7.6 Introduction to distributions

In general, random variables map events in the sample space to real numbers. In assigning numerical values to objects in the sample space, we can simplify the sample space substantially.

We may want to know what the probability is that this variable will take on certain values, or certain intervals of values. This is known as the variable's distribution. We can consider this for discrete or continuous random variables.

7.7 PDF and CDF

Two generally useful things to know about a distribution are its probability mass function (PMF) or probability density function (PDF) and its cumulative distribution function (CDF).

PMF: For a discrete random variable, the PMF is a function that gives the probability that a discrete random variable is exactly equal to some value.

$$f_X(x) = Pr(X = x)$$

PDF: For a continuous random variable, the PDF is a function, whose value at any point in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable falls in a particular range (given by integral of PDF over that range). The PDF is the derivative of the CDF.

$$P[x \leq X \leq b] = \int_a^b f(x)dx$$

CDF: The CDF of X, evaluated at x, is the probability that X will take a value less than or equal to x. For a a continuous distribution, it gives the area under the probability density function from $-\infty$ to x.

$$F_X(x) = P(X \leq x)$$

For a continuous random variable, this can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

7.8 Binomial Distribution

Consider coin tosses, where the probability of heads is 1/2. Let's define the random variable h as the number of heads. Now let's consider a case where we flip five coins.

Distribution of this random variable h can be written

$$P(h = k) = p_h(k) = \binom{5}{k} p^k(1-p)^{5-k}$$

for $k = 0, 1, \ldots, 5$. This is a special case of what's known as the binomial distribution.

Binomial distribution: (note that n and p are parameters, while k is different values that can be assumed by the random variable):

$$P(k; n, p) = \binom{n}{k} p^k(1-p)^{n-k}$$

Notation: We might write this $h \sim Binom(n, p)$

This is the probability mass function, which assigns a probability to each of a finite number of realizations of the random variables.

The cumulative distribution function is the probability that we achieve any value less than or equal to a particular value. E.g. for the case defined above:

$$F(k) = \sum_{i=0}^{k} \binom{5}{i} p^i (1-p)^{5-i}$$

where $p = 1/2$.

To find the probability that the random variable falls between two values, say $2 \leq h \leq 4$ use $F(4) - F(2-1)$. Need to subtract off probability one step below the bottom bound because it's discrete.
Example: Find cumulative distribution of random variable h.

7.9 Uniform Distribution

This is a continuous distribution where all points on the interval over which it has support are equally likely.

Uniform distribution: $f(x) = \frac{1}{b-a}$

The above is the PDF of the normal distribution. More formally, this is written

$$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Example: say we have uniform distribution with support of $[-1, 3]$. What is $Pr([1,2])$?
Answer: $\frac{2}{3-(-1)} - \frac{1}{3-(-1)} = \frac{1}{4}$

What is the CDF of the uniform distribution in general terms?

$$\int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}$$

Formally, the CFD of the normal distribution is

$$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$$

7.10 Normal Distribution

This is the standard bell curve that you are likely familiar with. It is a continuous distribution function with support across the entire real line. σ parameterizes spread of distribution (it's variance) and μ parameterizes position.

Normal distribution: $f(x; \sigma, \mu) = \frac{1}{2\pi\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}}$

Note: In order for something to be a valid probability distribution or density function, it needs to sum or integrate to 1 when taken over the entire space where it has support. This, essentially, satisfies the axiom that $Pr(S) = 1$.

Mixed distributions/atoms in the distribution? In formal theory you often see this being ruled out because it's mathematically inconvenient.

Note: Methods of estimation (e.g. maximum likelihood estimation, or MLE) are about being given data, assuming a process (distribution) that underlies that data, and then trying to determine the parameters of the distribution from that data.

What does it mean to say that variables are independently and identically distributed?

IID random variables each have the same probability distribution as the others and all are mutually independent, i.e. all drawn independently from the same distribution. This is important for statistics because observations in a sample are often assumed to be effectively IID for the purposes of statistical inference.

What are some ways that we can connect a linear model to a distribution? Say we have a linear model and some cumulative distribution $\phi(x)$ (this could be the standard normal distribution, for instance). We can wrap this distribution around the model by writing it $\phi(X\beta)$

Models like logit and probit allow us to do this in a way that ensures estimated $\hat{y}$s are between zero and one.

7.11 Expected Values

Intuitively the long-run average value of repetitions of the experiment it represents.

In discrete space: $E(x) = \mu = \Sigma_k k \cdot p_x(k)$

Example: What is the expected value of rolling a die?
Answer: sum the probabilities of each side, multiplied by the value of that side: $1(\frac{1}{6}) + (2\frac{1}{6})... = 3.5$

Example: $x \sim Binom(4, 0.5)$. What is the expected value of x (i.e. $E(x)$)?
Answer: $\Sigma_4 4(\frac{1}{2} = 2$

Analogous continuous case: $E(Y) = \mu = \int_{-\infty}^{\infty} y \cdot f(y)dy$. This is a measure of central tendency.

Example: find expected value of $x \sim U(-1, 3)$.
Answer: Using the formula above,

$$\int_{-\infty}^{\infty} x \cdot \frac{1}{3-1} dx$$

We re-express this, simplifying and adjusting the bounds to evaluate this distribution on its support $(1, 3)$.

$$\int_1^3 \frac{1}{2} x dx = \frac{1}{4} x^2 \Big|_1^3 = \frac{9}{4} - \frac{1}{4} = 2$$

Example: find expected value of $x \sim f(t) = 2t,\ t \in [0, 1]$

Answer: $\mathbb{E}(x) = \int_0^1 x \cdot 2x dx = \int_0^1 2x^2 dx = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3}$

Rules of expectation operator:

1. $E(a) = a$
2. $E(bX) = bE(X)$

3. $E(a + bX) = a + bE(X)$
4. $\Sigma E(g(X)) = E(\Sigma g(x))$
5. $E(E(X)) = E(X)$ This is the law of iterated expectations

Conditional expectation: $E(Y|X)$

Example: Dice when six has already been rolled. What is conditional expectation of the value?
Answer: $\mathbb{E}(6 + x) = 6 + \mathbb{E}(x)$ =$6 + 3.5 = 9.5$

Regression function: $E(y|x)$. Notice, this is a conditional expectation where your outcome y depends on soem matrix x.

7.12 Variance and Other Moments

A central moment is a moment of a probability distribution of a random variable about the random variable's mean; that is, it is the expected value of a specified integer power of the deviation of the random variable from the mean.

m^{th} moment of X is $E(X^m)$. m^{th} central moment is $E(X - E(X))^m$

- First moment - mean/expected value
- First central moment - 0
- Second central moment - variance
- Third moment - skewness
- Fourth moment - kurtosis (sharpness of the peak)

Variance: deviation from expected value. If you sum the deviations of individual points, will sum to 0. Normally calculated as quatratic loss i.e. the expectation of the squared deviation of a random variable from its mean. Average of squared deviations:

$$Var(X) = \frac{\Sigma_I (x_i - \mu)^2}{n}$$

Covariance: a measure of the joint variability of two random variables, i.e. do these variables move together? $E(x - E(x))(y - E(Y))$

Also can be expressed as $E(X^2) - (E(X))^2$. See proof below.

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E(X^2 - 2xE(X) + (E(X))^2) \\ &= E(X^2) - E(2xE(x)) + (E(X))^2 \\ &= E(X^2) - 2E(x)E(x) + (E(X))^2 \\ &= E(X^2) - (E(x))^2 \end{aligned}$$

7.13 Rules of Variance and Covariance

- $Var(a + bX) = b^2 Var(X)$
- $Var(a + bX + cY) = b^2 Var(X) + c^2 VAR(Y) + 2bcCov(X, Y)$
- $Var(c) = 0$
- $Cov(X, Y) = E(XY) - E(X)E(Y)$. Note, this is zero if X and Y are independent, as in this case $E(XY) = E(X)E(Y)$
- $Cov(X + c, Y + b) = Cov(X, Y)$
- $Cov(cX, bY) = cbCov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov(X, X) = Var(X)$

Correlation: Most commonly used to describe the extent to which two variables have a linear relationship with each other.

$$\frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma X \sigma Y}$$

7.14 Quick Introduction to Maximum Likelihood Estimation

- Up until now, we've been dealing with distributions where we know the value of all the parameters.
- What if we know some things about the distribution (or at minimum, suspect certain things) such as what family of distributions it is part of, but don't know others, such as the value of particular parameters?
- We may, in this case, want to estimate the value of particular parameters.
- Example: Imagine being handed a coin, and not knowing whether or not it's a fair coin. You know that coin flips will be binomially distributed (i.e. you know the family of the distribution) and you know how many times you've flipped it (n) but you don't know p. How might you "guess" p by flipping it a number of times and using the data you generate?
- Any function of the data whose objective is to "guess" the parameter is called an *estimator*, even if this estimator ends up throwing out a lot of the data.
- Computing the value of that function given particular data gives you an *estimate*.
- One approach is known as maximum likelihood estimation (MLE).
- Treat each realization of the coin flip as iid. This allows us to obtain a "likelihood function", which is in effect, the joint probability of your "sample".
- General form: $L(\theta) = \prod_{i=1}^{n} p(k_i; \theta)$, where θ is an arbitrary parameter or set of parameters.
- Intuition: we're just multiplying all the probabilities of the individual observations. For discrete distributions this gives you the joint probability of the sample; for continuous distributions, it gives you the joint density.
- In either case, a sensible approach to estimation is to choose the parameter that maximizes this function, as this is the choice that makes the sample most likely.
- Will often maximize a transformation of the function.

- In particular, recall that logarithms are monotonic functions, so taking the natural logarithm of the likelihood function will not change the optima.
- We can try an example: flip a coin 15 times. How can we use MLE to derive an estimator for p?

Binomial Distribution with $n = 15$
Say we're given $n = 15$, so the only unknown parameter is p. Thus we have:

$$= \binom{15}{x} p^x (1-p)^{15-x}$$

Resulting in the likelihood function (where m is the number of observations)

$$L(p|\boldsymbol{x}, n = 15) = \prod_{i=1}^{m} \binom{15}{x_i} p^{x_i} (1-p)^{15-x_i}$$

And the log-likelihood function (all Σs are $\Sigma_{i=1}^{m}$):

$$logL(p|\boldsymbol{x}, n = 15) = \Sigma x_i log(p) + \Sigma(15 - x_i) log(1-p) + \Sigma log\left(\binom{15}{x_i}\right)$$

Taking the derivative and setting to zero

$$\begin{aligned}
\frac{\partial logL}{\partial p} &= \frac{\Sigma x_i}{p} + \frac{\Sigma 15 - x_i}{1-p}(-1) = 0 \\
\leftrightarrow \frac{1-p}{p} &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\
\leftrightarrow \frac{1}{p} - 1 &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\
\leftrightarrow \frac{1}{p} &= \frac{\Sigma(15 - x_i) + \Sigma x_i}{\Sigma x_i} \\
\leftrightarrow p &= \frac{\Sigma x_i}{\Sigma(15) - \Sigma x_i + \Sigma x_i} \\
\leftrightarrow p &= \frac{\Sigma x_i}{m(15)}
\end{aligned}$$

- An estimator doesn't need to be good to be an estimator. "Seven" is an estimator. A fair bit of statistics work is on properties of estimators as ways to evaluate them.
- A big divide in political methodology is about how willing people are to make distributional assumptions. Randomized experiments don't require making distributional assumptions for results to hold, but much of the statistical modeling work with observational data does.
- Nonparametric methods try to avoid distributional assumptions, but have less power than parametric (distribution assuming) methods.
- Approaches coming from the causal inference literature (e.g. regression discontinuity designs, natural experiments, etc.) are often about trying to get "as if" random assignment to avoid making distributional assumptions.