CM2104: Computational Mathematics
Discrete Probability Theory 2: Random Variables

Prof. David Marshall

School of Computer Science & Informatics

Random variables

Consider again (**see last lecture**) the following example:

Example

Consider a dice which has a probability of $\frac{3}{4}$ of being fair, and a probability of $\frac{1}{4}$ of being biased such that

$$p(1) = p(2) = p(3) = p(4) = p(5) = \frac{1}{10} \qquad p(6) = \frac{1}{2}$$

What is the **probability** of rolling an **even number**?

The **sample space** is given by:

$$S = \{(i, f) \,|\, 1 \leq i \leq 6, f \in \{fair, biased\}\}$$

In our previous solution, however, we used notations such as $P(\{2, 4, 6\}|\textbf{fair})$ instead of now we use:

$$P\left(\{(i, f) \,|\, i \in \{2, 4, 6\}, f \in \{\textbf{fair}, \textbf{biased}\}\} \,|\, \{(i, \textbf{fair}) \,|\, 1 \leq i \leq 6\}\right)$$

Random variables

In practice we often want to refer to particular aspects of the **sample space**:

- In the previous example, we may **only** be interested to know **whether or not the dice is fair**
- When rolling two dice, we may **only** be interested in the **sum of the numbers** obtained

A **random variable** X is a **mapping** from S to $\mathbb{R}$.

- For each **outcome** s, the value $X(s)$ **corresponds** to a **particular feature of** s

Note that a **random variable** X defines a **new sample space** S_X given by:

$$S_X = \{X(s) \,|\, s \in S\}$$

with the associated **probability distribution** p_X defined for $x \in \mathbb{R}$ as:

$$p_X(x) = P(\{s \,|\, X(s) = x\})$$

Example

Consider the experiment of **rolling two dice**. The **sample space** S is given by:

$$\{(i,j) \,|\, 1 \leq i,j \leq 6\}$$

where the associated **probability distribution** p is **uniform**, *i.e.* $p((i,j)) = \frac{1}{36}$ **for each** $(i,j) \in S$

Let X be the random variable corresponding to the **total number obtained**, *i.e.*

$$\begin{aligned} X : \quad & S \to \mathbb{R} \\ & (i,j) \mapsto i+j \end{aligned}$$

Then $S_X = \{2, ..., 12\}$ and the associated **probability distribution** p_X is given by:

$$p_X(2) = p_X(12) = \frac{1}{36} \qquad p_X(3) = p_X(11) = \frac{2}{36} \qquad p_X(4) = p_X(10) = \frac{3}{36}$$

$$p_X(5) = p_X(9) = \frac{4}{36} \qquad p_X(6) = p_X(8) = \frac{5}{36} \qquad p_X(7) = \frac{6}{36}$$

Note that p_X is **not uniform**

MATLAB Code, Throwing 2 6-sided dice probability Distribution

two_dice_px.m

```
samples=[1 2 3 4 5 6];

% Make Sample Space S = {(i,j)|1 ? i,j ? 6}
sample_space = repmat(samples,6,1) + repmat(samples',1,6);

% Count Unique elements
[unique_elements, unique_idx, element_idx] = unique(sample_space);
length(element_idx);

count = hist(element_idx,unique(element_idx));  % hist only deals with number so use char_idx
px = count/length(element_idx);

for i = 1:length(unique_elements);
    fprintf('px(%d) = %1.4f\n',unique_elements(i),px(i));
end
```

- `repmat(A,M,N)` — Replicates and tiles an array, returning a large matrix consisting of an `M`-by-`N` tiling of copies of `A`: see `doc/help repmat()`
- `unique(A)` — returns the same values as in `A` but **sorted** with **no repetitions**: see `doc/help unique()`
- `hist(Y,X)` — returns the count of `Y` among bins with centres specified by values `X`: see `doc/help hist()`

Note: (See Exercises) Brute force enumeration is not the best way to evaluate this solution for $m > 2$ dice.

Exercise

- Consider the experiment of doing **three coin flips**, and let X be the random variable corresponding to the number of times **heads** was **obtained**.

  What is the associated **probability distribution** p_X?

- Let Y be the random variable which is **1** if all **three** coin flips had the **same result** (i.e. either three times heads or three times tails) and 0 otherwise.

  What is the associated **probability distribution** p_Y?

Notation

An advantage: **Random variables** often allow us to **compactly** describe events

We write $X = a$ for the event $\{s \,|\, X(s) = a\}$, $X \geq a$ for the event $\{s \,|\, X(s) \geq a\}$, and similar for $X > a$, $X \neq a$, *etc.*

We write $X = 1, Y = 2$ for the event $\{s \,|\, X(s) = 1 \textbf{ and } Y(s) = 2\}$

The usual calculation rules for the **intersection of events** apply, *e.g.*

$$P(X = a | Y = b, C = z) = \frac{P(X = a, Y = b, C = z)}{P(Y = b, C = z)}$$

$$P(X = a, Y = b) = P(X = a) \cdot P(Y = b | X = a)$$

If $X = a$ and $Y = b$ are **independent events** for all $a \in S_X$ and $b \in S_Y$ we say that X and Y are **independent random variables**, and we have accordingly that

$$P(X = a, Y = b) = P(X = a) \cdot P(Y = b)$$

Expected value

Often it is desirable to **summarise** the **characteristics** of a random variable with a few scalar values, known as the **moments** of the distribution.

The **expected value** $E[X]$ of a random variable X is the moment defined by:

$$E[X] = \sum_{x \in S_X} x \cdot p_X(x) = \sum_{s \in S} X(s) \cdot p(s)$$

The **expected value** is also known as the **mean** of the distribution

If X and Y are **two random variables** and $a, b, c \in \mathbb{R}$, we write $aX + bY + c$ for the random variable defined by $(aX + bY + c)(s) = aX(s) + bY(s) + c$, and similar for other operations on the reals. We have:

$$E[aX + bY + c] = a \cdot E[X] + b \cdot E[Y] + c$$

Note however that in general

$$E[f(X_1, ..., X_n)] \neq f(E[X_1], ..., E[X_n])$$

Expected value

Example: Consider the experiment of **throwing a dice**, and let the **random variable** X correspond to the number shown by the dice, then the **expected value** $E[X]$ is:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}$$

Note that the **expected value** is intuitively an **expectation of the average value** over a **large number of experiments**, rather than really a value that you would **"expect"**

Example: Consider the experiment of **throwing two dice**, and let the**random variables** X_1 and X_2 correspond to the **numbers** shown by the **first** and the **second dice** respectively. The **expected value** of the **sum of the two** dice obtained is given by

$$E[X_1 + X_2] = E[X_1] + E[X_2] = \frac{7}{2} + \frac{7}{2} = 7$$

MATLAB: Simple Dice Throwing Expectation Example

expectation.m

```
p=[1/6 1/6 1/6 1/6 1/6 1/6];
ln=length(p);
samples=[1 2 3 4 5 6];

expec=sum(p.*samples);

disp(['For the outcome of throwing a dice as: ',
          num2str(samples)]);
disp(['with probabilities: ' num2str(p)]);
disp(['The expected value of throwing a dice is: ',
          num2str(expec)]);
disp(['The expected value of throwing two dice is: ',
          2*num2str(expec)]);
```

Expected value

Example

A game is played in which **three fair dice** are **thrown independently**.
You **win £1** if **two** of the **dice show** the **same number**, and **£4** if all three **show** the **same number**; **otherwise you lose £1**.
Let **X** denote the **amount** you **win**.
Find the **expected value** of **X**.

The distribution of X is given by:

$$p_X(4) = \frac{6}{6 \cdot 6 \cdot 6} = \frac{1}{36}$$
$$p_X(-1) = \frac{6 \cdot 5 \cdot 4}{6 \cdot 6 \cdot 6} = \frac{20}{36} = \frac{4}{9}$$
$$p_X(1) = 1 - \frac{1}{36} - \frac{20}{36} = \frac{15}{36} = \frac{5}{12}$$

So we find

$$E[X] = 4 \cdot \frac{1}{36} - 1 \cdot \frac{20}{36} + 1 \cdot \frac{15}{36} = -\frac{1}{36}$$

Hence, in the long run, you are expected to lose money by playing this game

Exercise

Consider a **multiple-choice test** with **5 questions.**

- For each of the questions **three** options are presented, exactly **one** of which is the **correct answer**.
- For each question which is **correctly** answered, **3 marks** are awarded, but for each question which is **incorrectly** answered **2 marks** are lost as a correction for random guessing.
- For questions which remain **unanswered 0 marks** are given (*i.e.* no marks are lost).

Suppose you answer **all the questions** by **randomly guessing** one of the options, and let the random variable X denote our resulting score.
Find the **expected value** $E[X]$.

Variance

The **variance** is a moment of a random variable which measures its **variability**, *i.e.* how much the outcome would typically differ from the expected value:

$$Var[X] = E[(X - E[X])^2]$$

Note that $E[X - E[X]]$ would be **meaningless** as a measure of variability, since

$$E[X - E[X]] = E[X] - E[X] = 0$$

The **standard deviation** σ_X of a **random variable** X is defined as:

$$\sigma_X = \sqrt{Var[X]}$$

Variance

Proposition

It holds that

$$Var[X] = E[X^2] - (E[X])^2$$

Indeed, noting that $E[X] \in \mathbb{R}$, we have

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2 \cdot E[X] \cdot X + E[X]^2] \\ &= E[X^2] - 2 \cdot E[X] \cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Proposition

For $X_1, ..., X_n$ independent random variables, it holds that

$$Var[X_1 + ... + X_n] = Var[X_1] + ... + Var[X_n]$$

Note that the above proposition only holds for **independent random variables**!

Variance

Example: Consider again the experiment of **throwing a dice** with X the **number obtained**. We have:

$$E[X] = \frac{7}{2} \qquad E[X^2] = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} + \frac{16}{6} + \frac{25}{6} + \frac{36}{6} = \frac{91}{6}$$

Hence

$$Var[X] = \frac{91}{6} - \frac{49}{4} = \frac{182 - 147}{12} = \frac{35}{12} \approx 2.92$$

Now consider the same experiment, but using a **biased dice** for which $p_X(1) = p_X(2) = p_X(5) = p_X(6) = \frac{1}{8}$ and $p_X(3) = p_X(4) = \frac{1}{4}$. Then

$$E[X] = \frac{1+2+5+6}{8} + \frac{3+4}{4} = \frac{7}{2}$$
$$E[X^2] = \frac{1+4+25+36}{8} + \frac{9+16}{4} = \frac{66+50}{8} = \frac{29}{2}$$

Hence

$$Var[X] = \frac{29}{2} - \frac{49}{4} = \frac{58-49}{4} = \frac{9}{4} \approx 2.25$$

The **variance** is **smaller** than above, as the probability distribution is **more concentrated around** the **expected value**

MATLAB: Simple Dice Throwing Expectation Example

variance.m

```
p=[1/6 1/6 1/6 1/6 1/6 1/6];
ln=length(p);
samples=[1 2 3 4 5 6];

expec=sum(p.*samples);
expecsq = sum(samples.*samples.*p);
vardice = expecsq - expec*expec;
% or sum(p.*power(samples-expec,2))

disp(['For the outcome of throwing a dice as: ',
         num2str(samples)]);
disp(['with probabilities: ' ,
         num2str(p)]);
disp(['The expected value of throwing a dice is: ',
         num2str(expec)]);
disp(['The Variance  of throwing a dice is: ',
         num2str(vardice)]);
```

Exercise

Consider again the **multiple-choice test** with **5 questions** and the **random variable** X denoting the score.

Find the **variance** $Var[X]$.

Uniform distribution

In practical applications, most **random variables** have a **probability distribution** which belongs to a well-known **class** of distributions.

One example is the **uniform distribution**, where $p(x) = \frac{1}{|S_X|}$ for every $x \in S_X$. Note that when $S_X = \{1, 2, ..., n\}$, we then have:

$$E[X] = \sum_{i=1}^{n} \frac{1}{n} \cdot i = \frac{1}{n} \cdot \sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n \cdot (n+1)}{2} = \frac{n+1}{2}$$

$$E[X^2] = \sum_{i=1}^{n} \frac{1}{n} \cdot i^2 = \frac{1}{n} \sum_{i=1}^{n} i^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

which gives us

$$\begin{aligned} Var[X] &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{2n^2+3n+1}{6} - \frac{n^2+2n+1}{4} \\ &= \frac{n^2-1}{12} \end{aligned}$$

MATLAB DISCRETE Uniform Distribution Examples

uniform.m

```
% First get the domain over which we will
% evaluate the functions.

n=6;  % Dice
% n=52;  % Any Card
% n= 4;  % Card Suit
% n= 13;  %  Card Rank

% Draw n Samples from distribution
x = 1:n;

% Now get the probability density function
% values at x.
pdf = unidpdf(x,n);
% Now get the cdf.
cdf = unidcdf(x,n);
% Do the plots.
subplot(1,2,1),plot(x,pdf)
title('Probability Density Function')
xlabel('X'),ylabel('f(X)')
axis([x(1)-1 x(end)+1 0 1/n+0.1])
axis square
subplot(1,2,2),plot(x,cdf)
title('Cumulative Distribution Function')
xlabel('X'),ylabel('F(X)')
axis([x(1)-1 x(end)+1 0 1.1])
axis square
shg;
```

```
%  Expectation and Variance

[E V] = unidstat(n);

fprintf('Expectation and Variance of Discrete
            Uniform Distribution of size
            %d, E(X) = %1.4f,
            Var(X) = %1.4f\n', n, E,V);
```

- `unidpdf(X,N)` — returns the **discrete** uniform probability density function on `(1,2,...,N)` at the values in `X`: see `doc/help unicpdf`
- `unidcdf(X,N)` — returns the **discrete** uniform cumulative density function on `(1,2,...,N)` at the values in `X`: see `doc/help unidcdf`
- `unidstat(N)` — returns the **mean** and **variance** of the (discrete) uniform distribution on `(1,2,...,N)`: see `doc/help unidstat`

Binomial Distribution

A **Bernoulli trial** is an experiment whose outcome is either **success** or **failure**.

For example, **flipping a coin**, considering **heads** to be a **success** and **tails** to be a **fail**.

Let the random variable X be the number of successes in n successive **Bernoulli trials**, each of which has a **probability** of q to be **successful**. Then the **distribution** of X is the **binomial distribution** with parameters n and q, defined by:

$$p(k) = {}^{n}C_k q^k (1-q)^{n-k} = \binom{n}{k} q^k (1-q)^{n-k}$$

We write $X \sim B(n, q)$ to indicate that X has this **distribution**.

Let $X_i = 1$ if the ith trial was **successful** and $X_i = 0$ **otherwise**. Then $X = X_1 + ... + X_n$ and

$$E[X] = E[X_1 + ... + X_n] = E[X_1] + ... + E[X_n] = n \cdot q$$

Furthermore, it can be shown that:

$$Var[X] = n \cdot q \cdot (1-q)$$

MATLAB: Simple Binomial Distribution Example

binomial.m

A Quality Assurance inspector tests **200** circuit boards a day. If **2%** of the boards have **defects**, what is the **probability** that the inspector will find **no defective boards** on any given day?

What is the **most likely number** of **defective** boards the inspector will find?

```
% what is the probability that the inspector will find no defective boards on any given day?

binopdf(0,200,0.02)

% What is the most likely number of defective boards the inspector will find?
defects=0:200;
y = binopdf(defects,200,.02);
[x,i]=max(y);
defects(i)
```

- `binopdf(X,N,P)` — returns the binomial probability density function with parameters `N` and `P` at the values in `X`: see `doc/help binopdf`.
- see also `doc/help binocdf`, `binoinv`, `binornd`, `binostat`.

Exercise

Banach always carries **two** matchboxes with him

- **one** in his **left pocket** and **one** in his **right pocket**.

Every time he needs a match,

- he picks **one** of the **two matchboxes** at **random** and takes a match from that box.

Initially, the matchboxes **contained 10 matches each**.

One day, Banach reaches into his **left pocket** and **discovers** that the matchbox is **empty**.

What is the **probability** that the matchbox in his other pocket has **exactly 4 matches left**?

Geometric distribution

Let the **random variable** X be the **number of times** we have to **repeat** a **Bernoulli trial** before we have our **first success**, with q still being the **probability** that an individual trial is **successful**
The sequence is like this:

$$
\begin{aligned}
P(X=1) &= q && \text{Success on 1st trail} \\
P(X=2) &= (1-q)q && \text{Failure on 1st, Success on 2nd trail} \\
P(X=3) &= (1-q)^2 q && \text{Failure on 1st+2nd, Success on 3rd trail} \\
P(X=4) &= (1-q)^3 q \\
&\vdots \\
P(X=n) &= (1-q)^{n-1} q
\end{aligned}
$$

The **distribution** of X is then the **geometric distribution** with parameter q, defined by

$$p_X(n) = (1-q)^{n-1} q$$

and we write $X \sim Geom(q)$.

Expected value of the geometric distribution.

Let $p = q - 1$

n	1	2	3	4	...
$P(X = n)$	q	pq	p^2q	p^3q	...

Now, let's use the fact that

$$\begin{aligned} E[X] &= \sum_{n=1}^{\infty} n \cdot P(X = n) \\ &= q + 2pq + 3p^2q + 4p^3q + \ldots \\ &= q(1 + 2p + 3p^2 + 4p^3 + \ldots) \\ &= q(1-p)^{-2} \end{aligned}$$

From the **generalised Binomial Theorem**, $\sum_{i=0}^{n} \frac{1}{(1-x)^s} = \binom{s+n-1}{n} x^n \equiv \binom{s+n-1}{s-1} x^n$

So with $s = 2, x = p$ we see that $(1-p)^{-2} = 1 + 2p + 3p^2 + 4p^3 + \ldots$ so

$$\begin{aligned} E[X] &= \frac{q}{(1-p)^2} \\ &= \frac{q}{q^2} \\ &= \frac{1}{q} \end{aligned}$$

Geometric distribution: Expected Value + Variance

So

$$E[X] = \frac{1}{q}$$

- So, for example, if the success probability q is $\frac{1}{3}$, it will take on **average 3 trials** to get a success.
- **All this maths for a result that was intuitively clear all along!**

Geometric distribution: Variance

It can be also shown that:

$$Var[X] = \frac{1-q}{q^2}$$

Sketch of Proof (**See Lab Class Exercises**):

$$E[X^2] = \sum_{n=1}^{\infty} n^2 \cdot P(X = n)$$
$$= q + 4pq + 9p^2q + 16p^3q + \ldots$$

Factor and simplify similar to E[X] above and eventually get:

$$E[X^2] = \frac{1}{q} + \frac{2p}{q^2}$$

$$Var[X] = E[X^2] - E^2[X] = \frac{1}{q} - \frac{2p}{q^2} - \frac{1}{q^2} = \ldots = \frac{1-q}{q^2}$$

MATLAB: Geometric Distribution Example

geometric.m

Suppose you toss a **fair** coin repeatedly, and a **success** occurs when the coin lands on heads.
What is the probability of observing **exactly three tails** (**failures**) before tossing a heads?

```
% To solve, determine the value of the probability density
%  function (pdf)
% for the geometric distributon at x equal to 3.
% The probability of success (tossing a heads) p in any given
%  trial is 0.5.

x = 3;
p = 0.5;
y = geopdf(x,p)
```

- `geopdf(X,P)` — returns the **pdf** of the geometric distribution with probability parameter `P`, evaluated at the values in `X`: see `doc/help geopdf`.
- see also `doc/help geocdf`, `geoinv`, `geornd`, `geostat`.

Example

Boys and Girls

In a country where everyone wants a boy, each family continues having babies till they have a boy. After some time, what is the proportion of boys to girls in the country? (Assuming probability of having a boy or a girl is the same)

Note: this is a well-known google interview question (http://www.mytechinterviews.com/10-google-interview-questions)

We consider **two random variables**:

B the **number of boys** in a given family
G the **number of girls** in a given family

We need to find the value of:

$$\frac{E[B]}{E[G]}$$

Example: Boys and Girls Solution

Clearly $p_B(1) = 1$ and $p_B(x) = 0$ for $x \neq 1$, hence

$$E[B] = \sum_{n \in \mathbb{N}} x \cdot p_B(n) = 1 \cdot p_B(1) = 1$$

To find the **expected number of girls** in a family, we consider that for $X = G + 1$ we have $X \sim Geom(\frac{1}{2})$, hence

$$E[G] = E[X - 1] = E[X] - 1 = 2 - 1 = 1$$

In other words, we can **expect** that the **number of boys** will be approximately the **same** as the **number of girls**

- although in practice, because families cannot really have an infinite number of children, we may expect a slightly higher number of boys.

MATLAB Code: Boys and Girls Example

geometric_boy.m

```
% We consider two random variables:
%
% B -- the number of boys in a given family
% G -- the  number of girls in a given family
%
% We need to compute E[B]/E[G]

% probability of having a boy or a girl is the same
pb = 0.5;
pg = 1 - pb;

Eb = geostat(pb);

% For a Girl, X = G + 1;
% E[G] = E[X-1] = E[X] - 1;
Eg = geostat(pg) + 1 - 1;

Ratio_Eb2Eg = Eb/Eg

% If now we say it is 60% more chance of having a girl
pb = 0.4; % probability of having a boy
pg = 1 - pb; % probability of having a girl

Eb = geostat(pb);

% For a Girl, X = G + 1;
% E[G] = E[X-1] = E[X] -
Eg = geostat(pg) + 1 - 1;

Ratio_Eb2Eg = Eb/Eg
```

Poisson distribution

Suppose we are interested in **counting** the **number of occurrences** of a certain event during a given period of time, where

- You know the **average rate** with which the event **occurs** (*e.g.* the average number of occurrences per hour)
- The occurrence of an event is **independent** from the time since the last occurrence

For example:

- The **number of cars** passing a certain intersection
- The **number of HTTP** requests received by a web server
- The **number of customers** arriving in a shop

Poisson distribution

Let the expected number of occurrences during a given time interval be λ, and let the **random variable** X be the **actual number** of **occurrences** during that time interval.

The **distribution of** X is then the **Poisson distribution** with parameter λ, defined by

$$p_X(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

and we write $X \sim Pois(\lambda)$

To understand the definition of the **Poisson distribution**, let us try to **approximate** the answer using the **binomial distribution**

Let us assume that there are n time points (with $n > k$), during the considered interval, at which an event can occur, and let $X \sim B(n, q)$. The expected number of occurrences is λ, so we have

$$E[X] = n \cdot q = \lambda \Leftrightarrow q = \frac{\lambda}{n}$$

Poisson distribution

The **probability** of **seeing k events** then becomes

$$p_X(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

If n is small, this will only be a very rough approximation of the correct answer, but it can be shown that

$$\lim_{n \to \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

In other words, the **Poisson distribution** could be seen as an **extreme** case of the **binomial distribution**, where the number of **Bernoulli trials** is **infinite**

It can be shown that for $X \sim Pois(\lambda)$ (**See Lab Exercises**):

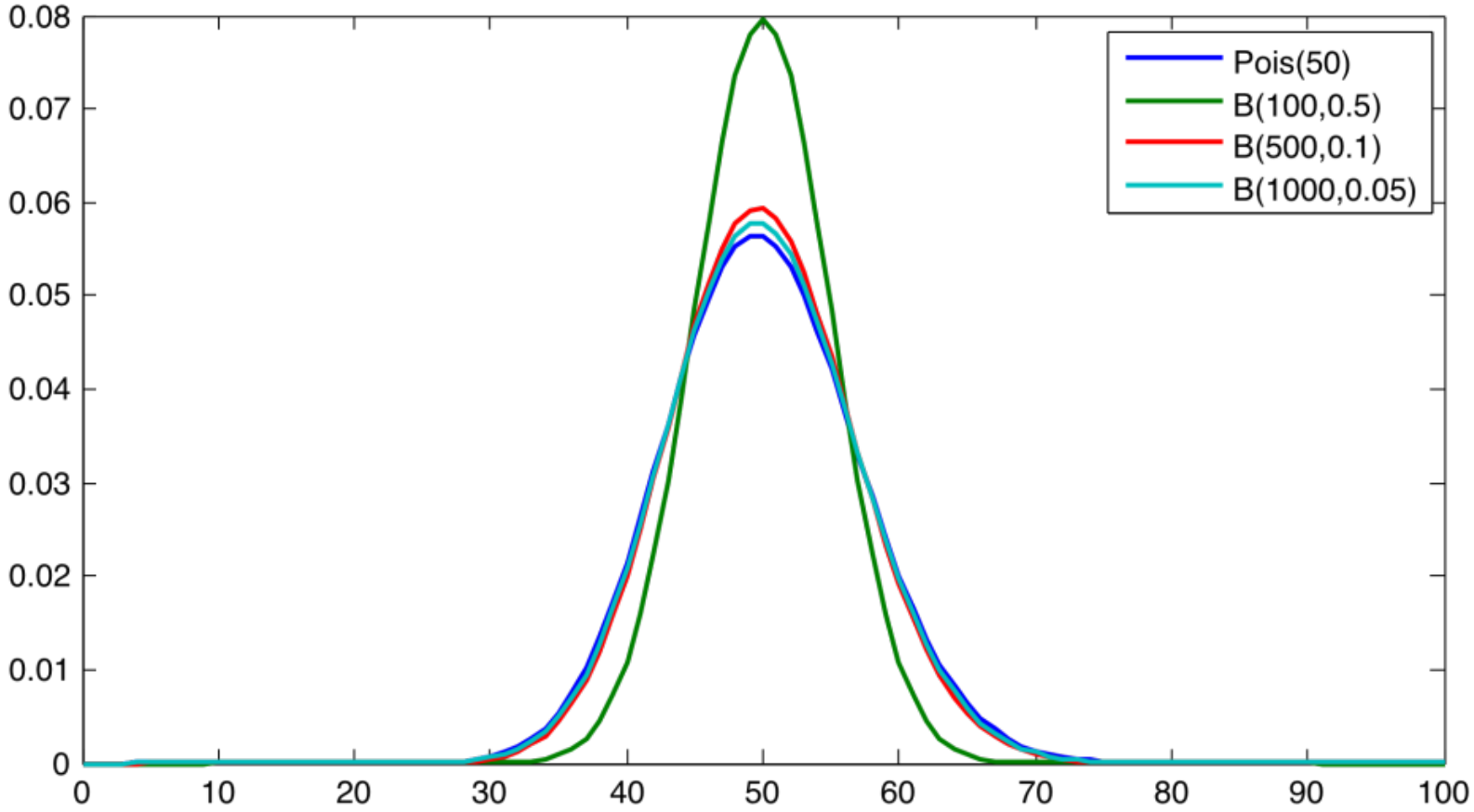
$$E[X] = \lambda \qquad\qquad Var[X] = \lambda$$

MATLAB Poisson distributions

poisson.m

```
% Example to show that an "infinite" number of Bernoulli tiral
% approximate a Poission distribution

x = 0:1:100;
y = poisspdf(x,50);
z100 = binopdf(x,100,0.5);
z500 = binopdf(x,500,0.1);
z1000 = binopdf(x,1000,0.05);
plot(x,y,x,z100,x,z500,x,z1000);
legend('Pois(50)', 'B(100,0.5)', 'B(500,0.1)', 'B(1000,0.05)');
```

Poisson distribution: Output from poisson.m



Poisson distribution

Car passing by

The **probability** of a **car passing** a certain intersection in a **20 minute window** is **0.9**.
What is the **probability** of a car **passing** the intersection in a **5 minute window**? (Assuming a constant probability throughout)

Let $X \sim Pois(\lambda)$ be the number of cars passing in a **20 minute window**. Then:

$$0.9 = P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$$

which means:

$$\lambda = -ln(0.1).$$

Now let Y be the number of cars passing in a **5 minute window**, then $Y \sim Pois(\frac{\lambda}{4})$, hence

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{\frac{ln(0.1)}{4}} \approx 0.438$$

MATLAB Code: Car passing by, Poisson Distribution

poisson_car.m

```
% The probability of a car passing a certain
% intersection in a 20 minute window is 0.9.
% What is the  probability of a car passing the
% intersection in a 5 minute window?
% (Assuming a constant probability throughout)

p20= 0.9;
lambda  = -log(1 - p20);

%  Prob of a car passing in a 5 minute window is:

p5 = 1 - poisspdf(0,lambda/4)
```

Poisson distribution

Web server load modelling

A web server is **receiving** on **average 100** requests per second, while it is currently able to handle up to 200 requests per second (if more requests come in during a second, the server fails).
What is the probability of the server failing during a particular second?

Let $X \sim Pois(100)$ be the number of requests that arrive during the considered second.
Then we need to find, $P(X > 200)$:

$$P(X > 200) = 1 - P(X \leq 200) = \sum_{i=0}^{200} p_X(i)$$

Matlab Code: poisson_webserver.m

```
x = 0:1:200;
y = poisspdf(x,100);
pfail= 1 - sum(y)
```

Joint distributions

Often we need **several random variables** to model a problem. Given a sequence of **random variables** $X_1, ..., X_n$ we can consider the **sample spaces** $S_{X_1}, ..., S_{X_n}$ and the **probability distributions** $p_{X_1}, ..., p_{X_n}$

However, we can also consider a **single sample space**:

$$S_{X_1,..,.X_n} = S_{X_1} \times ... \times S_{X_n} = \{(x_1, ..., x_n) \mid x_i \in S_{X_i}\}$$

The associated probability distribution is called the **joint probability distribution** of $X_1, ..., X_n$:

$$p_{X_1,...,X_n}(a_1, ..., a_n) = P(X_1 = a_1, ..., X_n = a_n)$$

Using the **joint probability distribution** $p_{X_1,...,X_n}$, we can calculate the **expected value** of $Y = f(X_1, ..., X_n)$ as:

$$E[Y] = \sum_{(a_1,...,a_n) \in S_{X_1,...,X_n}} f(a_1, ..., a_n) \cdot p_{X_1,...,X_n}(a_1, ..., a_n)$$

Joint distributions

Example: Consider the experiment of **throwing** a **fair dice**, and let the **random variables** X and Y be defined as ($s \in S$):

$$X(s) = \begin{cases} 1 & \text{if } s \text{ is odd} \\ 2 & \text{if } s \text{ is even} \end{cases} \qquad Y(s) = \begin{cases} 1 & \text{if } s \leq 3 \\ 2 & \text{if } s > 3 \end{cases}$$

Then we have

$$S_{X,Y} = \{(1,1),(1,2),(2,1),(2,2)\}$$

The **joint distribution** is given by:

$$p_{X,Y}(1,1) = P(\{1,3\}) = \frac{1}{3} \qquad p_{X,Y}(1,2) = P(\{5\}) = \frac{1}{6}$$

$$p_{X,Y}(2,1) = P(\{2\}) = \frac{1}{6} \qquad p_{X,Y}(2,2) = P(\{4,6\}) = \frac{1}{3}$$

Marginal distributions

If we **only** know the **joint distribution** $p_{X_1,\ldots,X_n}$ of the **random variables** $X_1, \ldots, X_n$ but **not** the **distributions** $p_{X_1}, \ldots, p_{X_n}$ we can derive them as

$$p_{X_i}(x_i) = \sum_{x_1 \in S_1} \cdots \sum_{x_{i-1} \in S_{i-1}} \sum_{x_{i-1} \in S_{i+1}} \cdots \sum_{x_n \in S_n} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

The **probability distributions** $p_{X_1}, \ldots, p_{X_n}$ are called the **marginal distributions of** $p_{X_1,\ldots,X_n}$.

Example: Consider again the joint distribution $p_{X,Y}$ defined by $p_{X,Y}(1,1) = p_{X,Y}(2,2) = \frac{1}{3}$ and $p_{X,Y}(1,2) = p_{X,Y}(2,1) = \frac{1}{6}$.

Then:

$$p_X(1) = p_{X,Y}(1,1) + p_{X,Y}(1,2) = \frac{1}{2} \quad p_X(2) = p_{X,Y}(2,1) + p_{X,Y}(2,2) = \frac{1}{2}$$

$$p_Y(1) = p_{X,Y}(1,1) + p_{X,Y}(2,1) = \frac{1}{2} \quad p_Y(2) = p_{X,Y}(1,2) + p_{X,Y}(2,2) = \frac{1}{2}$$

Marginal distributions

While we **can** recover the **marginal distributions** from the **joint distribution**, there is **not enough information** in the **marginal distributions** to **recover** the **joint distribution**.

If X and Y are **independent**, however, we do have that:

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

Covariance

The **covariance** of **two random variables X and Y** is defined as:

$$Cov[X, Y] = E[(X - E[X]) \cdot (Y - E[Y])]$$

Note that for any **random variable X**:

$$Cov[X, X] = Var[X]$$

We have

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X]) \cdot (Y - E[Y])] \\ &= E[X \cdot Y - X \cdot E[Y] - Y \cdot E[X] + E[X] \cdot E[Y]] \\ &= E[X \cdot Y] - E[X] \cdot E[Y] - E[Y] \cdot E[X] + E[X] \cdot E[Y] \\ &= E[X \cdot Y] - E[X] \cdot E[Y] \end{aligned}$$

Correlation

The **correlation coefficient** σ_{XY} of **two random variables** X and Y is defined as:

$$\sigma_{XY} = \frac{Cov[X, Y]}{\sigma_X \cdot \sigma_Y} = \frac{Cov[X, Y]}{\sqrt{Var[X] \cdot Var[Y]}}$$

If $\sigma_{XY} = 0$, the X and Y are called **uncorrelated**.

If X and Y are **independent**, they satisfy $E[X \cdot Y] = E[X] \cdot E[Y]$ (**see Lab Class Exercise**).

Therefore, **independent random variables** are **always uncorrelated**:

$$Cov[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y] = E[X] \cdot E[Y] - E[X] \cdot E[Y] = 0$$

However, not all **uncorrelated variables** are **independent**!

Note that $\sigma_{XY} = 0$ iff

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Correlation

It can easily be shown that for **two random variables** X and Y, we have

$$Var[X+Y] = Var[X] + 2 \cdot Cov[X,Y] + Var[Y]$$

(**see Lab Class Exercise**)

In particular, if X and Y are uncorrelated, we have

$$Var[X+Y] = Var[X] + Var[Y]$$

Note that $\sigma_{XY} \in [-1, 1]$.

If $\sigma_{XY} > 0$ (resp. $\sigma_{XY} < 0$), **higher** values of X typically co-occur with **higher** (resp. lower) values of Y.

Correlation

Example

Find the **covariance** and **correlation coefficient** of the **random variables** X and Y, whose **joint probability function** is given by:

		y		
		0	1	2
x	0	$\frac{1}{24}$	$\frac{3}{24}$	$\frac{2}{24}$
	1	$\frac{2}{24}$	$\frac{4}{24}$	$\frac{6}{24}$
	2	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{4}{24}$

The **marginal distributions** of X and Y are given by

$$p_X(0) = \frac{6}{24} \qquad p_X(1) = \frac{12}{24} \qquad p_X(2) = \frac{6}{24}$$

$$p_Y(0) = \frac{4}{24} \qquad p_Y(1) = \frac{8}{24} \qquad p_Y(2) = \frac{12}{24}$$

Correlation

To **compute** $Cov[X, Y]$, we need $E[X \cdot Y]$, $E[X]$ and $E[Y]$:

$$E[X] = (0 \cdot \frac{6}{24}) + (1 \cdot \frac{12}{24}) + (2 \cdot \frac{6}{24}) = 1$$
$$E[Y] = (0 \cdot \frac{4}{24}) + (1 \cdot \frac{8}{24}) + (2 \cdot \frac{12}{24}) = \frac{32}{24}$$
$$E[X \cdot Y] = (0 \cdot \frac{9}{24}) + (1 \cdot \frac{4}{24}) + (2 \cdot \frac{7}{24}) + (4 \cdot \frac{4}{24}) = \frac{34}{24}$$

which means

$$Cov[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y] = \frac{34}{24} - \frac{32}{24} = \frac{1}{12}$$

To compute σ_{XY} we also need $Var[X]$ and $Var[Y]$:

$$E[X^2] = (0 \cdot \frac{6}{24}) + (1 \cdot \frac{12}{24}) + (4 \cdot \frac{6}{24}) = \frac{36}{24}$$
$$E[Y^2] = (0 \cdot \frac{4}{24}) + (1 \cdot \frac{8}{24}) + (4 \cdot \frac{12}{24}) = \frac{56}{24}$$

Correlation

We obtain

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{2}$$
$$Var[Y] = E[Y^2] - E[Y]^2 = \frac{5}{9}$$

Hence

$$\sqrt{Var[X]} = 0.707 \qquad \sqrt{Var[Y]} = 0.745$$

so the **correlation coefficient** is given by

$$\rho_{XY} = \frac{Cov[X, Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}} = \frac{0.083}{0.707 \cdot 0.745} = 0.158$$

MATLAB Simple Covariance Example(1)

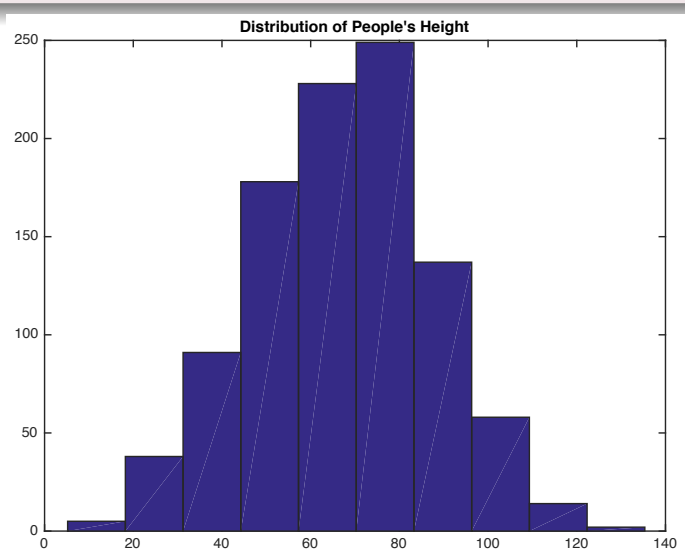
covariance.m

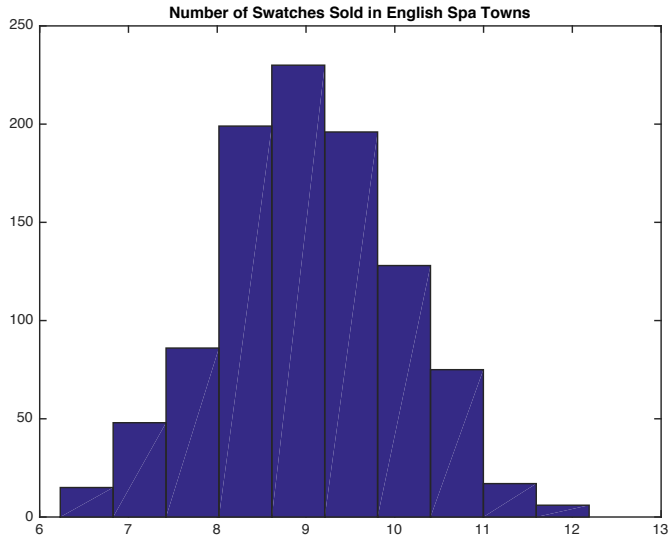
```
n = 1000;% Generate n samples

% Height of n people + Plot
x_height = normrnd(67,20,[1 n]);
figure; plot(hist(x_height));
title('Distribution of People''s Height')

% Number of Swatches sold in English Spa Towns
y_swatches = normrnd(9,1,[1 n]);
figure;plot(hist(y_swatches));
title('Number of Swatches sold in English Spa
       Towns')

% No correlation between x_height + y_swatches
cov1 = cov(x_height, y_swatches)
```

- Generate `n` samples from a normal distributions. See `doc/man normrnd`
- Use `cov(X,Y)` to compute covariance between the two random variables. See `doc/man cov`
- There is **no expected correlation** between
  - Person's height
  - Number of Swatches sold per day in English spa towns

$$cov1 \approx \begin{bmatrix} 400 & 0 \\ 0 & 1 \end{bmatrix}$$

See `doc/man hist` for displaying/counting statistics of samples.





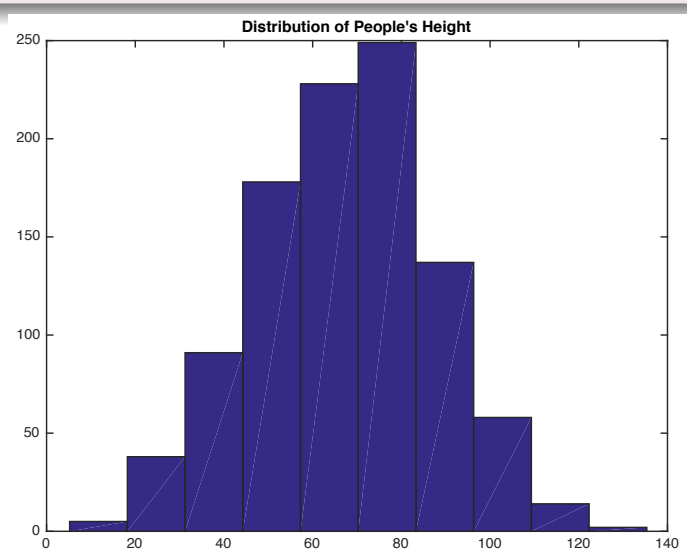
MATLAB Simple Covariance Example (2)

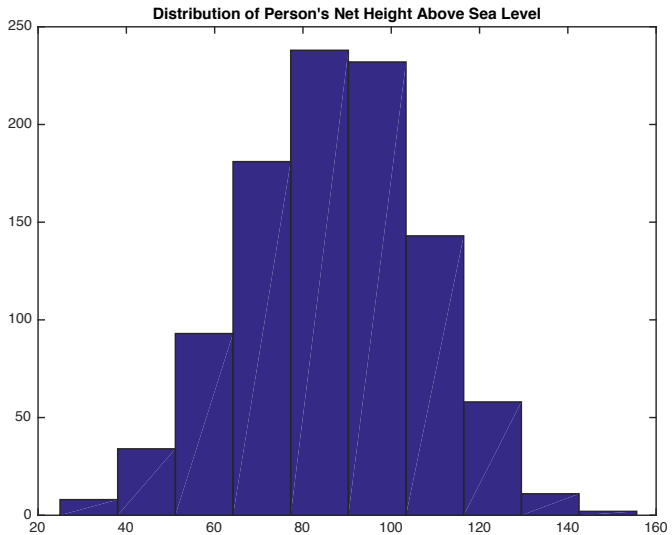
covariance.m

```
% Elevations above sea level
y_elevation = normrnd(20,1,[1 n]);

net_height = x_height +  y_elevation;

figure;
hist(net_height);
title('Distribution of Person''s Net Height
       Above Sea Level')

% Correlation between person's height and net
%  height above sea level

cov2 = cov(x_height, net_height)
```

- There **is expected correlation** between
  - Person's height
  - Net measured person's height above sea level (height + elevation)

$$cov2 \approx \begin{bmatrix} 400 & 400 \\ 400 & 400 \end{bmatrix}$$

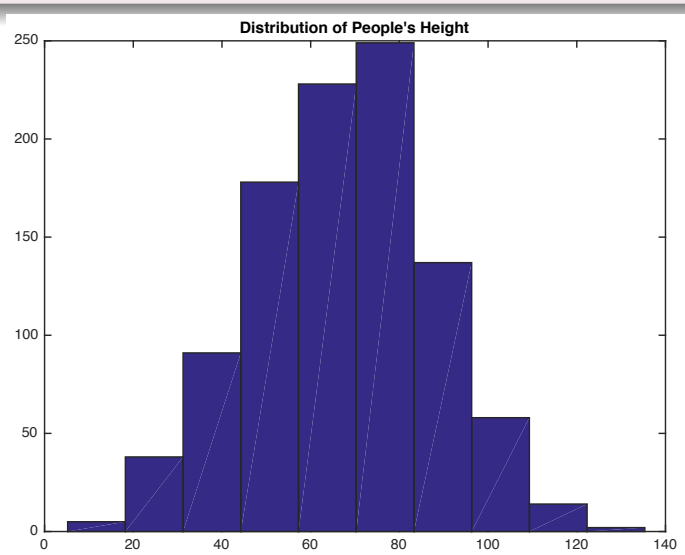




MATLAB Simple Covariance Example(3)

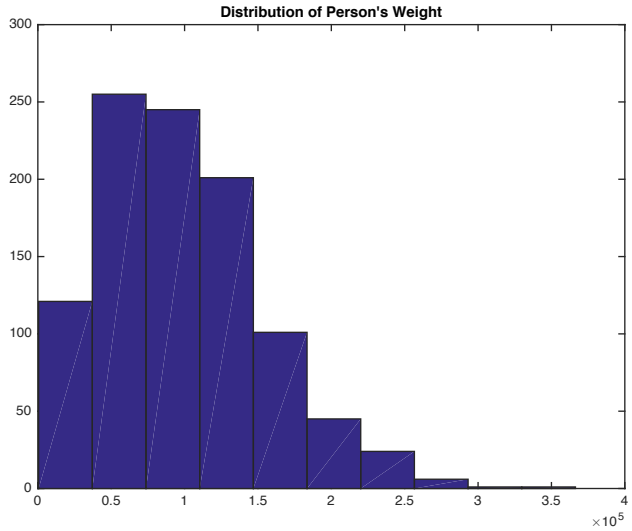
covariance.m

```
% Correlation between person's height and weight

Av_BMI = 20; % Body Mass Index

y_weight = 20*power(x_height,2);

figure;
hist(y_weight);
title('Distribution of Person''s Weight')

data3  = [x_height ; y_weight]';

cov3 = cov(x_height, y_weight)
```



- There **is expected correlation** between
  - Person's height
  - Person's weight
    - Assume normal BMI: $weight \approx 20.height^2$

$$cov3 \approx \begin{bmatrix} 0 & 11x10^6 \\ 11x10^6 & 3x10^9 \end{bmatrix}$$



Multinomial distribution

Consider a sequence of n **experiments**, each of which can yield **one** of the **outcomes** $s_1, \ldots, s_k$.

Let X_i be the **number of times outcome** s_i was **obtained**, then the **joint probability distribution** of $X_1, \ldots, X_k$ is called the **multinomial distribution**. It is defined as:

$$p_{X_1,\ldots,X_k}(X_1 = n_1, \ldots, X_k = n_k) = \begin{cases} \frac{n!}{n_1! \cdot \ldots \cdot n_k!} \cdot p_1^{n_1} \cdot \ldots \cdot p_k^{n_k} & \text{if } n = n_1 + \ldots + n_k \\ 0 & \text{otherwise} \end{cases}$$

where p_i is the **probability** of s_i in **each** of the **individual experiments**
It can be shown that ($i \neq j$):

$$E[X_i] = n \cdot p_i$$
$$Var[X_i] = n \cdot p_i \cdot (1 - p_i)$$
$$Cov[X_i, X_j] = -n \cdot p_i \cdot p_j$$

Note that for **higher** values of X_i, we would indeed expect **lower** values of X_j.
In the **special case** that $k = 2$, the **multinomial distribution** degenerates to the **binomial distribution** (considering that then $X_2 = n - X_1$)

Multinomial distribution

Example: When **throwing 6 dice**, the **probability** of getting **2 sixes, a five, and 3 fours** is given by:

$$p_{X_1,X_2,X_3}(2,1,3) = \frac{6!}{2! \cdot 1! \cdot 3!} \left(\frac{1}{6}\right)^3 \cdot \frac{1}{6} \cdot \left(\frac{1}{6}\right)^2 = \frac{10}{6^5}$$

multinomial.m

```
n = 6;
p = ones(1,n)/n; % Equal prob for all numbers;

x = [0,0,0,3,1,2];

% Compute the pdf of the distribution.
Y = mnpdf(x,p)
```

See `doc/help mnpdf`, `mnrfit`, `mnrval`, `mnrnd`

We'll see more of the multinomial distribution in the **next Section**.