Random Variables and Distributions

Snacks and Stats

October 2019

1 Terminology

This session will use several common statistical terms, most of which have very different colloquial definitions! Let's start with some definitions:

- **Outcome space (or sample space)** – All possible results of an experiment. If the experiment is flipping a coin, the outcome space is $\Omega = \{H, T\}$. If the experiment is flipping a coin twice, the outcome space is $\Omega = \{HH, HT, TH, TT\}$, etc.
- **Random variable** – Often we're more interested in some real-valued function of the outcome space rather than the actual outcome. For example, we may flip a coin 10 times and are interested in the *number of heads* we get. The number of heads is then a real-valued function of the outcome space. We'd call the number of heads a 'random variable'. This is a different use of the term 'random', and does not imply that all values are equally likely or that we can't evaluate their probability.
- **Discrete** – A random variable is discrete if its possible values are a countable set (eg. integers). The number of heads we get when flipping a coin 10 times is a discrete random variable. We can assign a probability to an individual value of a random variable.
- **Continuous** – A random variable is continuous if its possible values are not countable (eg. real numbers). The temperature of a gas is a continuous random variable. The probability of any individual value of a continuous random variable is zero, but we can assign probabilities to a range of values instead.

2 Distribution and Density Functions

A distribution function describes how probability is distributed over the possible values a random variable can take. There are many common ways that probability can be distributed, but there are also different ways to describe that distribution. We'll start with the latter, then talk about the former.

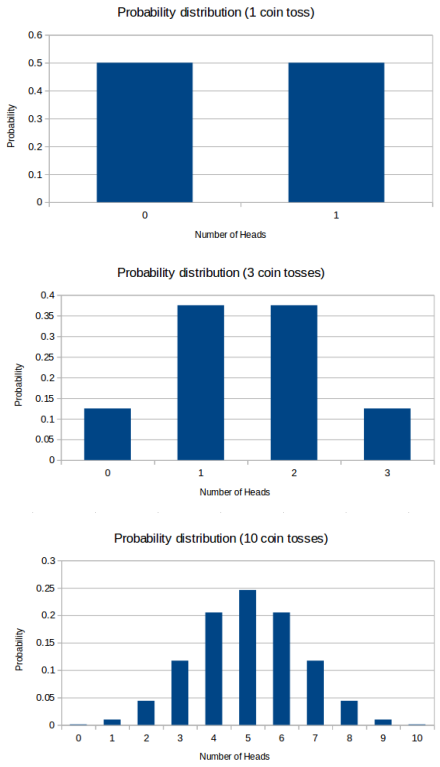

Figure 1: The probability distribution of number of heads for 1, 3 and 10 coin tosses.

2.1 Distributions and Densities

Perhaps the most straightforward way to describe a probability distribution would be to plot a graph of $P(X)$ vs. X: i.e. the probability of a random variable X vs. possible values of that random variable. This will only work for discrete random variables (can you see why?) but it's a good place to start. Staying with the example of tossing a coin, Figure 1 shows the distribution of probability over the number of heads when tossing 1, 3 or 10 coins. In each case a discrete number of heads has a finite, non-zero probability.

Another way of displaying essentially the same information is a cumulative probability distribution, as seen in Figure 2. In this case the probabilities displayed on the y-axis are not $P(x)$, but rather $P(X \leq x)$: i.e. the probability that the value of the random variable is *less than or equal to* x.

Continuous random variables are most often displayed as a probability density function, or 'p.d.f.'. The term 'density' in this context reflects the fact that individual values of a continuous random variable have a probability of zero, but *ranges* have non-zero probability. Therefore, in order to evaluate the prob-

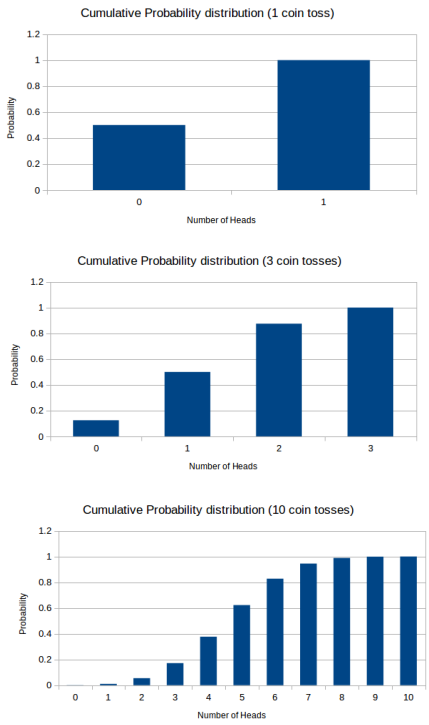

Figure 2: The cumulative probability distribution of number of heads for 1, 3 and 10 coin tosses.

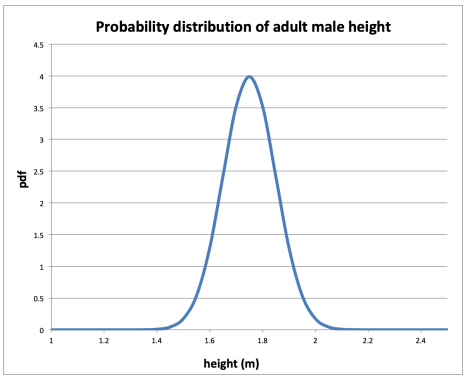

Figure 3: The probability distribution function of adult male height.

ability of a random variable taking a value within a range, one must integrate the probability density function over that range. A counter-intuitive property of probability density functions is that they *can have values greater than 1.*

Let's consider an example of a continuous random variable: the height of adult males. The probability distribution function of adult male height is shown in Figure 3.

To determine the probability that a given adult male has a height X, we can't just read the value straight off Figure 3 (can you see why?). Instead, we can only determine the probability that a given adult male has a height within the range $X+\Delta X$, by finding the definite integral of the probability distribution function over that range:

$$P(X + \Delta X) = \int_{X}^{X+\Delta X} \text{p.d.f } dX$$

That then implies that the total integral of the probability distribution function has to be 1:

$$\int_{-\infty}^{\infty} \text{p.d.f } dX = 1$$

This point is important. It implies that if you are to use any kind of probability distribution function in your research (i.e. a likelihood function) it must be *normalised* so that it integrates to 1 over parameter space.

Another useful way of displaying the probability distribution of a continuous random variable is a 'quantile function'. This is essentially a cumulative probability distribution with the axes flipped. Quantile functions are useful for answering questions like:

Q: 'What is the median adult male height?'

A: The height corresponding to the 0.5 quantile (1.75 m).

Q: 'How tall should a doorway be so that only 5% of the adult male population would have to duck their heads to enter?'

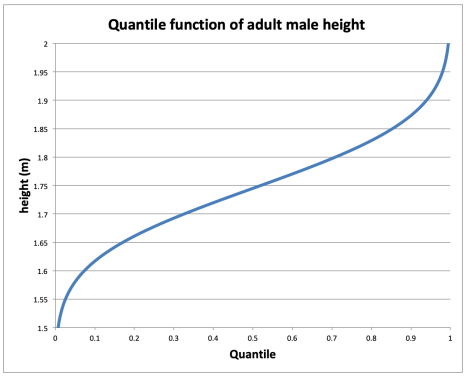

Figure 4: The quantile function of adult male height.

A: The height corresponding to the 0.95 quantile (1.9 m).

Quantile functions are the basis behind 'p-values', 'sigma levels' and other measurements of 'confidence'. For example, if a researcher says that they have developed a test that is successful and has a p-value of 0.04 (forgive my language here...), they are saying that the results they obtained would only be likely to happen 4% of the time if their test didn't actually work. In essence you are answering the question 'What is the probability of obtaining this result by chance?'. If the answer is very small, then 'chance' is likely not the cause. For another example, when a researcher says they have detected a new galaxy with confidence of $3\,\sigma$, they mean that their result is 3 standard deviations (σ) away from the mean expected if the galaxy wasn't there. That assumes that the probability distribution is gaussian... which we'll get to next.

2.2 Types of Distributions

Here we'll introduce some commonly seen distributions. If there's interest we can look into these more in our next session.

2.2.1 Discrete distributions

Bernoulli distribution
Example: Result of a True/False test

$$P(X = \text{True}) = p = 1 - P(X = \text{False})$$

Binomial distribution
Example: Number of times a True/False test returns True

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \text{ for } i = 0, 1, ..., n$$

where $\binom{n}{i} = \frac{n!}{i!(n-i)!}$

Poisson distribution
Example: Number of times a True/False test returns True if the probability of True is very small but the number of trials is large.

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \text{ for } i = 0, 1, ...$$

where rate $\lambda > 0$.

2.2.2 Continuous distributions

Uniform distribution
Example: All outcomes are equally likely

$$\text{p.d.f} = \frac{1}{b-a}, \text{ for } a < x < b$$

where a and b define the range of possible values of the random variable x.
Exponential distribution
Example: Half-life of a radioactive source

$$\text{p.d.f} = 1 - e^{-\lambda x}$$

Gaussian (Normal) distribution
Example: Velocity of molecules in a gas

$$\text{p.d.f} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with mean μ and standard deviation σ.

3 Central Limit Theorem

As the sample size or number of trials increase, many distributions will tend towards a Gaussian distribution. This is called the 'central limit theorem'. It's important to not take this for granted, as it's not always clear how large a sample must be before the distribution can be safely assumed to be Gaussian, or if there is some underlying phenomenon that prevents it from ever becoming Gaussian.

4 R!

```
# Set up 6 panel figure

par(mfrow=c(3,2))
```

```
# Plot upper left panel with three illustrative exponential p.d.f.
    distributions

xdens <- seq(0,5,0.02)
plot(xdens,dexp(xdens,rate=0.5), type='l', ylim=c(0,1.5), xlab='',
   ylab='Exponential p.d.f.',lty=1)

lines(xdens,dexp(xdens,rate=1), type='l', lty=2)
lines(xdens,dexp(xdens,rate=1.5), type='l', lty=3)
legend(2, 1.45, lty=1, substitute(lambda==0.5), box.lty=0)
legend(2, 1.30, lty=2, substitute(lambda==1.0), box.lty=0)
legend(2, 1.15, lty=3, substitute(lambda==1.5), box.lty=0)

# Help files to learn these function

help(seq) ; help(plot); help(par) ; help(lines); help(legend)

# Plot upper right panel with three illustrative exponential c.d.f.
    distributions

plot(xdens, pexp(xdens,rate=0.5), type='l', ylim=c(0,1.0), xlab='',
   ylab='Exponential c.d.f.', lty=1)
lines(xdens, pexp(xdens,rate=1), type='l', lty=2)
lines(xdens, pexp(xdens,rate=1.5),type='l',lty=3)
legend(3, 0.50, lty=1, substitute(lambda==0.5), box.lty=0)
legend(3, 0.38, lty=2, substitute(lambda==1.0), box.lty=0)
legend(3, 0.26, lty=3, substitute(lambda==1.5), box.lty=0)

# Plot middle panels with illustrative normal p.d.f. and c.d.f.

xdens <- seq(-5, 5, 0.02)
ylabdnorm <- expression(phi[mu~sigma^2] (x))
plot(xdens, dnorm(xdens, sd=sqrt(0.2)), type='l', ylim=c(0,1.0), xlab='',
   ylab=ylabdnorm,lty=1)
lines(xdens,dnorm(xdens, sd=sqrt(1.0)), type='l', lty=2)
lines(xdens,dnorm(xdens, sd=sqrt(5.0)), type='l', lty=3)
lines(xdens,dnorm(xdens, mean=-2.0, sd=sqrt(0.5)), type='l', lty=4)
leg1 <- expression(mu^' '==0, mu^' '==0, mu^' '==0, mu^' '==-2)
leg2 <- expression(sigma^2==0.2, sigma^2==1.0, sigma^2==5.0,
    sigma^2==0.5,)
legend(0.5, 1.0, lty=1:4, leg1, lwd=2, box.lty=0)
legend(3.0, 1.01, leg2, box.lty=0)

ylabpnorm <- expression(Phi[mu~sigma^2] (x))
plot(xdens,pnorm(xdens,sd=sqrt(0.2)), type='l', ylim=c(0,1.0), xlab='',
   ylab=ylabpnorm,lty=1)
lines(xdens,pnorm(xdens, sd=sqrt(1.0)), type='l', lty=2)
lines(xdens,pnorm(xdens, sd=sqrt(5.0)), type='l', lty=3)
lines(xdens,pnorm(xdens, mean=-2.0, sd=sqrt(0.5)), type='l', lty=4)
leg1 <- expression(mu^' '==0, mu^' '==0, mu^' '==0, mu^' '==-2)
```

```
leg2 <- expression(sigma^2==0.2, sigma^2==1.0, sigma^2==5.0,
    sigma^2==0.5,)
legend(0.5, 0.6, lty=1:4, leg1, lwd=2, box.lty=0)
legend(3.0, 0.61, leg2, box.lty=0)

# Plot bottom panels with illustrative lognormal p.d.f. and c.d.f.

xdens <- seq(0,3, 0.02)
plot(xdens, dlnorm(xdens, meanlog=0, sdlog=5), type='l', ylim=c(0,2),
    xlab='',
   ylab='Lognormal density', lty=1)
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1), type='l', lty=2)
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1/2), type='l', lty=3)
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1/8), type='l', lty=4)
leg1 <- expression(sigma==5, sigma==1, sigma==1/2, sigma==1/8)
legend(1.8,1.8,lty=1:4,leg1,box.lty=0)

plot(xdens,plnorm(xdens,meanlog=0,sdlog=5),type='l',ylim=c(0,1),xlab='x',
    ylab='Lognormal distribution',lty=1)
lines(xdens,plnorm(xdens,meanlog=0,sdlog=1),type='l',lty=2)
lines(xdens,plnorm(xdens,meanlog=0,sdlog=1/2),type='l',lty=3)
lines(xdens,plnorm(xdens,meanlog=0,sdlog=1/8),type='l',lty=4)
leg1 <- expression(sigma==5,sigma==1,sigma==1/2,sigma==1/8)
legend(1.5,0.6,lty=1:4,leg1,box.lty=0)

# Return plot to single-panel format

par(mfrow=c(1,1))
```