Distributions

Heather Dye

November 11, 2022

Distributions

Definition A **random variable** X takes on values determined by a random experiment. For example:

- Rolling a die and recording the outcome
- Selecting a person at random from the population and recording their weight.
- Counting the number of customers that arrive at store between 9 and 10 AM.

Distributions

The **probability distribution function** $f(x)$ associated to a random variable has the following properties.

- $0 \leq f(x) \leq 1$ for all x
- The area under the function $f(x)$ is 1
- The function $F(a)$ is called the **cumulative distribution function** or **cdf** for short. Note that:

$$F(a) = P(X \leq a).$$

- A discrete distribution function has the property that

$$F(a) = \sum_{f(x) \neq 0, x \leq a} f(x).$$

- A continuous distribution function has the property that

$$F(a) = \int_{-\infty}^{a} f(x)dx.$$

Discrete distribution functions

A discrete distribution function is called a **probability mass function** or **pmf**.

Example 1 Let X denote the outcome of the roll of a fair six sided die. Then

$$f(x) = \begin{cases} \frac{1}{6}, & x = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}$$

Now, the CDF is

$$F(a) = \begin{cases} 0, & a \leq 1 \\ \frac{1}{6}, & 1 \leq a < 2 \\ \frac{1}{6}, & 2 \leq a < 3 \\ \frac{3}{6}, & 3 \leq a < 4 \\ \frac{4}{6}, & 4 \leq a < 5 \\ \frac{5}{6}, & 5 \leq a < 6 \\ 1, & 6 \leq a \end{cases}$$

Example 2 Let X denote the number of heads in sequence of 4 flips of a fair coin. The associated pdf is

$$f(x) = \binom{4}{x} \left(\frac{1}{2}\right)^4, \text{ for } x \in \{1, 2, 3, 4\}.$$

Continuous Distributions

Example 3 - Uniform Distribution Let

$$f(x) = \begin{cases} x/2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

This is a continuous distribution.

Example 4 - Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} Exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \; -\infty < x < \infty$$

Notice: Parameters! Different families of distributions are described by parameters.

Expected Value and Variance

The expected value (or mean of a function) is denoted $E(X)$ or μ.

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

.

The population variance is denoted $Var(x)$ or σ^2

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

.

Problems: Compute the mean and variance of the distributions.

Problem 1 - Uniform Distribution - Discrete Let X denote the outcome of the roll of a fair six sided die. Then

$$f(x) = \begin{cases} \frac{1}{6} \; x = 1, 2, 3, 4, 5, 6 \\ 0 \; \text{ otherwise} \end{cases}$$

Now, the CDF is

$$F(a) = \begin{cases} 0, & a \leq 1 \\ \frac{1}{6}, & 1 \leq a < 2 \\ \frac{1}{6}, & 2 \leq a < 3 \\ \frac{3}{6}, & 3 \leq a < 4 \\ \frac{4}{6}, & 4 \leq a < 5 \\ \frac{5}{6}, & 5 \leq a < 6 \\ 1, & 6 \leq a \end{cases}$$

Problem 2 - Uniform Distribution - Continuous Let

$$f(x) = \begin{cases} x/2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Construct a R-markdown File

Computing probabilities The normal distribution is an example of a probability density function or **pdf**.

The standard normal distribution has mean 0 and standard deviation 1. If $X \sim N(0,1)$ then

$$f(x) = \frac{1}{\sqrt{2\pi}} Exp\left(-\frac{x^2}{2}\right), \ -\infty < x < \infty$$

This integral requires some effort to calculate! We will use R to facilitate our computations.

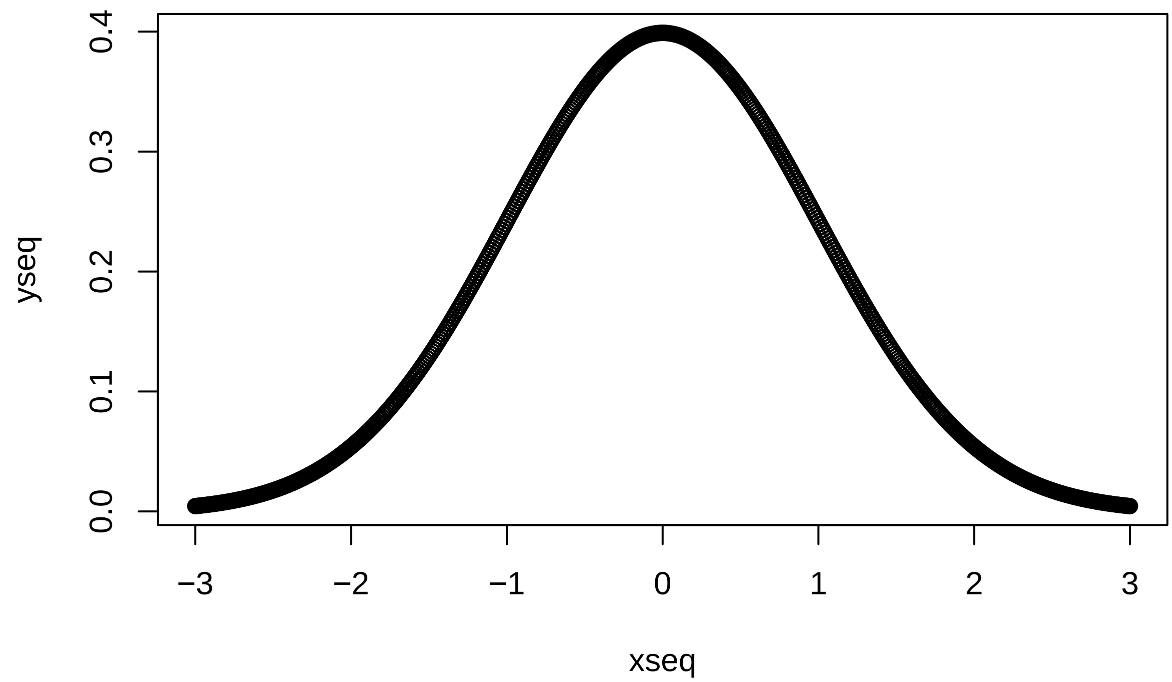
```r
# the cdf
pnorm(0)
```

```
## [1] 0.5
```

```r
pnorm(2)
```

```
## [1] 0.9772499
```

```r
# plot the density function
xseq=seq(-3,3, by=0.01)
yseq=dnorm(xseq)
plot(xseq, yseq)
```




```r
# quantiles
qseq = seq(0,1,by=0.01)
qnorm(qseq)
```

```
##   [1]        -Inf -2.32634787 -2.05374891 -1.88079361 -1.75068607 -1.64485363
##   [7] -1.55477359 -1.47579103 -1.40507156 -1.34075503 -1.28155157 -1.22652812
##  [13] -1.17498679 -1.12639113 -1.08031934 -1.03643339 -0.99445788 -0.95416525
##  [19] -0.91536509 -0.87789630 -0.84162123 -0.80642125 -0.77219321 -0.73884685
##  [25] -0.70630256 -0.67448975 -0.64334541 -0.61281299 -0.58284151 -0.55338472
##  [31] -0.52440051 -0.49585035 -0.46769880 -0.43991317 -0.41246313 -0.38532047
##  [37] -0.35845879 -0.33185335 -0.30548079 -0.27931903 -0.25334710 -0.22754498
##  [43] -0.20189348 -0.17637416 -0.15096922 -0.12566135 -0.10043372 -0.07526986
##  [49] -0.05015358 -0.02506891  0.00000000  0.02506891  0.05015358  0.07526986
##  [55]  0.10043372  0.12566135  0.15096922  0.17637416  0.20189348  0.22754498
##  [61]  0.25334710  0.27931903  0.30548079  0.33185335  0.35845879  0.38532047
##  [67]  0.41246313  0.43991317  0.46769880  0.49585035  0.52440051  0.55338472
##  [73]  0.58284151  0.61281299  0.64334541  0.67448975  0.70630256  0.73884685
##  [79]  0.77219321  0.80642125  0.84162123  0.87789630  0.91536509  0.95416525
##  [85]  0.99445788  1.03643339  1.08031934  1.12639113  1.17498679  1.22652812
##  [91]  1.28155157  1.34075503  1.40507156  1.47579103  1.55477359  1.64485363
##  [97]  1.75068607  1.88079361  2.05374891  2.32634787         Inf
```

Binomial Distribution The binomial distribution,$bin(5, 0.2)$ has the pmf:

$$f(x) = \binom{5}{x}(0.2)^x(0.8)^{5-x}.$$

```r
pbinom(2,5,0.2)
```

```
## [1] 0.94208
```

General results about distributions

The Central Limit Theorem If $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots X_n$ of size n from a distribution with a finite mean μ and a finite positive variance σ^2 then

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is $N(0,1)$ in the limit as $n \to \infty$.

Note that $E[\bar{X}] = \mu$. Similarly,

$$Var\left(\bar{X}\right) = Var\left(\frac{\sum X_i}{n}\right)$$

and

$$Var\left(\bar{X}\right) = \frac{1}{n^2} Var\left(\sum X_i\right)$$

so that:

$$Var\left(\bar{X}\right) = \frac{1}{n^2}\sigma^2.$$

What remains to be shown is that in the limit this distribution approaches a normal distribution. This is done by computing the mgf of $\bar{X}$ and comparing with the mgf of $N(0,1)$. Usually if $n > 30$ then we assume that $\bar{X}$ is approximately normal, without regard to the underlying distribution.

Chebychev's Theorem If the random variable X has a mean μ and a variance σ then, for every $k \geq 1$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Alternate: Within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$ of the distribution.

Proof sketch: Let $A = \{x : |x - \mu| \geq k\sigma\}$. Now,

$$\sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 f(x)$$

$$\sum (x-\mu)^2 f(x) = \sum_A (x-\mu)^2 f(x) + \sum_{A^C} (x-\mu)^2 f(x)$$

$$\sigma^2 \geq \sum_A (x-\mu)^2 f(x).$$

Notice that in A,$|x - \mu| \geq k\sigma$. Then,

$$\sigma^2 \geq \sum_A (x-\mu)^2 f(x) \geq \sum_A k^2\sigma^2 f(x).$$

As a result,

$$\sigma^2 \geq k^2\sigma^2 \sum_A f(x) \geq k^2\sigma^2 P(X \in A).$$

Then

$$\frac{1}{k^2} \geq P(X \in A) = P(|X - \mu| \geq k\sigma).$$