. (So in R, *size* is the largest possible number of successes we could observe, which confuses the conventional notation mentioned about.)
- **Distribution parameters:** $p =$ the probability of a single success, $n =$ the number of trials/coin flips
- **Sample space:** $S = \{0, 1, 2, \ldots, n\}$

Multinomial

- **Intuitive definition:** For categorical variables in general. As the binomial distribution models the number of heads (say) from a sequence of n coin tosses, the multinomial distribution models the different results of tossing a k-sided die n times. Here we also assume that the $1, \ldots, k$ possible outcomes for each of the n rolls are independent of one another.
- **Notation:** $X \sim MN(n, p_1, p_2, \ldots, p_k)$ or $X \sim Multi(n, p_1, p_2, \ldots, p_k)$
- **R functions:** `rmultinom(n, size, prob)` where, similar to the confusing `rbinom()` function, n represents the number of rolls/tosses of length *size* we want to simulate. *size* is an integer that specifies the number of times the die is rolled, and *prob* is a **vector** of k different numbers that represent the probability of landing on each one of the k different die faces.
- **Distribution parameters:** $n =$ the number of trials/rolls of the die, $p_1 =$ the probability of landing on the first die face, $p_2 =$ the probability of landing on the second die face, etc.
- **Sample space:** The sample space S is a bit difficult to visualize (since it's multidimensional) but it consists of a finte number of elements. Each element represents, over these n rolls, how many times did we observe 1, the first face of our die, how many times did we observe 2, the second face of our die, ..., how many times did we observe k, the last face of our die.

Common random variables: continuous

Uniform

- **Intuitive defintiion:** Randomly choosing any real number (rational or irrational) between two specific numbers.
- **Notation:** $X \sim U(a, b)$ or sometimes $X \sim Unif(a, b)$ (note parentheses are used because of convention but really these are inclusive of the numbers a and b so sometimes I'll write $X \sim U[a, b]$)
- **R functions:** `runif(n, min, max)` where n is the size of the sample you want to obtain, *min* is the lower bound of the interval, and *max* is the upper bound of the interval.
- **Distribution parameters:** a and b represent the smallest possible value that could be observed and the largest possible value that could be observed, respectively.
- **Sample space:** $S = [a, b]$

Normal/Gaussian

- **Intuitive definition:** For continuous quantitative variables that tend to follow a bell shaped curve. (Sometimes used even if the variable being considered can't be negative or if it's actually a discrete quantitative variable because of the Central Limit Theorem.)
- **Notation:** $X \sim N(\mu, \sigma)$ or sometimes $X \sim N(\mu, \sigma^2)$
- **R functions:** `rnorm(n, mean, sd)` where n is the size of the sample you wish to obtain, *mean* is the mean of the distribution μ, and *sd* is the standard deviation of the distribution, σ
- **Distribution parameters:** μ the population mean and σ the population standard deviation
- **Sample space:** $S = (-\infty, \infty)$

Chi-squared

- **Intuitive definition:** For continuous quantitative variables that can only take on positive numbers.
- **Notation:** $X \sim \chi^2_{(v)}$ (the symbol on the right hand side is the Greek letter *chi*)
- **R functions:** `rchisq(n, df)` where n is the size of the sample you wish to obtain and *df* is the population parameter the degrees of freedom.
- **Parameters:** v = degrees of freedom
- **Sample space:** $S = (0, \infty)$

Important facts and theorems

Two special facts about Normal random variables

1) If X is Normally distributed and Y is Normally distributed, then $aX + bY + c$ (a, b, c are any real numbers) is also Normally distributed.

Generally, if we have two random variables with the same distribution, say $X_1 \sim X_2$, the new random variable formed by adding or subtracting these two (e.g. $W = X_1 + X_2$) does **not** follow the same distribution as X_1 and X_2. In some special cases however, W may follow the same distribution as X_1 and X_2. **One special case** is if two random variable are Normally distributed then any linear combination of these two Normal random variables will also be a Normally distributed random variable (but with a different mean and variance).

2) The mean and variance completely and uniquely specify a Normal distribution.

The only two **parameters** we need to know about anything that is Normally distributed are the mean and the variance (or, equivalently, the mean and the standard deviation). For other distributions the model parameters needed to uniquely define the distribution may be something else, e.g. the Chi-squared distribution is defined according to a parameter called the *degrees of freedom*.

The Law of Large Numbers

"The relative frequency interpretation of probability is that if an experiment is repeated a large number of times [read "an infinite number of times"] under identical conditions and independently, then the relative frequency with which an event A actually occurs and the probability of A should be approximately the same."[1]

The Central limit theorem

If n is large [read "as n gets larger and larger and closer to infinity"], the standardized average ($\bar{z}$) of a random sample ($\{x_1, x_2, \ldots, x_n\}$) will itself be Normally distributed, regardless of the original distribution of X. Specifically, if X is a random variable with **any** distribution (e.g. Binomial, Uniform, Normal) and if we observe an infinitely large number of observations of X, $\{x_1, \ldots, x_n\}$, then for the standardized version of these observations, say $\{z_1, z_2, \ldots, z_n\}$ we have that

$$\bar{z} \sim N(0, 1).$$

This theorem is a key concept because it implies that statistical methods that work for Normal distributions can be applicable to many problems involving other types of distributions.[1,2]

---

[1] Encyclopedia Brittanica https://www.britannica.com/science/probability-theory/An-alternative-interpretation-of-probability#ref32778"

[2] Wikipedia https://en.wikipedia.org/wiki/Central_limit_theorem