

# Quantitative Methods in Geography

Geo 340

William D. McCoy

Department of Geosciences  
University of Massachusetts, Amherst

Autumn, 2007

## Random variable

A random variable is a rule that assigns a number to each elementary outcome in  $\mathcal{S}$ . In other words, a random variable is any numerically valued function defined over a sample space.

The probability distribution of a random variable may be represented by a table, a graph, or an equation that assigns a number to a variable.

We use an uppercase letter to denote a random variable and a lowercase letter to denote a particular value of the variable.

## Probability distributions

- ▶ random variables
- ▶ expectation and variance
- ▶ R functions for probability distributions
- ▶ discrete distributions
  - ▶ uniform discrete distribution
  - ▶ binomial distribution
  - ▶ poisson distribution
- ▶ continuous distributions
  - ▶ uniform continuous distribution
  - ▶ normal (Gaussian) distribution
- ▶ Central limit theorem

## Discrete and continuous random variables

$P(X)$  represents the probability function or the probability distribution for the random variable  $X$ .

$P(X = x)$  represents the probability that the random variable  $X$  takes on the specific value  $x$ .

A *discrete* random variable is one that can take on only a finite number of values. An example might be the number of people enrolled in a class at UMass.

A *continuous* random variable is one that can take on any number between some upper and lower limits. Even if the upper and lower limits are finite (and they don't need to be), there is an infinite number of real numbers between any two real numbers. An example might be temperature.

## Discrete probability distributions

The probability distribution of a discrete random variable is specified by a probability mass function (pmf) which is usually represented by a table, graph, or equation. There are two basic restrictions for any discrete probability distribution:

$$0 \leq P(X = x_i) \leq 1, \quad i = 1, 2, 3, \dots, k$$

$$\sum_{i=1}^k P(X = x_i) = 1$$

where there are  $k$  different possible discrete values for  $X$ .

## Discrete probability distributions (cont'd)

The cumulative distribution function (cdf) for a discrete random variable having pmf  $p(y)$  is:

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

for each number  $x$ .

## Expectation

The expected value of a discrete random variable is the weighted mean of the possible values,  $x_i$ :

$$E(X) = \mu_X = \sum_{i=1}^k x_i P(x_i)$$

The variance,  $V(X)$ , is:

$$V(X) = \sigma_X^2 = \sum_{i=1}^k [x_i - \mu]^2 P(x_i) = \sum_{i=1}^k x_i^2 P(x_i) - [E(X)]^2$$

The standard deviation,  $SD(X)$  or  $\sigma_X$ , is the square root of  $V(X)$ .

## Continuous probability distributions

The probability distribution of a continuous random variable is specified by a probability density function (pdf) which is usually represented by a graph or an equation. Again there are two fundamental requirements:

$$f(x) \geq 0, \quad \text{for all } x$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

The probability of  $X$  within an arbitrary interval is given by:

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

That is to say the probability is equal to the area under the curve of  $f(x)$  between  $a$  and  $b$ .

## Continuous probability distributions (cont'd)

The cumulative distribution function (cdf) for a continuous random variable having pdf  $f(y)$  is:

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

## R functions for probability distributions

R has a number of functions that return various properties of a large number of theoretical distributions. There is a common system of prefixes that are used to name these functions. The names of the functions use a root name (such as “norm” for the normal distribution) with one of the following prefixes:

- d Returns the probability density at a given point for a continuous distribution or the probability of a given value for a discrete distribution
- p Returns the cumulative probability up to a given value
- q Returns the quantile of a distribution for a given probability
- r Returns random numbers drawn from the distribution

## Examples of specific R functions

`dnorm(-0.5)` returns the probability density of the standard normal distribution at the point  $-0.5$  (one-half of a standard deviation below the mean).

`pnorm(-0.5)` returns the probability of a value equal to or less than  $-0.5$  for the standard normal distribution, *i.e.* the area in the tail of the “bell” curve to the left of  $-0.5$ .

`qnorm(0.25)` returns the first quartile of the standard normal distribution, *i.e.* the value in the standard normal distribution to the left of which is one-quarter of the area under the curve.

`rnorm(5)` returns a vector of five numbers drawn at random from the standard normal distribution.

## Discrete uniform distribution

A simple, but often used distribution is the discrete uniform distribution. The distribution arises when there are a discrete number,  $k$ , of equally likely elementary outcomes (such as the roll of a single die). The probability mass function for the distribution can be described mathematically as:

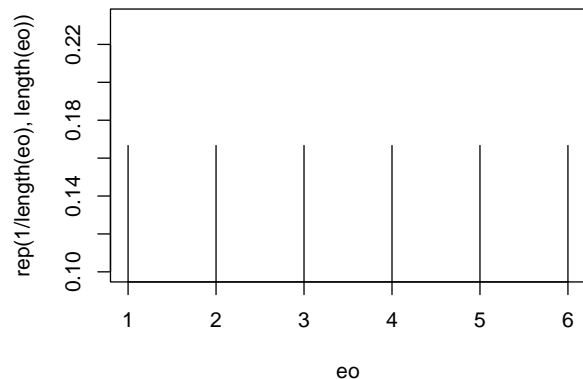
$$P(X = x) = \begin{cases} \frac{1}{k} & \text{for } x = 1, 2, 3, \dots, k, \\ 0 & \text{for all other } x. \end{cases}$$

$$E(X) = \mu = \frac{k+1}{2}$$

$$V(X) = \sigma^2 = \frac{k^2-1}{12}$$

## Plot of a uniform discrete distribution

```
> eo <- 1:6
> plot(eo, rep(1/length(eo), length(eo)),
+      type = "h")
```



## Mean and variance of a binomial distribution

The expected value of a binomial distribution is:

$$E(X) = \mu = np$$

and the variance of a binomial distribution is:

$$V(X) = \sigma^2 = np(1 - p)$$

So the standard deviation of a binomial distribution is:

$$SD(X) = \sigma = \sqrt{np(1 - p)}$$

## Binomial distribution

A Bernoulli trial is a *single* try of a probabilistic experiment having exactly two outcomes. The binomial distribution is used to determine the probability of  $x$  number of successes in a set of  $n$  Bernoulli trials where the probability of success of each try is  $p$ . The probability of failure for each try is, of course,  $1 - p$ . The probability mass function is:

$$P(X = x) = \begin{cases} C(n, x)p^x(1 - p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n, \\ 0 & \text{for all other } x. \end{cases}$$

(See the handout to go through a demonstration of how this equation can be derived.)

## Functions for the binomial distribution in R

Using **R**, we can easily solve this binomial probability mass function using `dbinom()`. For example, to find the probability of three successes in five trials where the probability of success is 0.3:

```
> dbinom(3, 5, 0.3)
```

```
[1] 0.1323
```

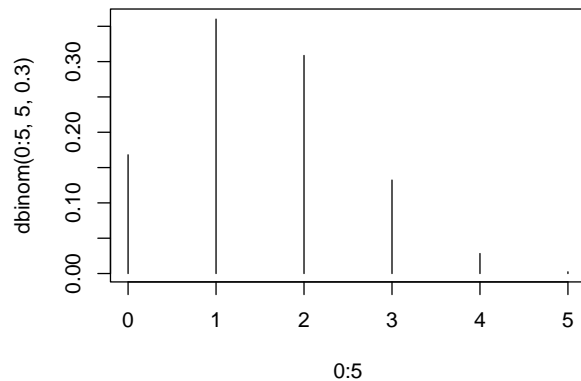
And the probability of three or fewer successes can be found with the cumulative probability function `pbinom()`:

```
> pbinom(3, 5, 0.3)
```

```
[1] 0.96922
```

## Plot of a binomial distribution

```
> plot(0:5, dbinom(0:5, 5, 0.3), type = "h")
```



## Functions for the Poisson distribution in R

We can easily solve the Poisson probability mass function in **R** using `dpois()`. For example, to find the probability of three occurrences in an interval when the mean number of occurrences per interval,  $\lambda$ , is 2:

```
> dpois(3, 2)
```

```
[1] 0.1804470
```

And the probability of three or fewer occurrences can be found with the cumulative probability function `ppois()`:

```
> ppois(3, 2)
```

```
[1] 0.8571235
```

## Poisson distribution

In situations where  $n$  is very large (say  $\geq 100$ ) and  $p$  is very small (say  $\leq .01$ ) and the product  $np$  approaches a value,  $\lambda$ , which is the mean of the distribution, we can use the Poisson distribution as a good approximation:

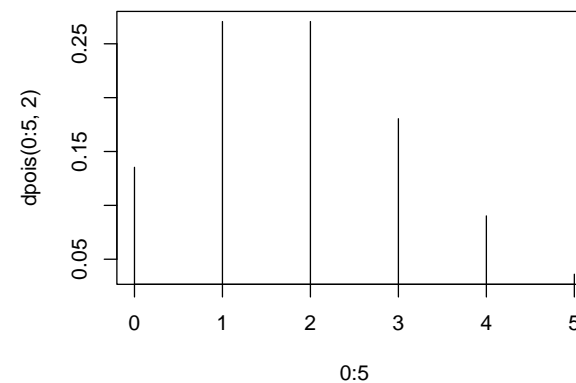
$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \text{ for } x = 0, 1, 2, \dots$$

$$E(X) = \lambda$$

$$V(X) = \lambda$$

## Plot of a Poisson distribution

```
> plot(0:5, dpois(0:5, 2), type = "h")
```



## Uniform continuous distribution

The uniform continuous distribution has the pdf:

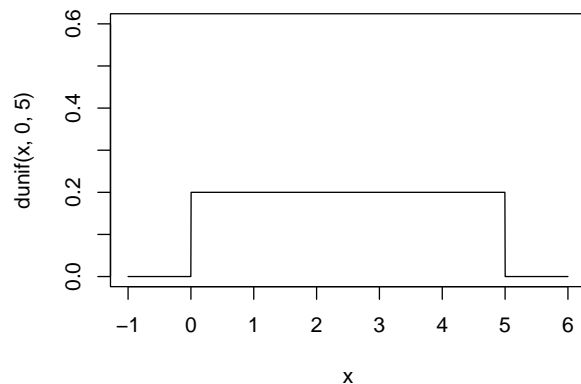
$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a, x > b \end{cases}$$

$$E(X) = \mu = \frac{a+b}{2}$$

$$V(X) = \sigma^2 = \frac{(b-a)^2}{12}$$

## Plot of a uniform continuous distribution

```
> curve(dunif(x, 0, 5), -1, 6, n = 1501,  
+       ylim = c(0, 0.6))
```



## cdf of a uniform continuous distribution

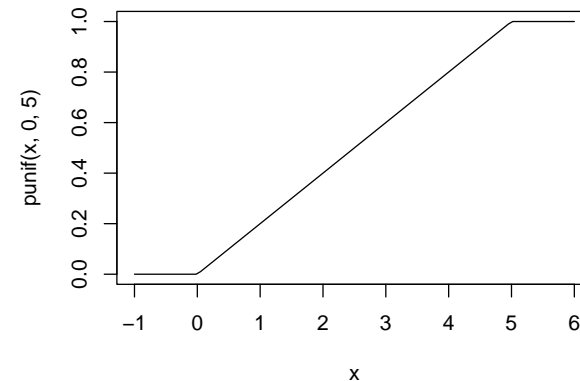
The cdf of a uniform continuous distribution is:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

And a plot of the cdf looks like this (next frame):

## Plot of cdf of a uniform distribution

```
> curve(punif(x, 0, 5), -1, 6)
```



## Normal (Gaussian) distribution

The normal distribution is a continuous distribution defined by the following pdf for all real  $x$

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

$$E(X) = \mu$$

$$V(X) = \sigma^2, \quad SD(X) = \sigma$$

We can also easily find values for the cumulative probability up to any value of  $x$  using `pnorm()`. For example, to find the cumulative probability at  $X = 0.4$  in the standard normal distribution:

```
> pnorm(0.4)
[1] 0.6554217
```

And to find a cumulative probability for a normal distribution with a different mean and standard deviation, just specify the mean and s.d. in the call to `pnorm()`:

```
> pnorm(6, 5, 2)
[1] 0.6914625
```

## Functions for the normal distribution in R

In R, we can easily find values for the probability density for any value of  $x$  using `dnorm()`. For example, to find the probability density at  $X = -0.3$  in the standard normal distribution, *i.e.*  $\mu = 0$  and  $\sigma = 1$ :

```
> dnorm(-0.3)
[1] 0.3813878
```

To find a probability density for a normal distribution with a different mean and standard deviation, just specify the mean and s.d. in the call to `dnorm()`:

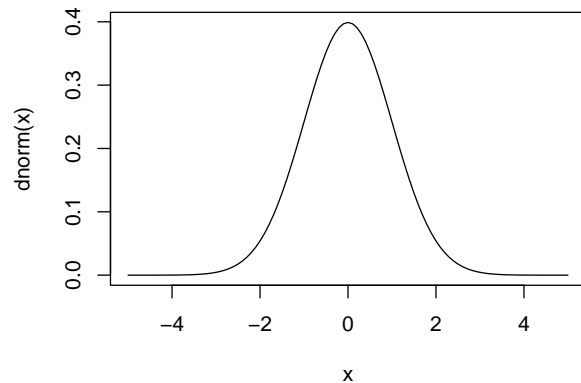
```
> dnorm(1.5, 5, 2)
[1] 0.04313866
```

To find the value of  $x$  that has a cumulative probability  $p$  (*i.e.* a quantile) in the standard normal distribution use the `qnorm()` function:

```
> qnorm(0.95)
[1] 1.644854
```

## Plot of the standard normal distribution

```
> curve(dnorm(x), -5, 5)
```



## Central limit theorem

If the sample size,  $n$ , is sufficiently large, the distribution of sample means for samples drawn from *any* population (with mean  $\mu$  and variance  $\sigma^2$ ) will be approximately normally distributed with  $\mu_{\bar{x}} = \mu$  and  $\sigma_{\bar{x}}^2 = \sigma^2/n$ . And as  $n$  gets larger, the sampling distribution gets ever closer to a normal distribution.  $\sigma_{\bar{x}}$  is called the standard error of the mean. The CLT gives us these rules:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

## Plot of cdf of a normal distribution

```
> curve(pnorm(x), -4, 4)
```

