STA237 Week 3 to 5 Recap

Pranav Rao

October 10, 2023

1 Random Variables

- A **random variable** is a variable whose value is unknown or assigns values to each of an experiment's outcomes
- Two types of random variable:
  - **Discrete:** A random variable that takes on only countable values ($X = 0, 1, 2, 3$)
  - **Continuous:** A random variable that can take on an infinite number of values ($X \in [1, 5]$)

2 Distribution Functions

2.1 Definition

- A **distribution** is a function that shows the possible values for a variable and how often they occur.

2.2 PMFs vs PDFs vs CDFs

For each distribution, there are two main functions that can be used to describe it.

- The first function we can use to describe a distribution changes depending on the type of variable the distribution is based on. For distributions with based on a:
  - *Discrete variable*: the first function is called a **probability mass function** (PMF for short). From Wikipedia, a PMF is a "function that gives the probability that a discrete random variable is exactly equal to some value."
  - *Continuous variable*: the first function is called a **probability density function** (PDF for short, not to be confused with the filetype). This function is like a PMF, but it will be continuous because it is based on a continuous variable.

- The second function used to describe a distribution has the same name for both type of functions. This function is called a **cumulative distribution function** (CDF). The CDF is a function such that, when evaluated at some x, it gives the cumulative probability that the random variable X will take a value less than or equal to X. Depending on the type of variable, the CDF is calculated differently:
  - For a discrete random variable X, the CDF $F_X(x)$ is calculated as such:

$$F_X(x) = P(X \leq x) = \sum_{y \leq x} P(X = y)$$

  - For a continuous random variable X, the CDF $F_X(x)$ is calculated as such:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)dy$$

3 Expected Values, Variance, Standard Deviation, and MGFs

3.1 Expected Values

- The **expected value** (also known as the **mean, expectation, average, etc.**), according to Wikipedia, is informally defined as "the arithmetic mean of a large number of independently selected outcomes of a random variable". The expected value, often represented as $E[X]$ or μ, can be calculated as such:
  - For a *discrete* random variable X, given that the PMF of X is $p(a)$ for some value a:

$$E[X] = \mu = \sum_i a_i P(X = a_i) = \sum_i a_i p(a_i)$$

  - For a *continuous* random variable X, given that the PDF of X is $f(x)$ for some value x:

$$E[X] = \mu = \int_{-\infty}^{\infty} x f(x)dx$$

3.2 Variance and Standard Deviation

- *Intuitive definition of variance:* Intuitively, the **variance** of a random variable X measures how much the values of X tend to spread out or vary from the mean (average) value. Like expectation, the variance also has a symbol commonly associated with it, which is σ^2.

- *Mathematical definition of variance*: Mathematically, the **variance** of any random variable X with mean μ is defined as:

$$Var(X) = \sigma^2 = E[(X - \mu)^2]$$

- An alternative definition for variance (which is often easier to use in calculations) is:

$$Var(X) = \sigma^2 = E[X^2] - (E[X])^2$$

- *Definition of standard deviation*: The **standard deviation**, according to Wikipedia, is "measure of the amount of variation or dispersion of a set of values". The standard deviation, often represented as σ is calculated as the square root of the variance, namely:

$$\sigma = \sqrt{\sigma^2} = \sqrt{E[(X-\mu)^2]} = \sqrt{E[X^2] - (E[X])^2}$$

3.3 Moment-Generating Functions

- A **moment** (in statistics) is a way to quantify characteristics of a given probability distribution.
  - The first moment $E[X]$ is the *expected value.*
  - The second moment $E[X^2]$ can be used to calculate variance (see above).
  - Similarly, n'th moment $E[X^n]$ can help provide some other useful information.

- *Intuitive definition of MGF:* Intuitively, a **moment-generating** function (MGF) is a function that can be used to generate the function for a specific *moment* of a random variable X.

- *Mathematical definition of MGF:* Let X be a random variable with CDF F_X. Mathematically, the moment-generating function (denoted as $M_X(t)$) is calculated as:

$$M_X(t) = E[e^{tX}]$$

  We say that the MGF exists if there exists a positive constant a such that $M_X(s)$ is finite for all $[-a, a]$.

- The moment-generating function (provided the expectation exists for some t in a neighbourhood of 0) is calculated differently depending on the type of variable in question:
  - For a *discrete random variable* X with PMF p_X, the moment-generating function is defined as:

$$M_X(t) = E(e^{tX}) = \sum_k e^{tk} p_X(k)$$

- For a *continuous random variable* X with PDF f_X, the moment-generating function is defined as:

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$$

- We can get the n'th moment from a moment-generating formula $M_X(t)$ by taking n derivatives of the MGF and then evaluating at 0. That is to say:

$$E(X^n) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

4 Common Discrete Distributions

In statistics, there are common distributions that can be used to model certain types of events. Memorizing these distributions can make calculating probabilities for specific events a lot easier. This section will look at some of the most common distributions used in this course, their pre-calculated PMF, mean, variance, (i.e. what you would get if you tried to calculate them yourself) and their associated R functions (see the special section on R distribution functions later this document). NOTE: all of these random variables are discrete because we have not explicitly covered continuous distributions yet.

4.1 Discrete Uniform Distribution

A note: this distribution should be used sparingly; the continuous version of this appears to be far more common.

- *Use case*: where all of the n discrete outcomes are equally likely to occur
- *Example*: rolling a fair six-sided die
- *Notation*: $Unif(a, b)$, where a is the first discrete value and b is the last
- *PMF*: $1/n$, where n is the number of possible outcomes
- *Mean*: $\frac{a+b}{2}$, where a is the first discrete value and b is the last
- *Variance*: $\frac{n^2-1}{12}$, where n is the number of outcomes
- *Associated R functions*: None in the standard library.

4.2 Bernoulli Distribution

- *Use case*: where there are only two possible outcomes (one success, one failure)
- *Example*: flipping a fair coin (where $p = \frac{1}{2}$)
- *Notation*: $Bern(p)$, where p is the probability of the success occurring
- *PMF*: $\begin{cases} p \text{ if the it is the first outcome} \\ 1-p \text{ if the second outcome} \end{cases}$
- *Mean*: p, where p is the probability of the success occurring
- *Variance*: $p(1-p)$, where p is the probability of the success outcome occurring
- *Associated R functions*: `dbern, pbern, qbern, rbern`

4.3 Binomial Distribution

- *Use case*: where there are n Bernoulli trials run back to back (still only 2 outcomes, one success, one failure)
- *Example*: "what is the probability of getting exactly 3 heads if you flip a coin 5 times?" (where $n = 5$, $k = 3$, $p = \frac{1}{2}$)
- *Notation*: $Bin(n, p)$ or $B(n, p)$, where n is the number of trials and p is the probability of the success outcome occurring
- *PMF*: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where k is the number of times you are hoping the success outcome will occur, p is the probability of the success outcome on a single trial, and n is the number of trials
- *Mean*: np, where p is the probability of the success outcome occurring and n is the number of trials
- *Variance*: $np(1-p)$, where p is the probability of the success outcome occurring and n is the number of trials
- *Associated R functions*: `dbinom, pbinom, qbinom, rbinom`

4.4 Geometric Distribution

- *Use case*: when you want to figure out the probability of a success happening within the first k independent Bernoulli trials
- *Example*: "what is the probability that I will get a heads in the first two times I flip a fair coin?" (where $k = 2$, $p = \frac{1}{2}$)

- *Notation*: $Geo(p)$, where p is the probability of the success outcome occurring
- *PMF*: $P(X = k) = (1-p)^{k-1}p$ where p is the probability of success and k is the desired number of trials for the success event to occur
- *Mean*: $\frac{1}{p}$, where p is the probability of the success outcome occurring
- *Variance*: $\frac{1-p}{p^2}$, where p is the probability of the success outcome occurring
- *Associated R functions*: `dgeom, pgeom, qgeom, rgeom`

4.5 Negative Binomial Distribution

- *Use case*: when you want to figure out the probability that r successes appear in the first x independent Bernoulli trials
- *Example*: "what is the probability that, if I continuously flip a fair coin, I will get three heads within the first 5 trials" (where $r = 3$, $k = 5$, $p = \frac{1}{2}$)
- *Notation*: $NB(r, p)$, where r is the desired number of successes and p is the probability of the success event occurring
- *PMF*: $P(x = k) = \binom{k-1}{r-1}p^{r-1}(1-p)^{(k-1)-(r-1)}$, where p is the probability of the success outcome, r is the desired number of successes, and k is the number of trials within which you want to achieve r successes
- *Mean*: $\frac{r(1-p)}{p}$, where p is the probability of the success outcome, r is the desired number of successes
- *Variance*: $\frac{r(1-p)}{p^2}$, where p is the probability of the success outcome, r is the desired number of successes
- *Associated R functions*: `dnbinom, pnbinom, qnbinom, rnbinom`

4.6 Hypergeometric Distribution

- *Use case*: when you want to take a sample of size n from a combination of 2 groups (say, a "success" group of size b and the "failure" group), in which are there a total of N entities, without replacement (that is, these are not independent and therefore not Bernoulli trials), and you want to know the probability that, out of that sample, k people are part of the "success group"
- *Example*: "6 doctors and 19 nurses attend a small conference. If all 25 names are put in the hat and 5 names are randomly picked without replacement, what is the probability that 4 doctors and 1 nurse are picked?" (where $N = 25$, $n = 5$, $b = 6$, and $k = 4$)

- *Notation*: (I could not find a satisfactory answer for this, so I'm guessing) $H(N, b, n)$, where N is the total size of both groups, b is the size of the success group, and n is the sample size
- *PMF*: $P(X = k) = \frac{\binom{b}{k} \cdot \binom{N-b}{n-k}}{\binom{N}{n}}$, where N is the total size of both groups, b is the size of the "success" group, and k is the desired number of elements to be drawn from the success group
- *Mean*: $n \cdot \frac{b}{N}$, where n is the sample size, b is the size of the success group, and N is the size of both groups combined
- *Variance*: $n \cdot \frac{b}{N} \cdot \frac{N-b}{N} \cdot \frac{N-n}{N-1}$, where N is the total sample size, b is the size of the "success" group, and N is the size of both groups combined
- *Associated R functions*: `dhyper, phyper, qhyper, rhyper`

4.7 Poisson Distribution

- *Use case*: when you want to find to calculate the probability of number of events occurring in a fixed interval of space or time if you know those events occur with a known constant mean rate
- *Example*: "One nanogram of plutonium will have an average of 2.3 radioactive decays per second, and the number of decays follow a Poisson distribution. What is the probability that in a 2-second period, there are exactly 3 radioactive delays?" (where $k = 3$ and $\lambda = 2.3 \cdot 2 = 4.6$)
- *Notation*: $Pois(\lambda)$, where λ is the rate of occurrence
- *PMF*: $P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- *Mean*: λ, where λ is the rate of occurrence
- *Variance*: λ, where λ is the rate of occurrence
- *Associated R functions*: `dpois, ppois, qpois, rpois`

5 R Distribution Functions

- R follows a similar format for all of its functions that have to do with distributions.
- Each distribution function has four variations, which start with four prefixes: `d, p, q, r`. Here is what each one does:
  - `d` variation: returns the value of the **distribution's PMF** at the given parameters.

  - `p` variation: returns the value of the **distribution's CDF** at the given parameters.
  - `q` variation: returns the value of the **distribution's inverse CDF** at the given parameters.
  - `p` variation: returns a **vector of random variables**, distributed depending on the type of distribution.