Random Variables

A random variable is a quantity whose value is determined by the outcome of an experiment.

Before the experiment is carried out, all we know is the range of possible values.

Birthday example

The birthday $X_1$ of the first person is a random variable, with range $\{1, 2, \ldots, 365\}$.

The birthday $X_2$ of the second person is another random variable, associated with the same experiment.
Birthday example

The quantity $X$ defined by

$$X = \begin{cases} 
0 & \text{if no two people have the same birthday} \\
1 & \text{otherwise} 
\end{cases}$$

is a random variable, and $X = 1$ if and only if $A$ occurs.

Bernoulli Random Variable

Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable.
Example 3.4

9-volt batteries are tested until one with an acceptable voltage is obtained. The sample space is $S = \{A, UA, UUA, \ldots \}$. Define $X$ by

$$X = \text{number of batteries tested}.$$  

Then

$$X(A) = 1$$
$$X(UA) = 2$$
$$X(UUA) = 3$$

$$\ldots$$

The range of $X$ is $\{1, 2, 3, \ldots \}$, a countably infinite set.
Example 3.5

$Y$ is the altitude at a randomly chosen location in the U.S. The range of $Y$ is the real number interval $[-282, 14494]$, an uncountably infinite set.
Discrete and Continuous Random Variables

Discrete Random Variable:
A random variable whose range is finite or countably infinite.

Continuous Random Variable:
A random variable $Y$ satisfying:
1. its range is the union of one or more real number intervals;
2. $P(Y = c) = 0$ for every $c$ in the range of $Y$.

Note
For a continuous random variable $Y$, $P(c - \epsilon \leq Y \leq c + \epsilon)$ could be positive for any $\epsilon > 0$, but decreases to 0 as $\epsilon$ becomes smaller.
Probability Distributions for Discrete Random Variables

To calculate the probability of any event defined by a discrete random variable $X$, we need a list of the possible values (the range) of $X$, and the probability of each.

Probability Mass Function

The *probability mass function* (pmf) $p$ of a discrete random variable $X$ is

$$p(x) = P(X = x) = P\{s \in S : X(s) = x\}$$

for any $x$ in the range of $X$. 
Every pmf must satisfy:

\[ p(x) \geq 0 \text{ for all } x \text{ in the range of } X \]

and

\[ \sum_{x \in \text{range of } X} p(x) = 1. \]

Any function \( p \) with these properties could be the pmf of some discrete random variable.
Parameter of a Distribution

Recall that a Bernoulli random variable $X$ is one that takes only the values 0 and 1.

Suppose that $P(X = 1) = \alpha$, for some $\alpha$ between 0 and 1; then the pmf of $X$ is:

$$p(x) = \begin{cases} 
\alpha & x = 1 \\
1 - \alpha & x = 0.
\end{cases}$$

This is a different pmf for different $\alpha$; we write it as $p(x; \alpha)$, and call it a family of distributions (or pmfs), indexed by the parameter $\alpha$. 
Geometric Distribution

Recall Example 3.4: 9-volt batteries are tested until one with an acceptable voltage is obtained. The sample space is $S = \{A, UA, UUA, \ldots \}$. Define $X$ by

$$X = \text{number of batteries tested.}$$

Then

$$X(A) = 1$$
$$X(UA) = 2$$
$$X(UUA) = 3$$
$$\vdots$$
So

\[ p(1) = P(X = 1) = P(A) = p \]
\[ p(2) = P(X = 2) = P(UA) = P(U)P(A) = (1 - p)p \]
\[ p(3) = P(X = 3) = P(UUA) = P(U)P(U)P(A) = (1 - p)^2p \]
\[ \vdots \]

In general,

\[ p(x) = p(x; p) = p(1 - p)^{x-1}, \quad x = 1, 2, \ldots \]

This is the geometric family of distributions, with parameter \( p \).
Looking Ahead

In many problems we observe a random variable $X$, and we may know that its distribution belongs to a particular parametric family.

But we typically do not know the value of the parameter; a central problem in statistics is making inferences about the values of parameters: parameter estimation, etc.
Cumulative Distribution Function

For a given random variable $X$, we often need the probability $P(X \leq x)$ for some real number $x$.

It is convenient to define it as a function, the \textit{cumulative distribution function} (cdf)

$$ F(x) = P(X \leq x), -\infty < x < \infty. $$

If $X$ is discrete, we can construct $F(x)$ from the pmf $p(x)$:

$$ F(x) = \sum_{y \text{ in the range of } X \text{ with } y \leq x} p(y). $$
The graph of $F(x)$ against $x$ is constant between values of $X$, with jump of size $P(X = x) = p(x)$ at each possible value of $X$.

So from the locations of the jumps in $F(x)$ we can identify the range of $X$, and from the sizes of the jumps we can identify its pmf.

That is, the pmf and cdf carry the same information. Either can be derived from the other, and each provides a complete description of the probability distribution of $X$. 
Expected Values

The Expected Value of $X$

Let $X$ be a discrete random variable with range $D$ and pmf $p(x), x \in D$. The expected value of $X$ is

$$E(X) = \sum_{x \in D} xp(x).$$

The expected value is sometimes written $\mu_X$, and sometimes called the *mean* value.

Expected value is just a *weighted average* of the values of $X$, weighted by their probabilities.
Interpreting Expected Value

Recall the interpretation of probability as the relative frequency in a large number $n$ of trials.

For each possible value $x$ of $X$, let $n(x)$ be the number of times that $X$ takes the value $x$. Then the sum of the observed values of $X$ is

$$\sum_{x \in D} xn(x),$$

and their average is

$$\frac{\sum_{x \in D} xn(x)}{n} = \sum_{x \in D} \frac{n(x)}{n}.$$
But when \( n \) is large, we expect

\[
\frac{n(x)}{n} \approx p(x),
\]

so

\[
\sum_{x \in D} x \frac{n(x)}{n} \approx \sum_{x \in D} xp(x) = E(X)
\]

That is, we expect the sample average in many trials to be close to the expected value.
Bernoulli Family
If $X$ has the Bernoulli distribution with parameter $\alpha$, then $D = \{0, 1\}$, and

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = \alpha.$$  

Geometric Family
If $X$ has the geometric distribution with parameter $p$, then $D = \{1, 2, \ldots \}$, and

$$E(X) = \sum_{x=1}^{\infty} xP(X = x) = \sum_{x=1}^{\infty} xp(1 - p)^{x-1} = \frac{1}{p}.$$  

Note that when $D$ is infinite, the series defining $E(X)$ may converge to infinity, or not converge at all.
Expected Value of a Function of $X$

If $X$ is a random variable, and $h(x)$ is some function (such as $h(x) = x^2$ or $h(x) = \cos(2\pi x)$), then $Y = h(X)$ is also a random variable.

Not surprisingly,

$$E(Y) = E[h(X)] = \sum_{x \in D} h(x)p(x).$$
Rules of Expected Value

In the special case where \( h(x) \) is a \textit{linear} function, say

\[
h(x) = ax + b,
\]

we find

\[
E[h(X)] = E(aX + b) = \sum_{x \in D} (ax + b)p(x) = \left[ a \sum_{x \in D} xp(x) \right] + \left[ b \sum_{x \in D} p(x) \right] = aE(X) + b.
\]
Variance
If the random variable $X$ has expected value $\mu$, then its variance is

$$V(X) = E[(X - \mu)^2].$$

Standard Deviation
The variance of $X$ is sometimes written $\sigma_X^2$, and its standard deviation is

$$\sigma_X = \sqrt{V(X)}.$$

Standard deviation is in the same units as $X$, and represents (in a root-mean-square sense) how far you can expect $X$ to differ from $\mu$. 
Shortcut Formula for Variance

\[ V(X) = E[(X - \mu)^2] \]
\[ = E(X^2 - 2\mu X + \mu^2) \]
\[ = E(X^2) - 2\mu E(X) + \mu^2 \]
\[ = E(X^2) - \mu^2. \]

Bernoulli Family

If \( X \) has the Bernoulli distribution with parameter \( \alpha \), then \( \mu = \alpha \), and \( X^2 = X \), so

\[ V(X) = \alpha - \alpha^2 \]
\[ = \alpha(1 - \alpha). \]
Rules of Variance and Standard Deviation

Again consider the linear function $Y = aX + b$.

Now $\mu_Y = a\mu_X + b$, so

$$(Y - \mu_Y)^2 = a^2(X - \mu_X)^2$$

and so

$$V(Y) = V(aX + b) = a^2 V(X).$$

As a result,

$$\sigma_Y = |a|\sigma_X.$$