Random Variables

CSU, Chico Math 314

2018-09-10

outline

outline

Population Level

What we are about to have a theoretical discussion about how we think about populations in statistics.

- Populations are described with functions
- These functions produce random data
- Using these data (from functions that describe the population) we estimate characteristics about the population

outline

Random Variables

Statistics is formed around the idea of a **random variable**, which means something a bit different than we normally think of when we use the word variable.

random variable

A random variable is a function that maps outcomes of the sample space to a number.

Random Variables Types

As before with different types of variables, we again have different types of random variables.

discrete
A discrete random variable has a sample space that is at most countably infinite.

continuous
A continuous random variable has a sample space that can take on any value within any given interval.

Discrete Random Variable

We theorize that data are generated from a function that only takes on mass at each value in the sample space. Probability statements are made from area under the function.

Discrete RV, motivation

The probability model of coin flipping.

x	1 (head)	0 (tail)
$P(X = x)$	p	$1 - p$

Discrete Random Variable, example

A **discrete random variable** is described using a function, but is sometimes more easily seen using a table. Such tables denote the values the random variable can take on and the probability of each value (when known). Consider the random variable X (capitlized) that takes on the following values x (lower case).

x	1	2	3	4
$P(X = x)$	0.4	0.3	0.2	0.1

Probability Mass Function

We call the function that describes an arbitrary discrete random variable X, the **probability mass function**.

probability mass function

The PMF of a discrete random variable X lists the probabilities associated with each value x the random variables X takes on.

Discrete Random Variable, example

Consider again, the random variable X from above (not the coin).
What is $P(X \leq 3)$?

x	1	2	3	4
$P(X = x)$	0.4	0.3	0.2	0.1

Cumulative Distribution Function

Probabilities of events of the form $\{X \leq x\}$ are defined generally with the term **cumulative distribution function**, for both discrete and continuous random variables.

cumulative distribution function

The CDF of a random variable X is defined to be
$F_X(x) = P(X \leq x)$, for all numbers x.

Cumulative Distribution Function, example

What then is the CDF of the random variable X above?

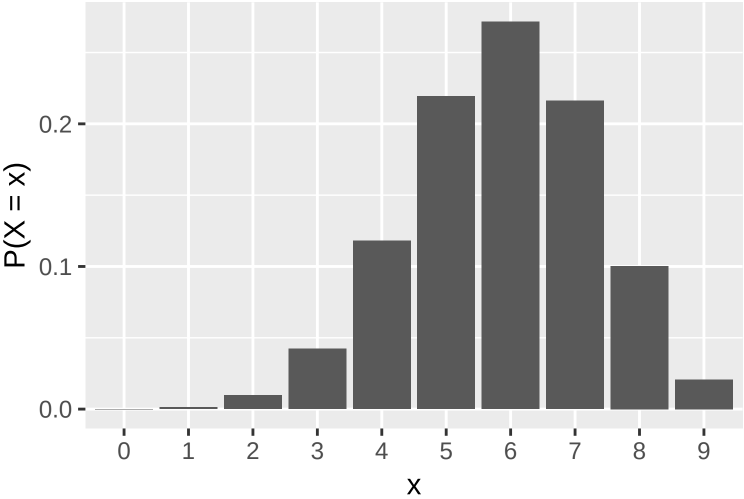
x	1	2	3	4
$F_X(x)$	0.4	0.7	0.9	1.0

Notes on PMF

Some notes on PMFs

- $P(X = x) \geq 0$, for all x
- $\sum_{x \in \mathcal{S}_X} P(X = x) = 1$
  - re CDF, $F_X(\infty) = 1$.
- $P(a < X \leq b) = \sum_{a < x \leq b} P(X = x) = F(b) - F(a)$
- watch the inequalities, $P(X \leq 3) \neq P(X < 3) = P(X \leq 2)$

PMF, via picture



Probability Density Function

The continuous analogue to the PMF is the **probability density function**.

probability density function

The PDF of a continuous random variable X is a nonnegative function f_X under which we measure probability on intervals of interest, say $[a, b]$.

Probability Density Function

The PDF of the random variable Y allows us to calculate, $P(a < Y < b)$. Specifically,

$$P(a < Y < b) = \int_a^b f_Y(y)dy.$$

Cumulative Distribution Function

The CDF of the continuous random variable Y is defined similarly,

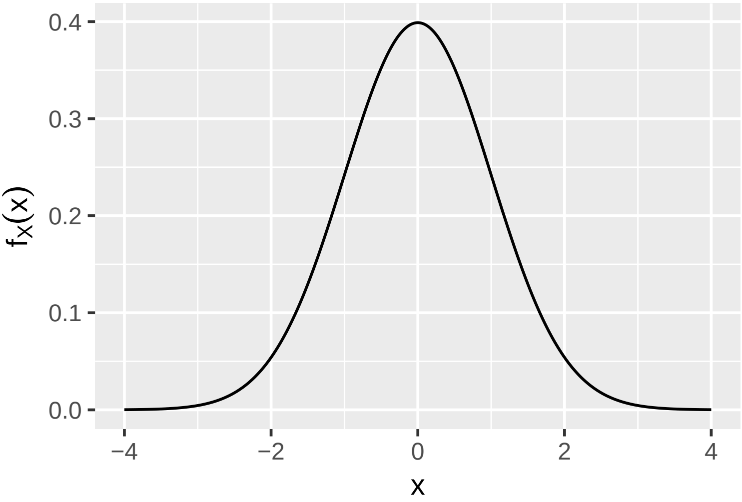
$$F_Y(y) = \int_{-\infty}^{y} f_Y(s) ds.$$

Notes on PDF

Some notes on PDFs

- $f_Y(y) \geq 0$, for all y
- $\int_{-\infty}^{\infty} f_Y(y)dy = 1$
  - re CDF, $F_Y(\infty) = 1$.
- $P(Y = y) = 0$, for all y
- $P(a < Y \leq b) = \int_a^b f_Y(y)dy = F(b) - F(a)$
- inequalities don't matter, $P(Y \leq 3) = P(Y < 3)$

PDF, via picture



outline

Motivating Parameters

In fact, all the sample statistics you've ever heard of are first defined for populations and then translated into data-terms. The population (not sample) mean is called the **expected value**. We should think about the center of the function that describes the random variable of interest.

Motivating Parameters, coin filp example

For instance, a coin flip's function is

$$f(x|p) = p^x(1-p)^{1-x},\ \text{for } x \in \{0, 1\}.$$

Motivating Parameters, coin flip example

The measure of center for this function is probably more easily seen with a table or a graph.

x	1 (head)	0 (tail)
$P(X = x)$	p	$1 - p$

Motivating Parameters, coin flip example

If a coin lands on heads (1) half the time and lands on tails (0) half the time, where is the center?

$$0.5 * 1 + 0.5 * 0 = 0.5$$

Expected Value, discrete random variable

Let's generalize the idea of the expected value of a coin flip to all discrete random variables. The population mean of a discrete random variable is called the **expected value**.

expected value

The expected value of a (discrete) random variable X is defined to be

$$E(X) = \mu_X = \sum_{x \in \mathcal{S}_X} x \cdot P(X = x).$$

Expected Value, discrete random variable, example

What is the expected value of the random variable we assume fits die rolling?

$$
\begin{aligned}
E(X) &= \sum_{x \in \mathcal{S}_X} x \cdot P(X = x) \\
&= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + \ldots + 6 \cdot P(X = 6) \\
&= 1/6 + 2/6 + \ldots + 6/6
\end{aligned}
$$

Expected Value, notes

Note about the expected value

- the most common **parameter** in statistics,
- the population analogue to sample mean; a measure of center
- this is essentially where the sample mean comes from. We have n randomly sampled (all occur with equal probability) outcomes, so that $P(X = x_i) = 1/n$.

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

Expected Value, continuous random variable

The expected value for continuous random variables is defined similarly, but with integrals in place of sums.

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x) df$$

Variance

The population variance is defined via the expected value and the expected value property about any function h.

variance

The variance is defined to be

$$Var(X) = \sigma_X^2 = E[(X - \mu_X)^2].$$

Variance and Standard Deviation

The relationship between the population variance and the population standard deviation is, again, simple.

standard deviation

The standard deviation is defined to be

$$\sigma_X = \sqrt{\sigma_X^2}.$$

outline

Bernoulli Distribution

The expected value of the random variable $X \sim Bernoulli(p)$ is simple to calculate. We'll calculate its variance on Friday.

x	0	1
$P(X = x)$	$1 - p$	p

Uniform Discrete

The discrete uniform distribution $X \sim \mathcal{U}(a, b)$ has PMF

$$f(x|a, b) = \frac{1}{b - a + 1} = \frac{1}{n},$$

where $n = b - a + 1$.

Uniform Continuous

The continuous uniform distribution $X \sim \mathcal{U}(a, b)$ has probability density function (PDF)

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

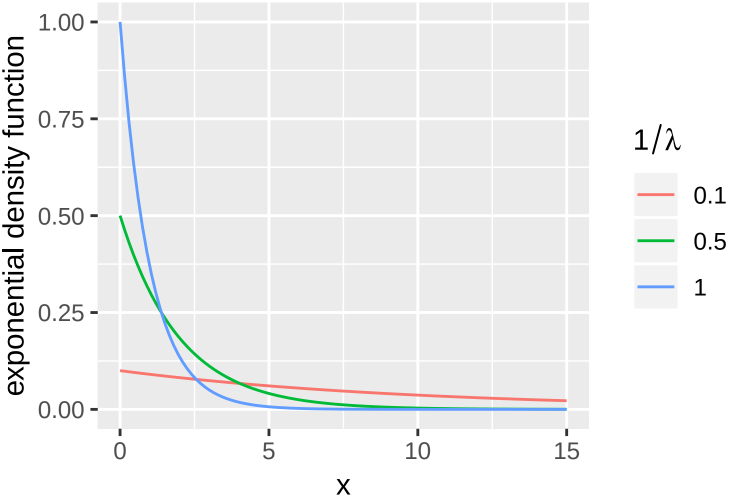
Exponential

The exponential distribution $X \sim Exp(\lambda)$ has PDF

$$f(x|\lambda) = \begin{cases} \lambda \exp\left(-\lambda x\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Exponential, notes

The exponential distribution is often used to compute the survival time of things.



PDF, Normal

The normal (Gauβian) distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ has PDF

$$f(x|\mu,\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}.$$

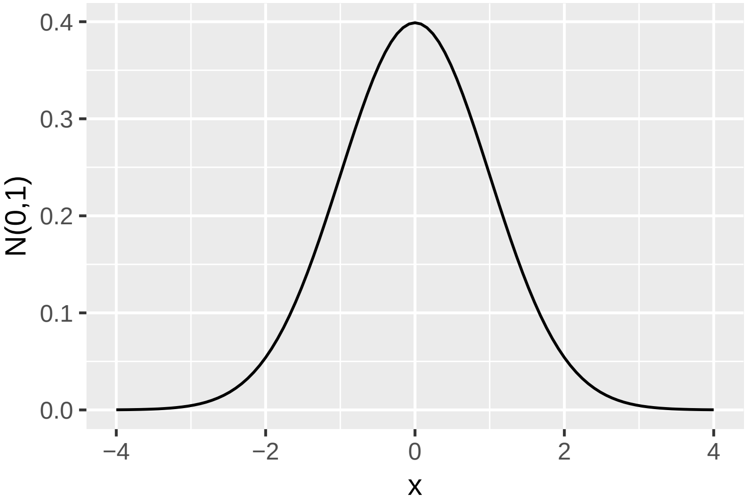
Normal, notes

Notes on the normal distribution

- area under the curve is equal to 1,
- unimodal (one hump)
- perfectly symmetric about μ,
- parameters: centered at μ with standard deviation σ,
- often $\mu = 0$, shifts the distribution,
- often $\sigma = 1$, scales the distribution,
- we write in short hand $N(\mu, \sigma^2)$,
- use Z to denote $N(0, 1)$
- Gauβian distribution

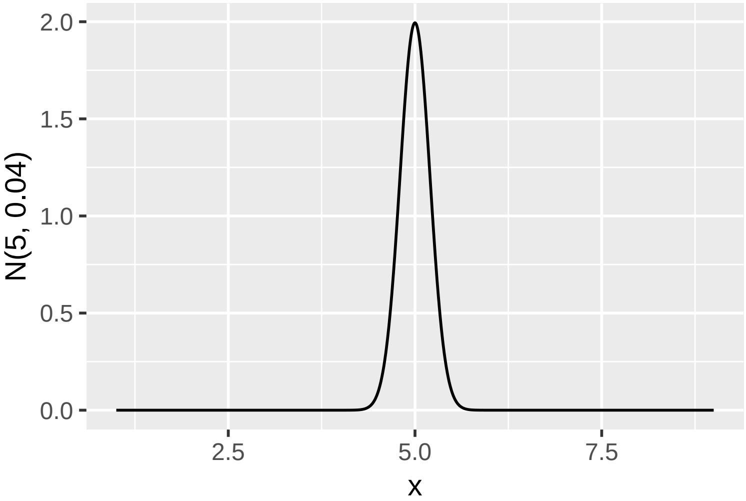
Normal, notes

The normal distribution shows up everywhere.



Normal, notes

Normal distribution with different parameters.



outline

Take Away

- Random variables are the main underlying object of interest to statisticians.
- RVs allow us to calculate probability statements: from flipping a coin to probability we encounter a person taller than $7'9$".
- RVs have distribution functions that we assume describe populations of interest,
- these distributions have known functional forms indexed by unkown parameters.
- Our goal is to estimate these parameters from sample data.

outline

references I

Michael G. Akritas. *Probability and Statistics with R for Engineers and Scientists*. Pearson Education, Inc., 2016.

Rao V. Dukkipati. *Probability and Statistics: For Scientists and Engineers*. New Academic Science, 2013.