Lecture 8

Part 2 Probability and Distributions

Example

- A student sitting a statistics quiz decides to answer each of the ten multiple choice questions entirely by chance.
- Each question has five options, only one of which is correct.
- Let X be the number of questions the student answers correctly.
- Then $X \sim Bin(n = 10, p = 0.2)$.

Example

What is the probability the student gets half the answers correct?

$$P(X = 5) = \frac{10!}{5!(10-5)!} \times 0.2^5 \times (1-0.2)^5 = 0.0264$$

What is the probability that the student passes, i.e., gets five or more correct?

$$P(X \geq 5) = P(X = 5) + P(X = 6) + P(X = 7)$$

$$+P(X = 8) + P(X = 9) + P(X = 10)$$

$$= \text{a lot of calculations!}$$

Binomial Tables

- There are tables available that list $P(X \leq k)$ for different values of k, n and p.
- From tables, look up $n = 10$ and $p = 0.2$.

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - 0.9672 \text{ (from tables)} \\ &= 0.0328 \end{aligned}$$

Binomial Tables

- What is the probability the student gets half the answers correct?

$$\begin{aligned} P(X = 5) &= P(X \leq 5) - P(X \leq 4) \\ &= 0.9936 - 0.9672 \text{ (from tables)} \\ &= 0.0264 \end{aligned}$$

Binomial Tables

- The binomial tables are a tool to make life easier by helping us calculate binomial probabilities for frequently used values of n and p.
- However, they are not a substitute for knowing and being able to use the binomial probability distribution formula - not all values of n or p will be tabulated!

That's all for discrete probability distribution, let's talk about continuous ones now.

Continuous Random Variable

- A continuous random variable takes on an uncountable number of possible values.
- Cannot list all possible values in any systematic way.
- It is impossible to assign a non-zero probability to each possible value *and* still have all probabilities add up to 1.
- Therefore, for a continuous random variable X, the following is true for *any* value of x:

$$P(X = x) = 0$$

Continuous Probability Distribution?

- So what do we do about the probability distribution for a continuous random variable?
- Although $P(X = x) = 0$ for any value x, it turns out we can find probabilities of the form:

$$P(a < X < b)$$

Note:

- Discrete:

$$P(X \leq x) \neq P(X < x)$$

- Continuous:

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

Flashback to Histograms

- Histograms were a useful way to visually display the *distribution* of continuous data.
- Constructing a histogram involved:
  - Dividing range of possible values into intervals or "classes".
  - Counting number of observations that fall into each interval.
  - Setting height of each interval to be the frequency (count).

Flashback to Histograms

- Let's change the height of each interval in the histogram.
- Suppose we instead set the height of each interval to be:

$$\text{Interval Height} = \frac{\text{Count}}{\text{Total Count} \times \text{Interval Width}}$$

$$= \text{Proportion} \times \frac{1}{\text{Interval Width}}$$

Flashback to Histograms

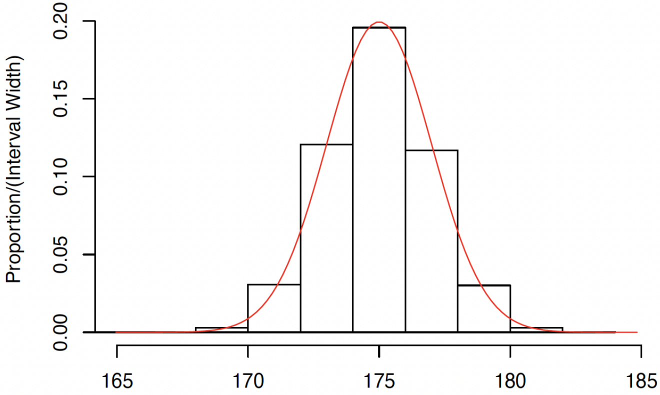
- The area of a rectangle corresponding to any particular interval is equal to:

$$\begin{aligned} \text{Area} &= \text{Interval Height} \times \text{Interval Width} \\ &= \text{Proportion} \times \frac{1}{\text{Interval Width}} \times \text{Interval Width} \\ &= \text{Proportion} \end{aligned}$$

- That is, the area of each rectangle is equal to the probability of an observation falling into that interval.

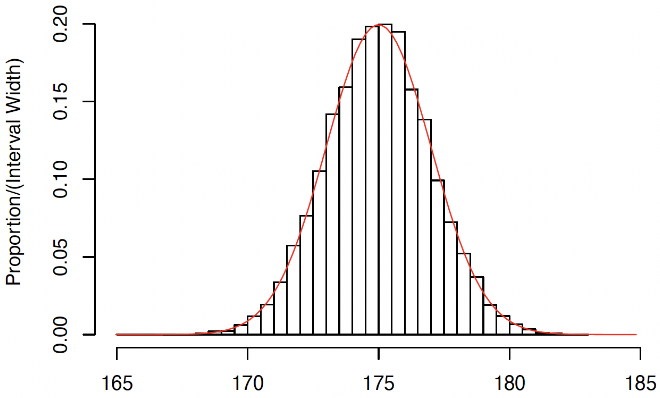
Measure 10,000 Heights

- Histogram with 10 intervals.



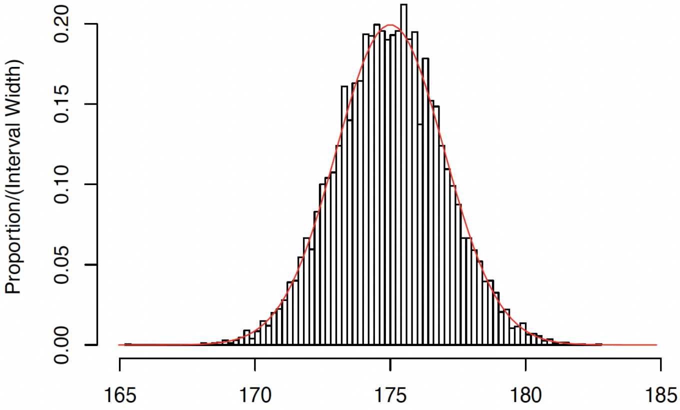
Measure 10,000 Heights

- Histogram with 50 intervals.



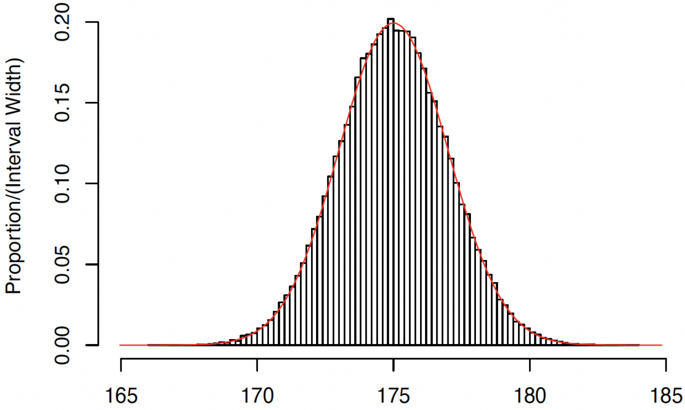
Measure 10,000 Heights

- Histogram with 100 intervals.



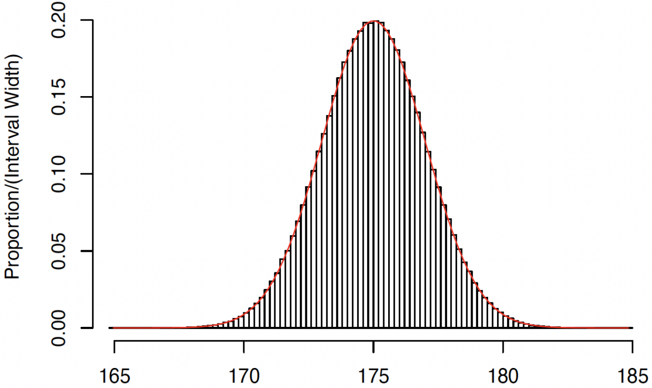
Measure 100,000 Heights

- Histogram with 100 intervals.



Measure 1,000,000 Heights

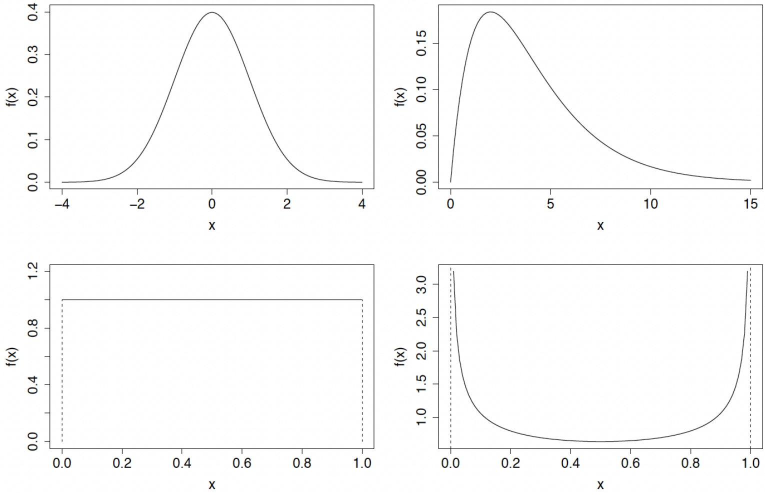
- Histogram with 100 intervals.



Probability Density Function

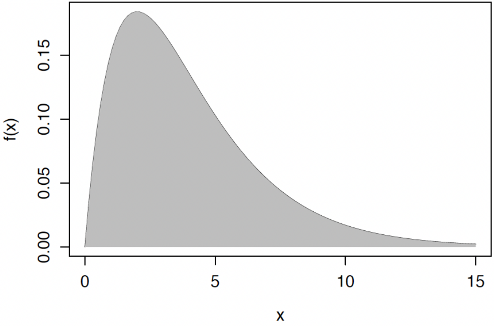
- As the number of observations and intervals both approach infinity, histograms of continuous data will approach a smooth curve (red line).
- The function that describes this curve is called the probability density function (PDF) and is denoted by $f(x)$.
- Can be thought of as the continuous analogue of the discrete probability distribution.

Examples of PDFs



Important Properties

- The PDF, $f(x)$, of a continuous random variable X must satisfy:
  1. $f(x) \geq 0$ for all x (non-negative)
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$ ((total area under curve equals 1).

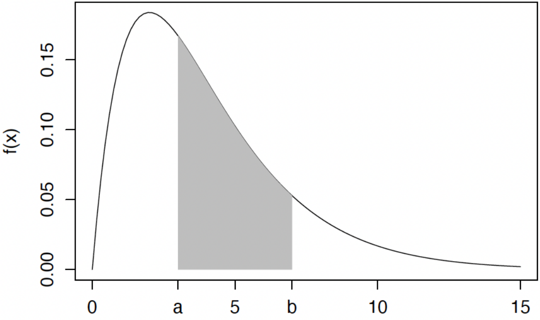


Why is the PDF Important?

- Just like with the discrete probability distribution, probability density functions represent populations.
- Once we know the probability density function of a continuous random variable, we know everything about that variable.
- We can use it to calculate probabilities and also population parameters like the mean (expected value) and variance.

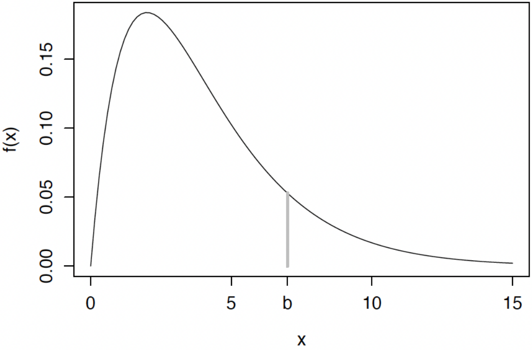
Calculating Probabilities

- The probability that X lies between a and b is equal to area under the PDF between the points a and b.
- It is calculated by $P(a < X < b) = \int_a^b f(x)dx$



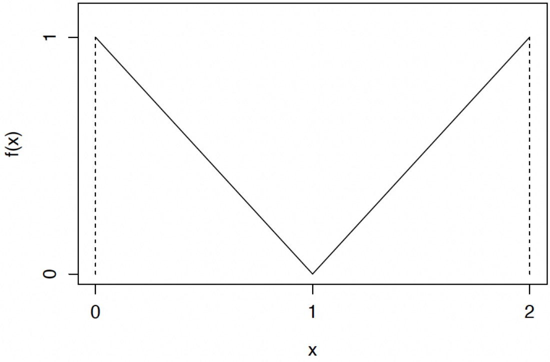
Calculating Probabilities

- Recall that the probability X will equal any specific value is always zero, i.e., $P(X = x) = 0$ for all x.
- We can see that as $a \to b$, then the area $\to 0$.



Calculating Probabilities

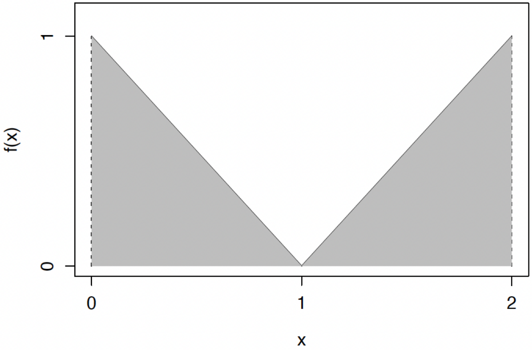
$$f(x) = \begin{cases} -x+1, & 0 \leq x < 1 \\ x-1, & 1 \leq x \leq 2 \end{cases}$$



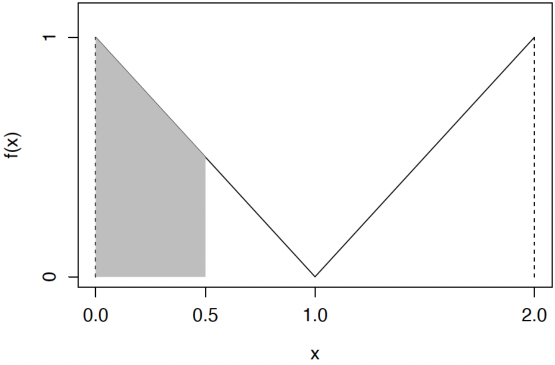
Is $f(x)$ a Valid PDF?

- From the graph, $f(x) \geq 0$ for all $0 \leq x \leq 2$.
- The total area under the curve is equal to:

$$\frac{1}{2} \times 1 \times 1 + \frac{1}{2} \times 1 \times 1 = 1$$

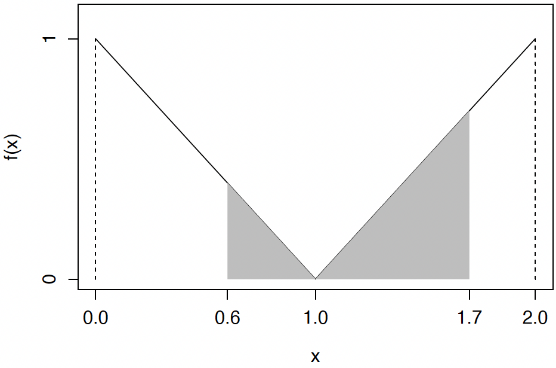


Find $P(X < 0.5)$



$$
\begin{aligned}
P(X < 0.5) &= P(X < 1) - P(0.5 < X < 1) \\
&= \frac{1}{2} \times 1 \times 1 - \frac{1}{2} \times 0.5 \times 0.5 = \frac{3}{8}
\end{aligned}
$$

Find $P(0.6 < X < 1.7)$



$$
\begin{aligned}
P(0.6 < X < 1.7) &= P(0.6 < X < 1) + P(1 < X < 1.7) \\
&= \frac{1}{2} \times 0.4 \times 0.4 + \frac{1}{2} \times 0.7 \times 0.7 = \frac{13}{40}
\end{aligned}
$$

Expected Value

- Let X be a continuous random variable with PDF $f(x)$. The **expected value** (or **population mean**) of X is defined to be:

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

- The **expected value** of $g(X)$, where $g(X)$ is some function of X, is defined to be:

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

Variance

- Let X be a continuous random variable with PDF $f(x)$ and $\mu = E(X)$. The **(population) variance** of X is defined to be:

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

- A shortcut formula for the variance is given below:

$$V(X) = E(X^2) - (E(X))^2 = \left(\int_{-\infty}^{\infty} x^2 f(x)dx\right) - \mu^2$$

Special Continuous Distributions

- There are a number of special continuous distributions, some of which we will encounter in this course:
  - Uniform distribution.
  - Normal distribution.
  - t-distribution.
  - F-distribution.
  - Chi-squared distribution.
  - Exponential distribution.
  - Cauchy distribution.

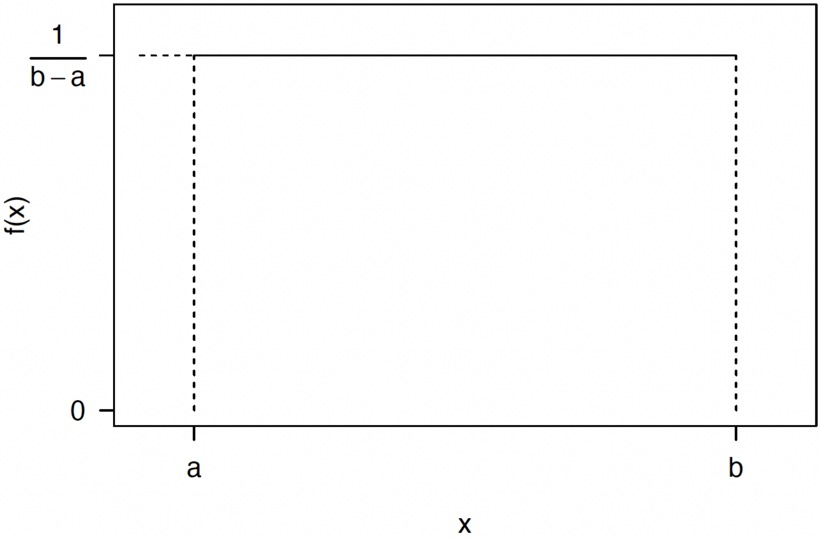
Uniform Distribution

- A continuous random variable X is said to have a **uniform distribution** between a and b if its PDF is given by the following function:

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

- We use the notation $X \sim U(a, b)$.
- There are two *parameters* that define a uniform distribution, namely, a and b.

Probability Density Function of Uniform Distribution



Expected Value and Variance

- Let $X \sim U(a, b)$.
- The **expected value** of X is given by:

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

- The **variance** of X is given by:

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

Expected Value

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_a^b x \times \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \left[\frac{x^2}{2} \right] \Big|_a^b \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{a+b}{2}
\end{aligned}
$$