Random Variables

Random variable: A function from a sample space S into R, $X : S \to R$.

Discrete random variable: The *range* of X is discrete (either finite or countable).

Let X be discrete and finite $S \to R$. Let the range of X be R_X.

Each value in the range defines an *event*:

$A_x = \{s \in S : X(s) = x\}$.

- $A_x, x \in R_X$ are disjoint.
- $\cup_{x \in R_X} A_x = S$.

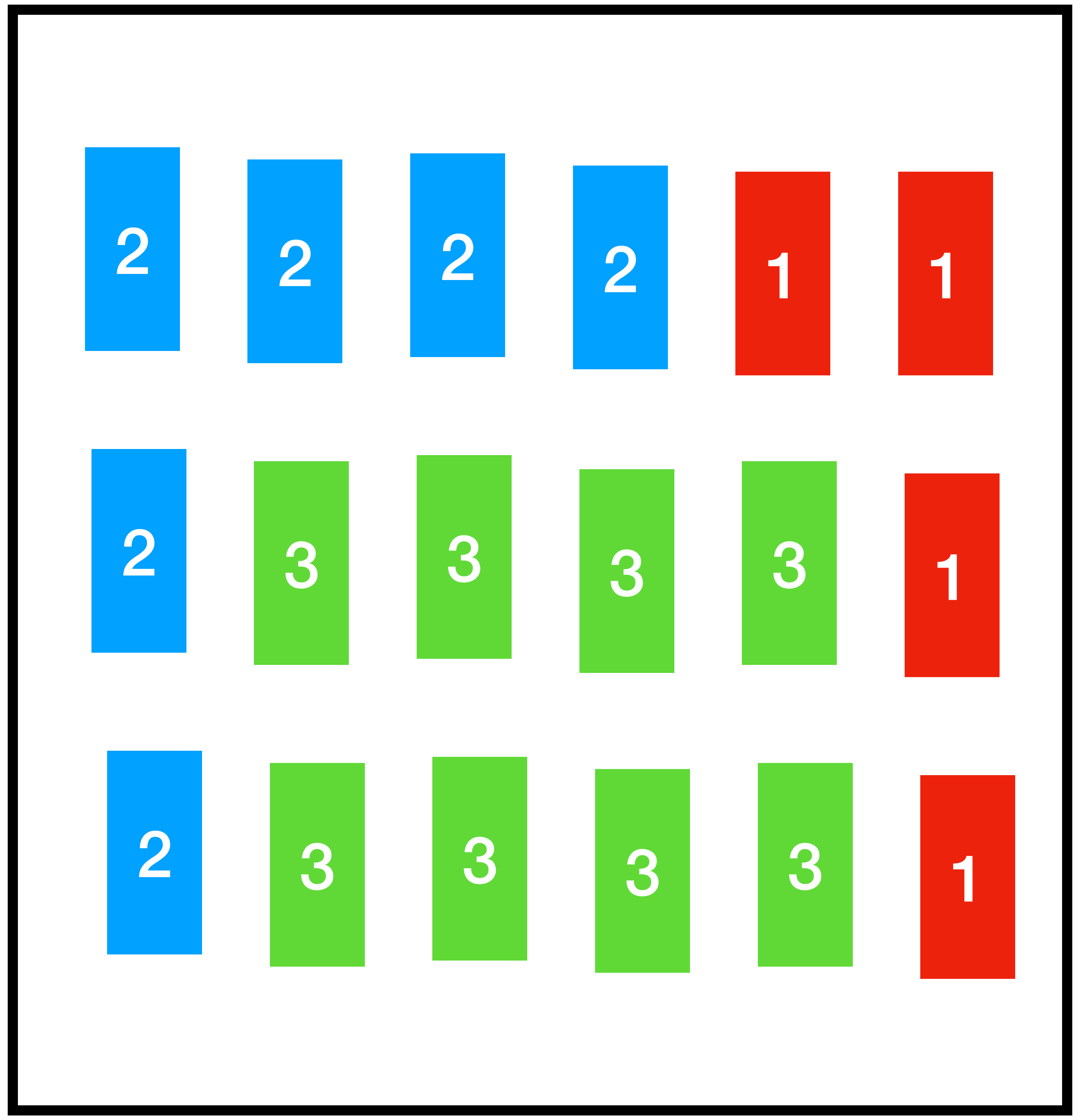
Define $f_X(x) = P(A_x) = \ldots$ (shorthand) $P(X = x)$.

$f_X(x), x \in R_X$ is the distribution of X. $\sum_{x \in R_X} f_X(x) = 1$.

Many different random variables with range R_X can have the same *distribution*.

P(Red) = 4/18, P(Blue) = 6/18, P(Green)=8/18

Red = 1, Blue=2, Green=3



$$\frac{1}{18}(4 \cdot 1 + 6 \cdot 2 + 8 \cdot 3) = 40/18$$ - average of box

$$f_X(1) = \frac{4}{18}, f_X(2) = \frac{6}{18}, f_x(3) = \frac{8}{18}.$$ - distribution of X

$$E(X) = 1 \cdot 4/18 + 2 \cdot 6/18 + 3 \cdot 8/18 = 40/18$$

$$Var(X) = (1 - \frac{40}{18})^2 \cdot \frac{4}{18} + (2 - \frac{40}{18})^2 \cdot \frac{6}{18} + (3 - \frac{4}{18})^2 \cdot \frac{8}{18} = .61724 =$$

MSD(BOX)

Bernoulli and Binomial distributions I

Bernoulli distribution: $R_X = \{0, 1\}, f_X(1) = p, f_X(0) = 1 - p$.
Experiment: Draw a card from a box with N cards, pN red, $(1 - p)N$ black.
If card is red $X = 1$.
$EX = 1 \cdot p + 0 \cdot (1 - p) = p$.
$Var(X) = (1 - p)^2 p + (0 - p)^2 (1 - p) = p \cdot (1 - p)$.

Bernoulli and Binomial distributions II

Binomial distribution:
$R_X = \{0, 1, \ldots, n\}, f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, \ldots, n$

Let X be the number of reds in n independent draws with replacement from above box.
To observe exactly k reds in n draws we need to observe $n-k$ blacks.
The probability of a particular sequence of k reds and $n-k$ blacks is always $p^k(1-p)^{n-k}$.
There are $\binom{n}{k}$ different sequences of length n with k reds.
Consequently $f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ and X has the binomial distribution.

Poisson distribution - unbounded domain I

$P(X = k) = f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, \ldots, \infty.$
Check that this is a distribution.
$\sum_{k=0}^{\infty} f_X(k) = 1.$?
Is used to model number of occurrences of a repeated event in a given interval of time.
For example - number of calls arriving at a call center per hour.

Functions of a random variable I

If $X : S \to R$ is a random variable, and $g : R \to R$ is a function then $Y = g(X)$ is also a random variable with range $R_Y = g(R_X)$.
Example: For the box example from last week take
$g(x) = \begin{cases} 3 & \text{if } x = 1 \\ 1 & \text{if } x > 1 \end{cases}.$
Then $Y = g(X)$ has range $R_Y = \{1, 3\}$ and distribution
$f_Y(1) = P(Y = 1) = P(X = 3 \text{ or } X = 5) = 6/10.$
$f_Y(3) = P(Y = 3) = P(X = 1) = 4/10.$
So we can compute the mean: $E(Y) = 1 \cdot 6/10 + 3 \cdot 4/10 = 1.8$

Sometimes it is difficult to compute the new distribution of Y.
It's still easy to compute $E(Y)$ in terms of the distribution of X:
$E(Y) = Eg(X) = \sum_{k=1}^{K} g(x_k) f_X(x_k).$
So in our example we compute

$$\begin{aligned} E(Y) =& g(1) \cdot 2/5 + g(3) \cdot 1/2 + g(5) \cdot 1/10 \\ =& 3 \cdot 2/5 + 1 \cdot 1/2 + 1 \cdot 1/10 = 1.8 \end{aligned}$$

Continuous Random Variables I

Let X = the time it takes for a read/write head to locate a desired record on a computer disk (once the head has been positioned over the correct track).

Suppose the disk makes one complete revolution every 10 milliseconds. We may want to answer questions like:

1. What is the chance of locating the record in less than 2ms?
2. What is the chance of locating the record in between 5ms and 7ms?
3. What is the average access time?
   (That is, the average time to locate the record.)

Continuous Random Variables II

X is a random variable: it assigns a number to each outcome of an experiment.

What are the possible values for X?
$R_X = [0, 10)$

This sample space is *not* a *discrete* set. S is not countable, listable. So, X is not a discrete random variable.

X is a **continuous random variable** since the sample space is an interval.

We want to describe the probability distribution for X.

Probability Distribution for Continuous r.v.s I

Problem: Recall that for random variables defined on an interval, we cannot have that for each $x \in R_X$
$P(X = x) = c > 0 \rightarrow P(R_X) = \infty \neq 1$
We cannot simply list outcomes and give their probabilities.

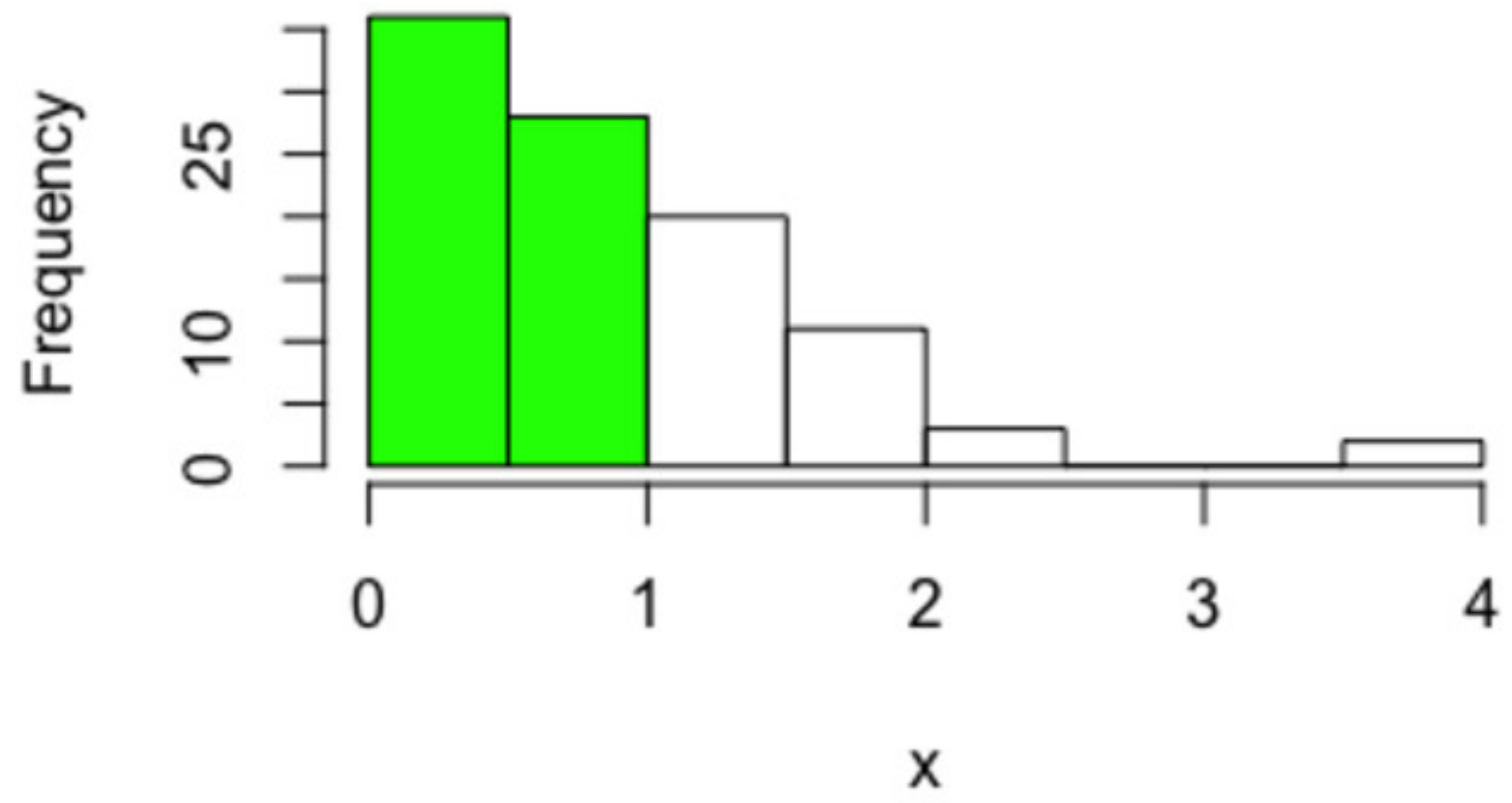
Need a different way to describe distribution for a continuous r.v.

Solution: Consider the cumulative distribution function:

Discrete case:

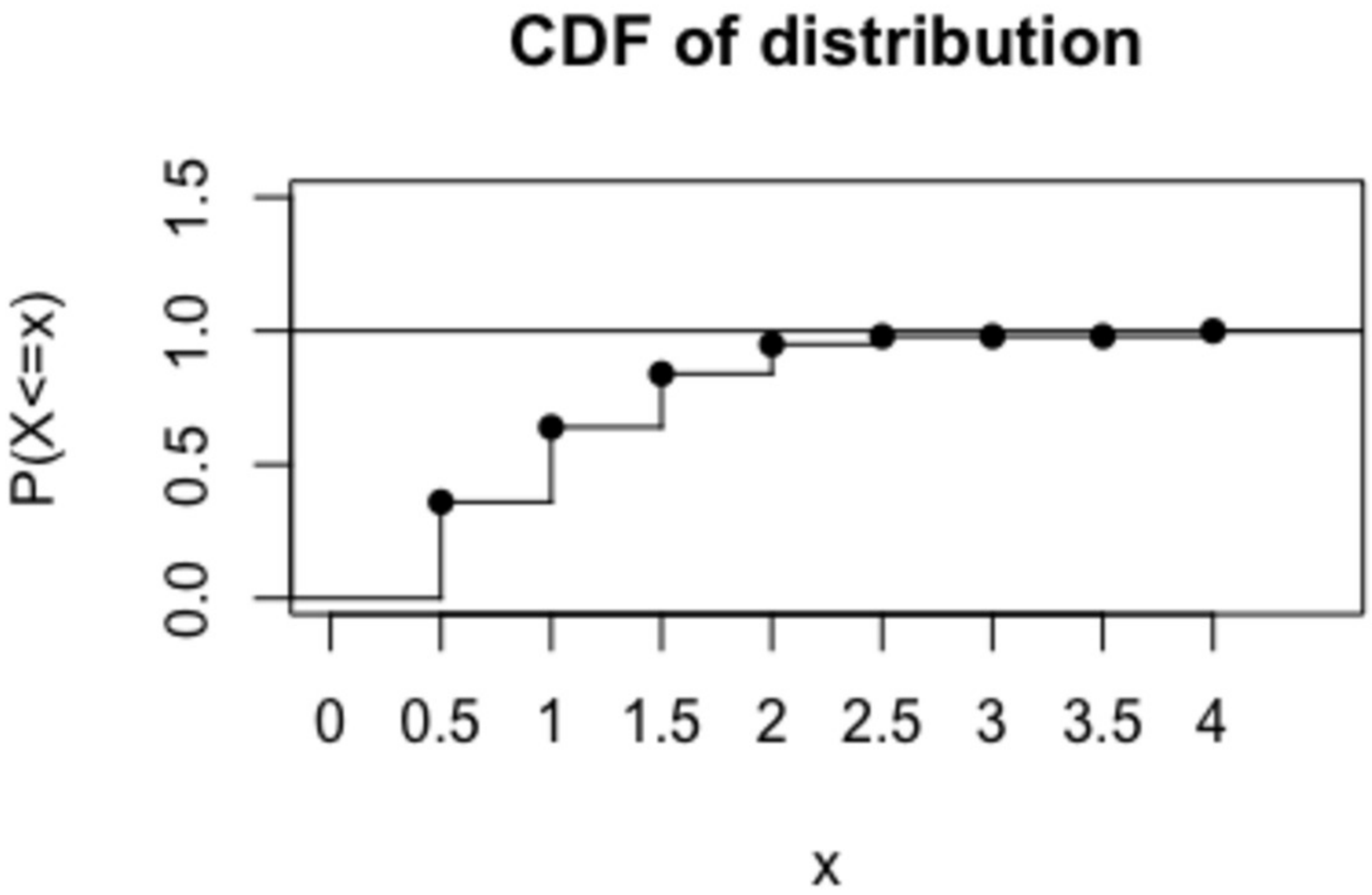
$$\begin{aligned} F_X(x_0) &= P(X \leq x_0) \\ &= \sum_{x \in R_X \ \& \ x \leq x_0} P(X = x) \\ &= \sum_{x \leq x_0} f_X(x) \end{aligned}$$

Probability Distribution for Continuous r.v.s II



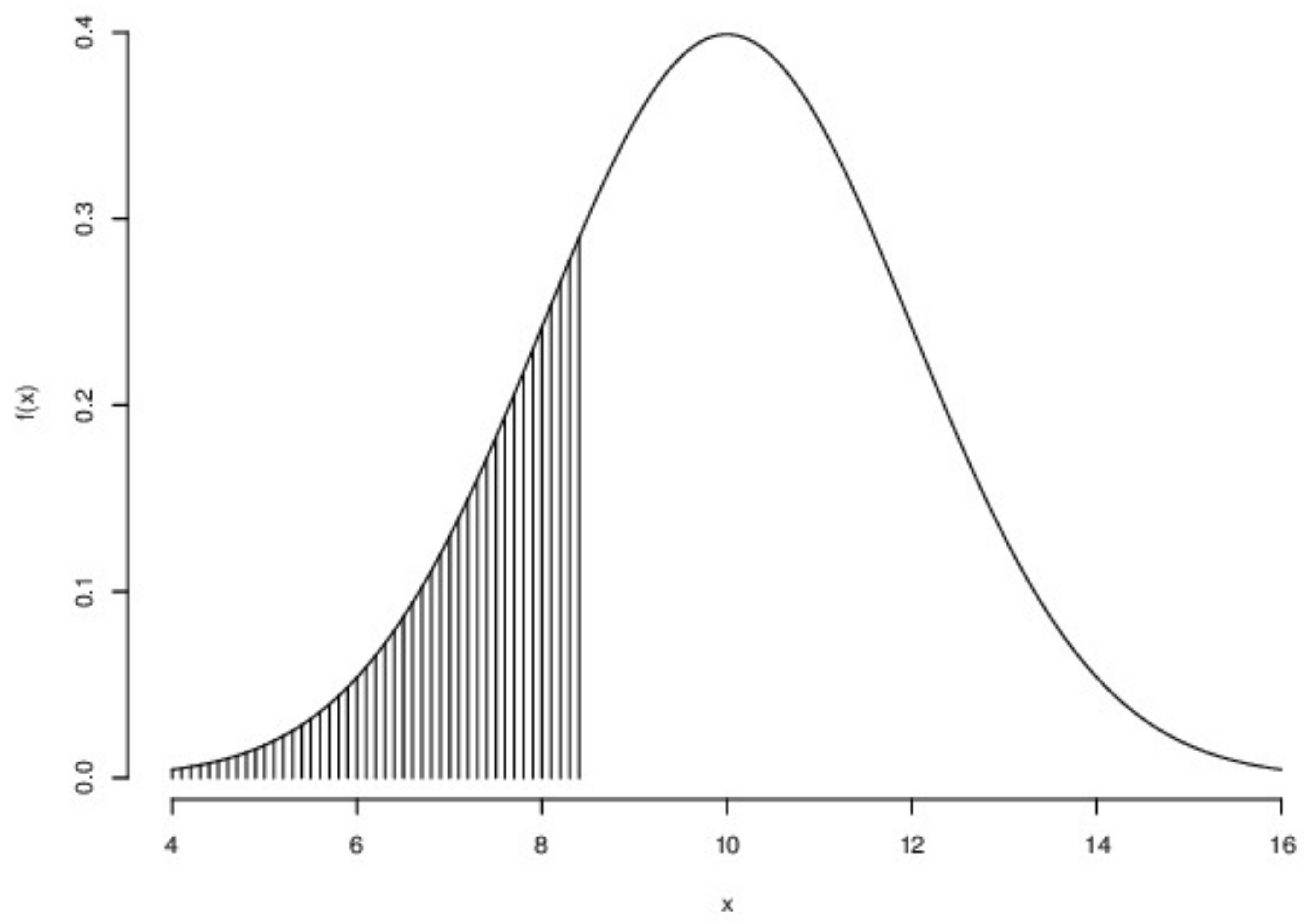
Here, $f(x)$ is a **probability mass function** (pmf) - the distribution of a discrete random variable.
Range $R_X = \{.50, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$
Probabilities $.36, .28, .20, .11, .03, 0.0, 0.0, .02$
Green area $F_X(1) = P(X \leq 1) = .36 + .28 = .64$

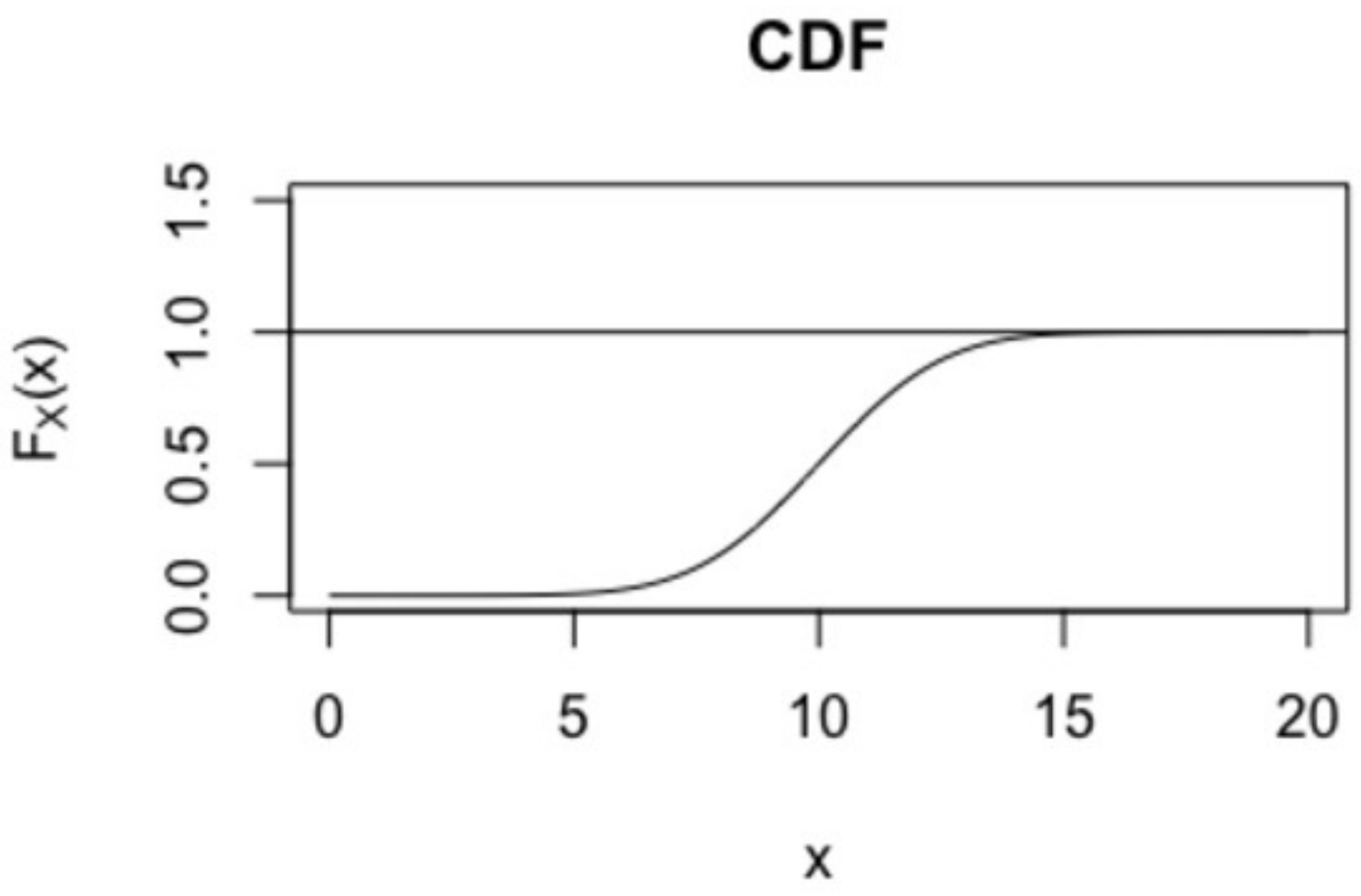
Probability Distribution for Continuous r.v.s III



Probability Distribution for Continuous r.v.s IV

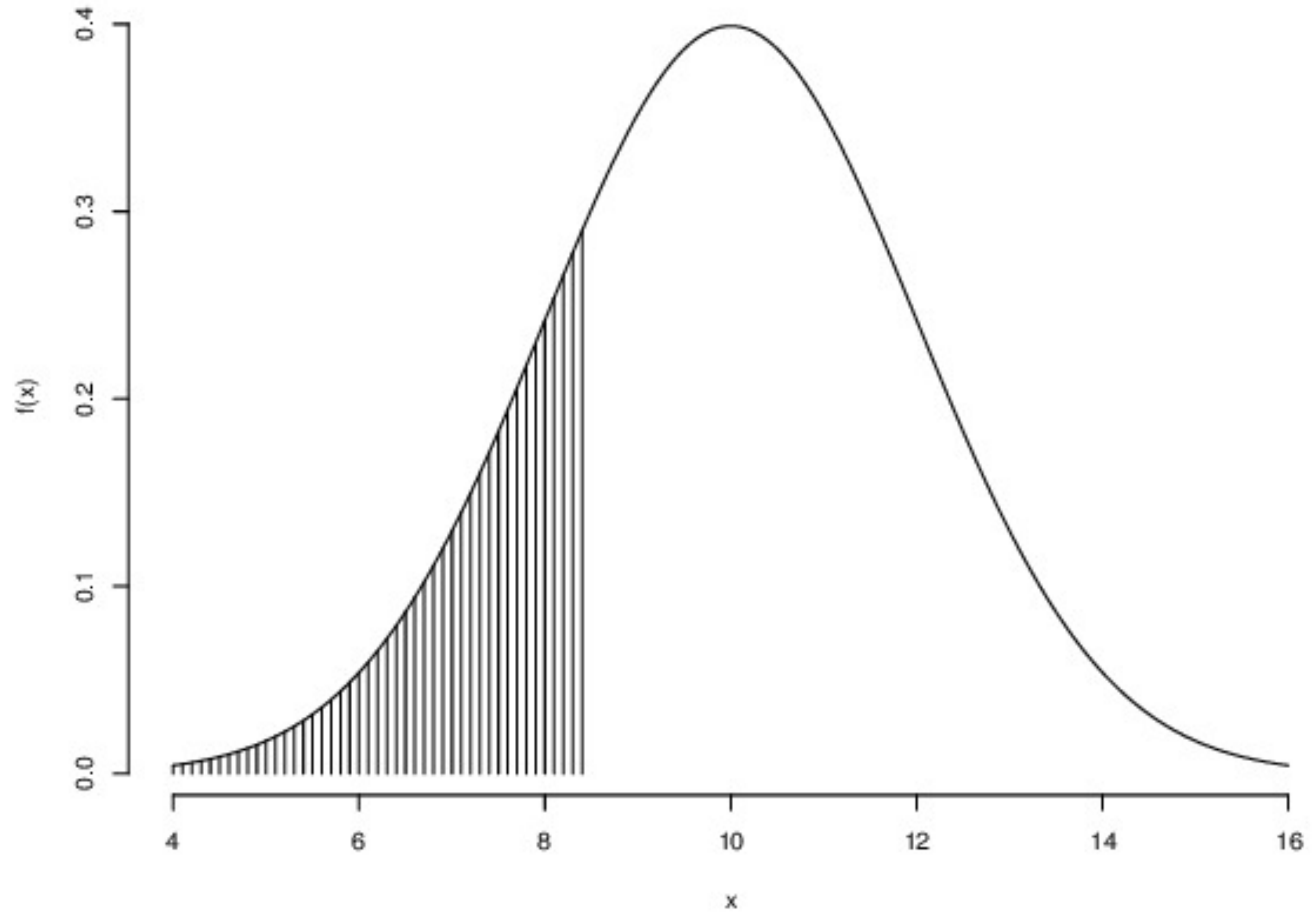
Continuous case: $F(x_0) = P(X \leq x_0) = \int_{x \leq x_0} f(x)dx$

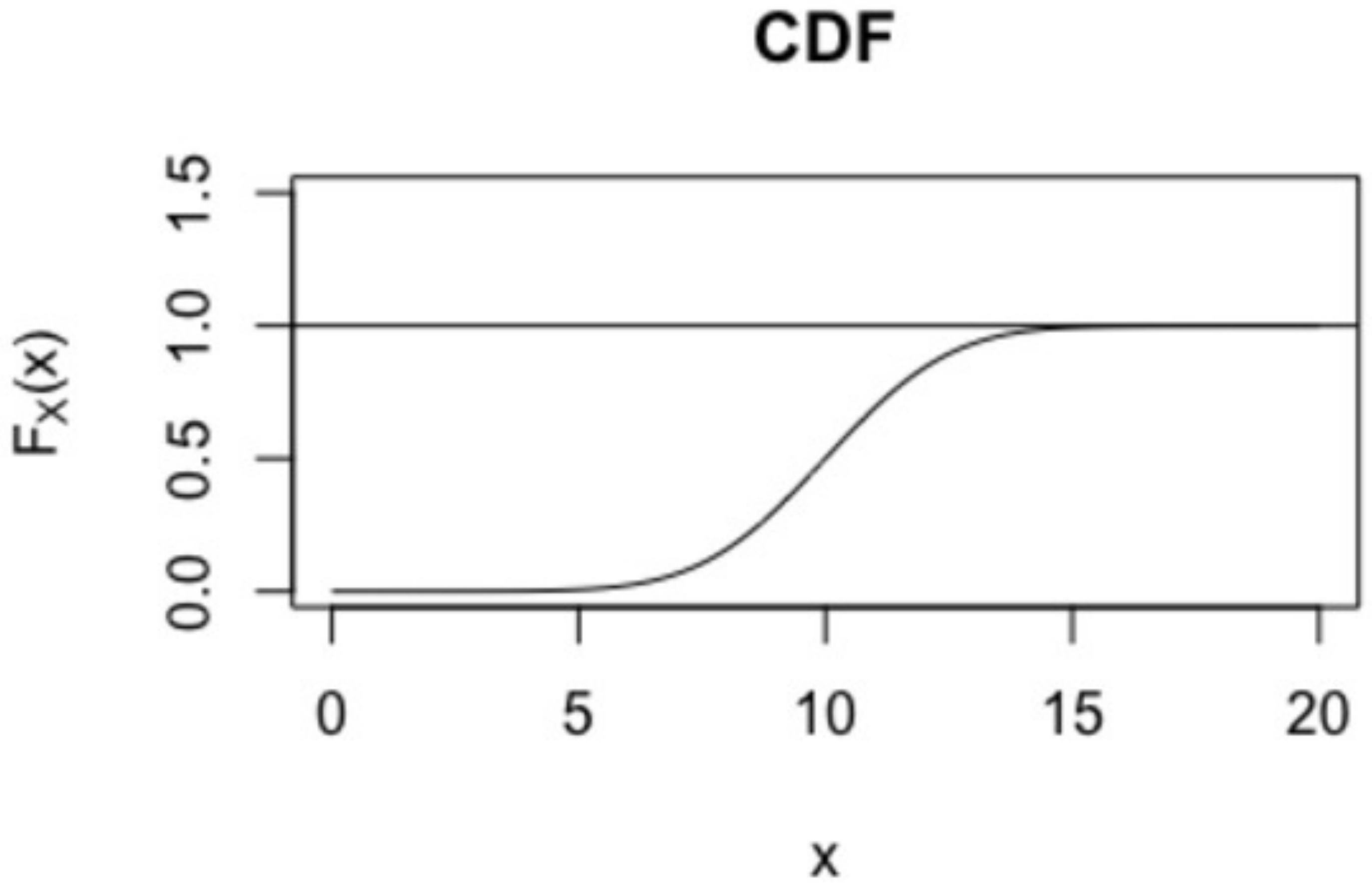


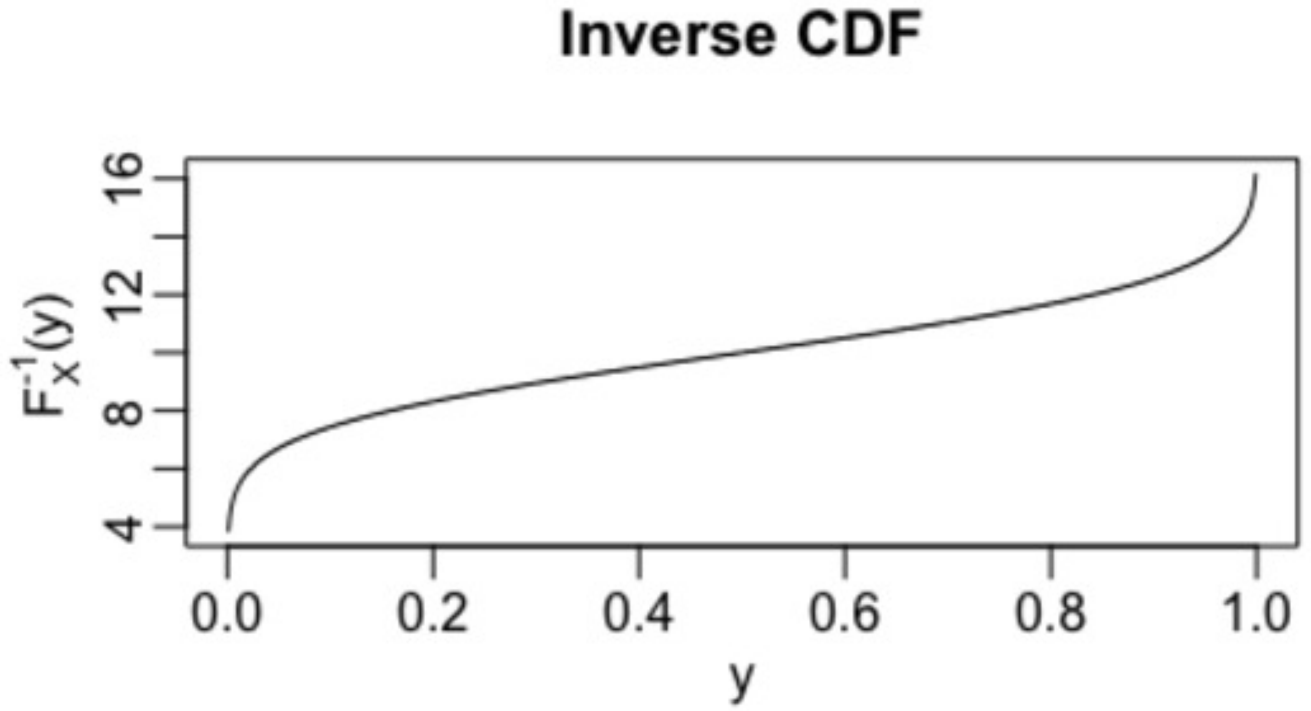


Here, $f_X(x)$ is a **probability density function** (pdf)
$F_X(x)$ is a CDF.

Probability Distribution for Continuous r.v.s V

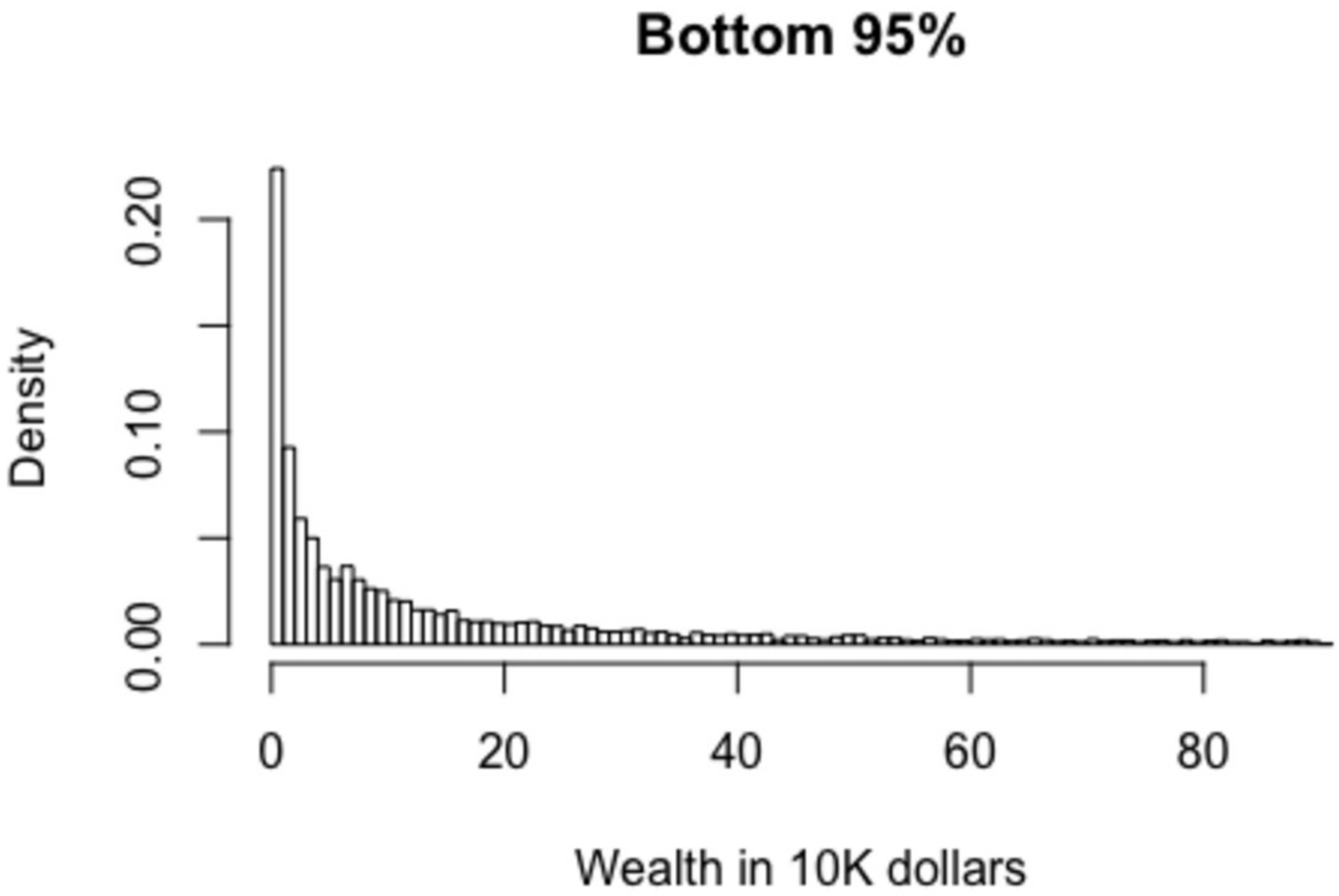






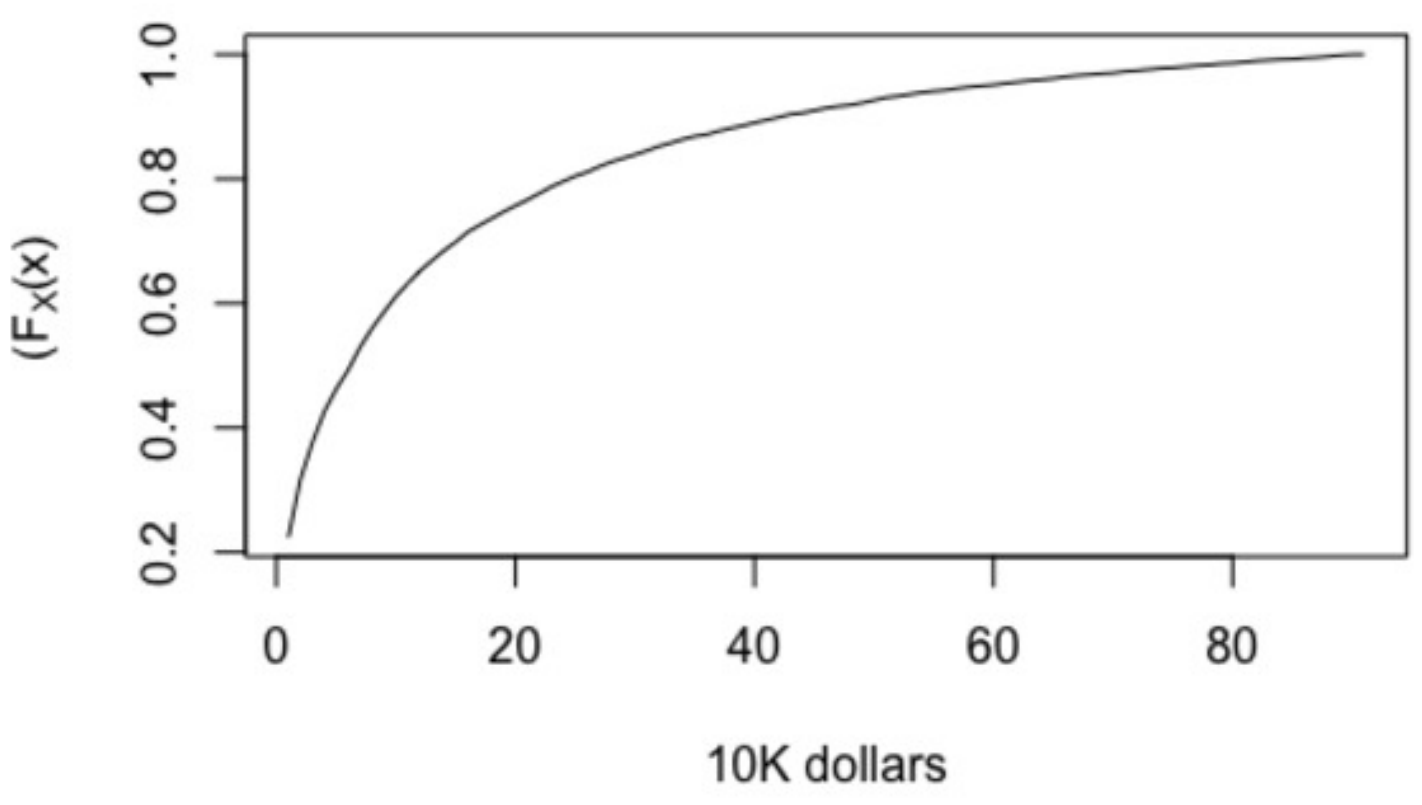
Probability Distribution for Continuous r.v.s VI

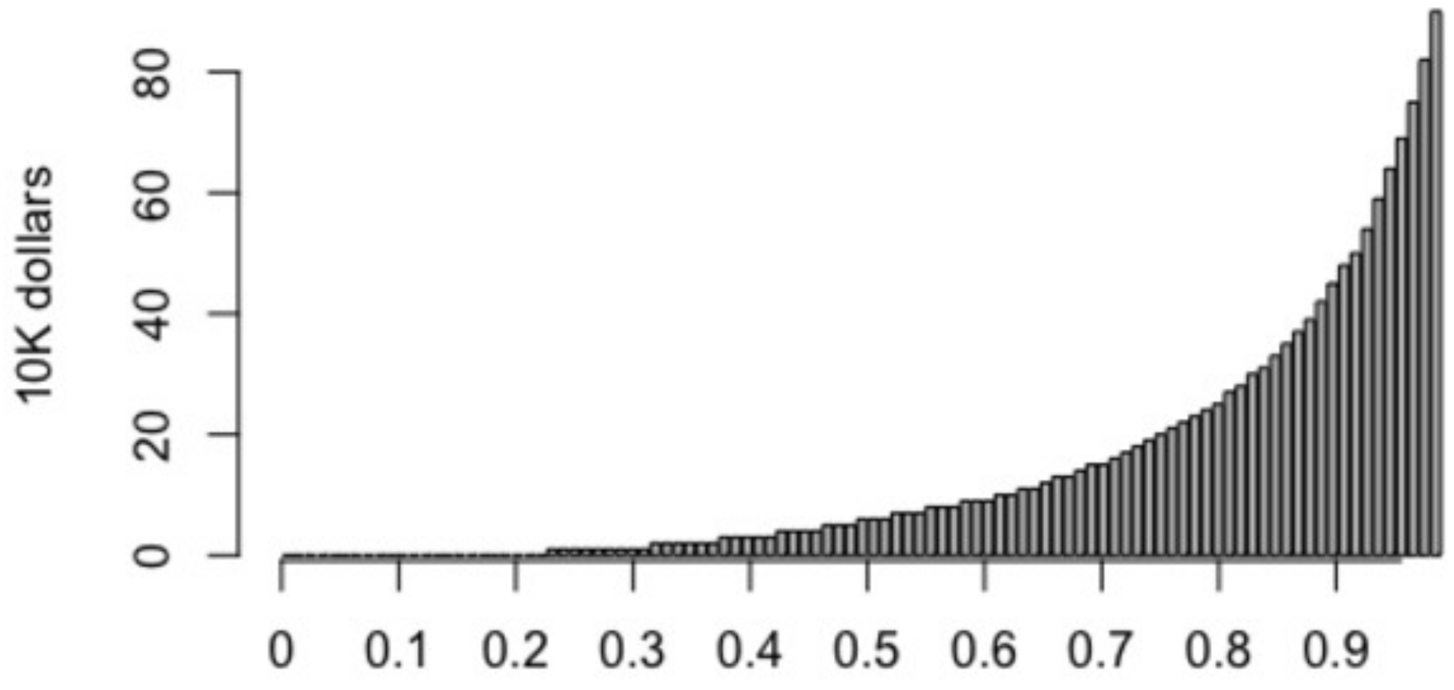
Back to wealth distribution - histogram:



Probability Distribution for Continuous r.v.s VII

The CDF and inverse CDF obtained from the histogram:





https://www.youtube.com/watch?v=QPKKQnijnsM

Continuous Random Variables I

By the Fundamental Theorem of Calculus,

$$f(x) = \frac{d}{dx} \int_{-\infty}^{x} f(t)\, dt = \frac{d}{dx} F(x) = F'(x)$$

We will want to find probabilities of the form $P(a \leq X \leq b)$.

Example: Access time of read/write head: What is the chance of locating the record in between 5ms and 7ms? $P(5 \leq X \leq 7)$

Continuous Random Variables II

$$
\begin{aligned}
P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\
&= F(b) - F(a) \\
&= \int_{-\infty}^{b} f(x)\, dx - \int_{-\infty}^{a} f(x)\, dx \\
&= \int_{a}^{b} f(x) dx
\end{aligned}
$$

Note that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

since $P(X = a) = \int_a^a f(x)\, dx = 0$ and $P(X = b) = 0$.

Continuous Random Variables III

Be careful! This is **not** true for discrete random variables where
$P(a \leq X \leq b) = F(b) - F(a-1)$, but
$P(a < X \leq b) = F(b) - F(a)$.

Example of a Continuous Random Variable I

Let X = the time it takes for a read/write head to locate a desired record on a computer disk (once the head has been positioned over the correct track).

The sample space for X is $R_X = [0, 10)$, in milliseconds.

Consider $P(2 \leq X \leq 4), P(1 \leq X \leq 3)$ and $P(5 \leq X \leq 7)$.
All are intervals of length 2ms.

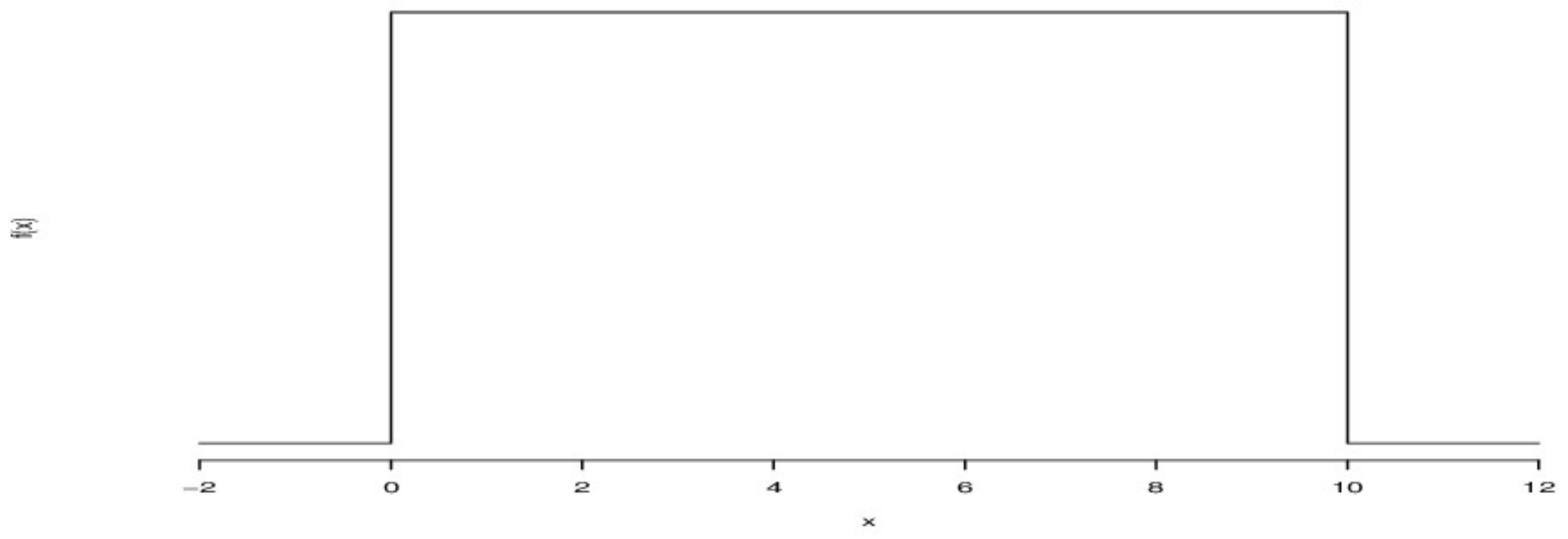
Is there any reason why one of these intervals is more likely to occur than any other?

Then, let all intervals of equal length contained in $[0, 10)$ have the same probability.

This (reasonable) assumption allows us to propose a form for the probability density function, $f(x)$.

Example of a Continuous Random Variable II

$f(x) = c, x \in [0, 10]$



What is the height of this function? We must have
$P(R_X) = \int_0^{10} f(x)\,dx = 1$
$\int_0^{10} f(x)\,dx = \int_0^{10} c\,dx = 10c$
Setting $10c = 1 \rightarrow c = 1/10 = 0.10$

Example of a Continuous Random Variable I

1. What is the chance of locating the record in less than 2ms?

$$P(X < 2) = \int_0^2 \frac{1}{10}\,dx = \left.\frac{x}{10}\right|_0^2 = 2/10 = 0.20$$

2. What is the chance of locating the record in between 5ms and 7ms?

$$P(5 \le X \le 7) = \int_5^7 \frac{1}{10}\,dx = \left.\frac{x}{10}\right|_5^7 = 2/10 = 0.20$$

3. What is the average access time?

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^{10} \frac{x}{10}\,dx = \left.\frac{x^2}{20}\right|_0^{10} = 10^2/20 = 100/20 = 5$$

Example of a Continuous Random Variable II

Since the uniform distribution is symmetric and the mean is always the balancing point of the distribution, we must have $E(X) = 5$, the midpoint of the range of X.

In general, if $X \sim \mathrm{Uniform}(a, b)$, then

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

$$E(X) = \frac{a+b}{2} \qquad \text{(midpoint of the range of } X\text{)}$$