Recap of Basic Probability Theory

Uffe Høgsbro Thygesen

Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby – Denmark
Email: uht@imm.dtu.dk

Elements of basic probability theory

- Stochastic experiments
- The probability triple $(\Omega, \mathcal{F}, \mathbb{P})$:
  - Ω: The sample space, $\omega \in \Omega$
  - $\mathcal{F}$: The set of events, $A \in \mathcal{F} \Rightarrow A \subset \Omega$
  - $\mathbb{P}$: The probability measure, $A \in \mathcal{F} \Rightarrow \mathbb{P}(A) \in [0, 1]$
- Random variables
- Distribution functions
- Conditioning

Why recap probability theory?

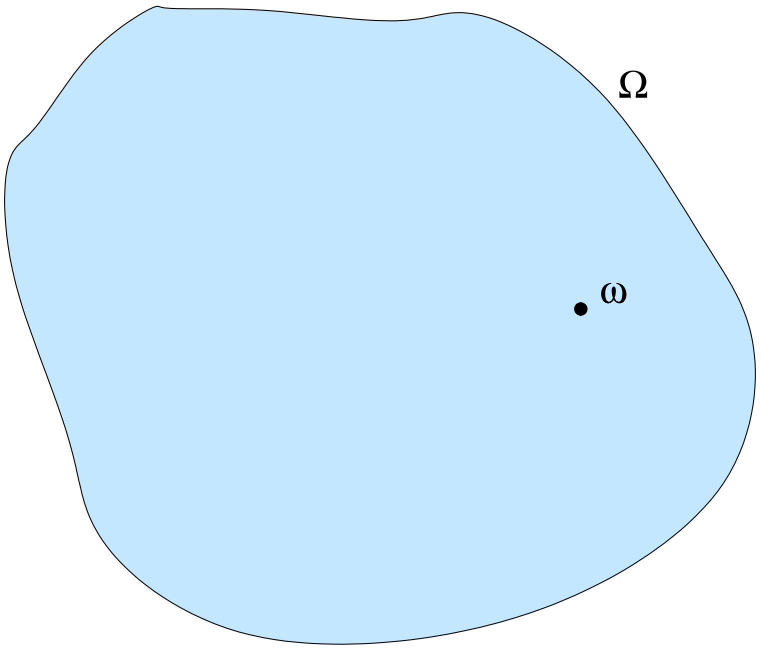
- Stochastic processes is *applied probability*
- A firm understanding of probability (as taught in e.g. 02405) will get you far
- We need a more solid basis than most students develop in e.g. 02405.

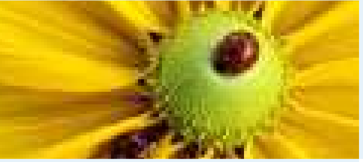
What to recap?

The concepts are most important: What is a stochastic variable, what is conditioning, etc.
Specific models and formulas: That a binomial distribution appears as the sum of Bernoulli variates, etc.

The set-up of probability theory

We perform a *stochastic experiment.*
We use ω to denote the outcome.
The **sample space** Ω is the set of all possible outcomes.





The sample space Ω

Ω can be a very simple set, e.g.

- $\{H, T\}$ (tossing a coin a.k.a. Bernouilli experiment)
- $\{1, 2, 3, 4, 5, 6\}$ (throwing a die once).
- $\mathbf{N}$ (typical for single discrete stochastic variables)
- $\mathbf{R}^d$ (typical for multivariate continuous stochastic variables)

or a more complicated set, e.g.

- The set of all functions $\mathbf{R} \mapsto \mathbf{R}^d$ with some regularity properties.

Often we will not need to specify what Ω is.

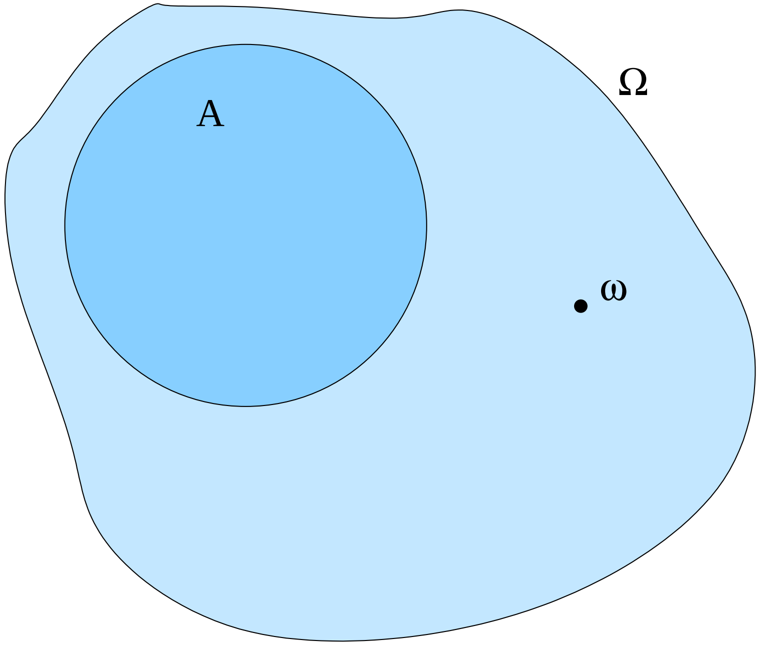
Events are sets of outcomes/subsets of Ω

Events correspond to statements about the outcome.

For a die thrown once, the event

$$A = \{1, 2, 3\}$$

corresponds to the statement "the die showed no more than three".



Probability

A Probability is a set measure of an event

If A is an event, then

$$\mathbb{P}(A)$$

is the probability that the event A occurs in the stochastic experiment - a number between 0 and 1.

(What exactly does this **mean**? C.f. G&S p 5, and appendix III)

Regardless of interpretation, we can pose simle conditions for **mathematical consistency**.

Logical operators as set operators

An important question: Which events are "measurable", i.e. have a probability assigned to them?
We want our usual logical reasoning to work!
So: If A and B are legal statements, represented by measurable subsets of Ω, then so are

- Not A, i.e. $A^c = \Omega \backslash A$
- A or B, i.e. $A \cup B$.

Parallels between statements and sets

Set	Statement
A	"The event A occured" $(\omega \in A)$
A^c	Not A
$A \cap B$	A and B
$A \cup B$	A or B
$(A \cup B) \setminus (A \cap B)$	A exclusive-or B

See also table 1.1 in Grimmett & Stirzaker, page 3

An infinite, but countable, number of statements

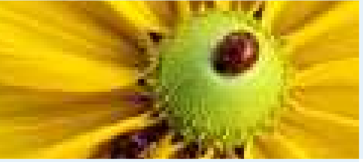
For the Bernoulli experiment, we need statements like

> At least one experiment shows heads

or

> In the long run, every other experiment shows heads.

So: If A_i are events for $i \in \mathbf{N}$, then so is $\cup_{i \in \mathbf{N}} A_i$.



All events considered form a σ-field $\mathcal{F}$

Definition:

1. The empty set is an event, $\emptyset \in \mathcal{F}$
2. Given a countable set of events $A_1, A_2, \ldots$, its union is also an event, $\cup_{i \in \mathbf{N}} A_i \in \mathcal{F}$
3. If A is an event, then so is the complementary set A^c.

(Trivial) examples of σ-fields

1. $\mathcal{F} = \{\emptyset, \Omega\}$
   This is the deterministic case: All statements are either true $(\forall\omega)$ or false $(\forall\omega)$.
2. $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$
   This corresponds to the Bernoulli experiment or tossing a coin: The event A corresponds to "heads".
3. $\mathcal{F} = 2^\Omega =$ set of all subsets of Ω.

When Ω is finite or enumerable, we can actually work with 2^Ω; otherwise not.

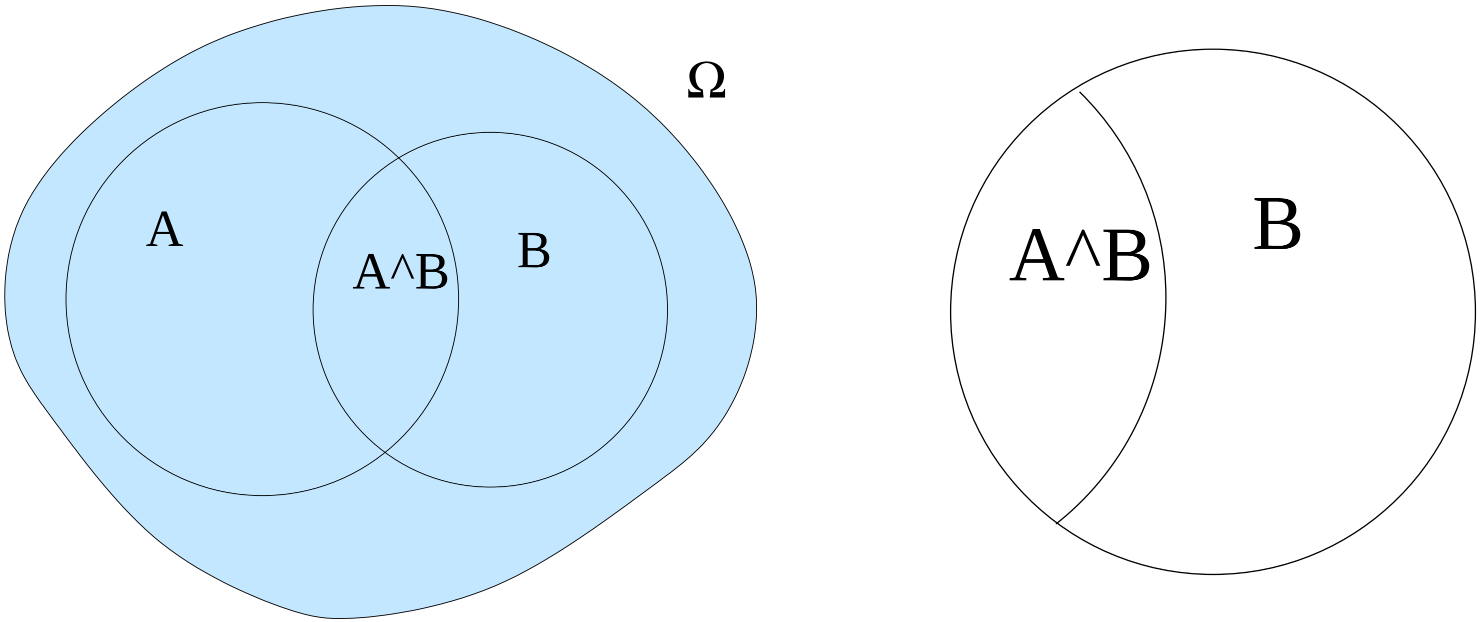
1. $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$
2. If $A_1, A_2, \ldots$ are mutually excluding events (ie. $A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

A $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ satisfying these is called a probability measure.
The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space.**

Conditional probabilities

In stochastic processes, we want to know what to expect from the future, conditional on our past observations.



$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Be careful when conditioning!

If you are not careful about specifying the events involved, you can easily obtain wrong conclusions.

Example: A family has two children. Each child is a boy with probability $1/2$, independently of the other. Given that at least one is a boy, what is the probability that they are both boys?

Be careful when conditioning!

If you are not careful about specifying the events involved, you can easily obtain wrong conclusions.

Example: A family has two children. Each child is a boy with probability $1/2$, independently of the other. Given that at least one is a boy, what is the probability that they are both boys?

The meaningless answer:

$$\mathbb{P}(\text{two boys}|\text{at least one boy}) = \mathbb{P}(\text{other child is a boy}) = \frac{1}{2}$$

Be careful when conditioning!

If you are not careful about specifying the events involved, you can easily obtain wrong conclusions.

Example: A family has two children. Each child is a boy with probability $1/2$, independently of the other. Given that at least one is a boy, what is the probability that they are both boys?

The meaningless answer:

$$\mathbb{P}(\text{two boys}|\text{at least one boy}) = \mathbb{P}(\text{other child is a boy}) = \frac{1}{2}$$

The right answer:

$$\mathbb{P}(\text{two boys}|\text{at least one boy}) = \frac{\mathbb{P}(\text{two boys})}{\mathbb{P}(\text{at least one boy})} = \frac{1/4}{3/4} = \frac{1}{3}$$

Lemma: The law of total probability

Let $B_1, \ldots, B_n$ be a *partition* of Ω

(ie., mutually disjoint and $\cup_{i=1}^n B_i = \Omega$)

Then

$$P(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Independence

Events A and B are called independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

When $0 < \mathbb{P}(B) < 1$, this is the same as

$$\mathbb{P}(A|B) = \mathbb{P}(A) = \mathbb{P}(A|B^c)$$

A family $\{A_i : i \in I\}$ of events is called independent if

$$\mathbb{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i)$$

for any finite subset J of I.

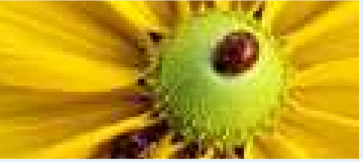
Stochastic variables

Informally: A quantity which is assigned by a stochastic experiment.
Formally: A mapping $X : \Omega \mapsto \mathbf{R}$.

Informally: A quantity which is assigned by a stochastic experiment.
Formally: A mapping $X : \Omega \mapsto \mathbf{R}$.

A Technical comment We want the probabilities $\mathbb{P}(X \leq x)$ to be well defined. So we require

$$\forall x \in \mathbf{R} \ : \ \{\omega : X(\omega) \leq x\} \in \mathcal{F}$$



Examples of stochastic variables

Indicator functions:

$$X(\omega) = I_A(\omega) = \begin{cases} 1 \text{ when } \omega \in A, \\ 0 \text{ else.} \end{cases}$$

Bernoulli variables:

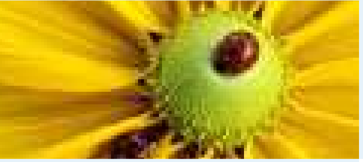
$$\Omega = \{H, T\}, \qquad X(H) = 1, \qquad X(T) = 0.$$

F_X, the cumulated distribution function (cdf)

$$F(x) = \mathbb{P}(X \leq x)$$

Properties:

1. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$.
2. $x < y \Rightarrow F(x) \leq F(y)$
3. F is right-continuous, ie. $F(x+h) \to F(x)$ as $h \downarrow 0$.



Discrete and continuous variables

Discrete variables: $\mathbf{Im}X$ is a countable set. So F_X is a step function. Typically $\mathbf{Im}X \subset \mathbf{Z}$.
Continuous variables: X has a *probability density function* (pdf) f, i.e.

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

so F is differentiable.

A variable which is neither continuous nor discrete

Let $X \sim U(0,1)$, i.e. uniform on $[0,1]$ so that

$$F_X(x) = x \qquad \text{for } 0 \leq x \leq 1.$$

Toss a fair coin. If heads, then set $Y = X$. If tails, then set $Y = 0$.

$$F_Y(y) = \frac{1}{2} + \frac{1}{2}x \qquad \text{for } 0 \leq y \leq 1.$$

We say that Y has an *atom* at 0.

Mean and variance

The mean of a stochastic variable is

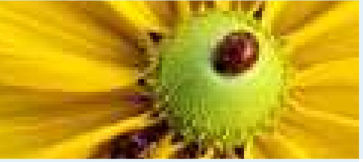
$$\mathbb{E}X = \sum_{i \in \mathbf{Z}} i\mathbb{P}(X = i)$$

in the discrete case, and

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

in the continuous case. In both cases we assume that the sum/integral exists absolutely.
The variance of X is

$$\mathbb{V}X = \mathbb{E}(X - \mathbb{E}x)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$



Conditional expectation

The conditional expectation is the mean in the conditional distribution

$$\mathbb{E}(Y|X=x) = \sum_y y f_{Y|X}(y|x)$$

It can be seen as a stochastic variable: Let $\psi(x) = \mathbb{E}(Y|X=x)$, then $\psi(X)$ is the conditional expectation of Y given X

$$\psi(X) = \mathbb{E}(Y|X)$$

We have

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$$

Conditional variance $\mathbb{V}(Y|X)$

is the variance in the conditional distribution.

$$\mathbb{V}(Y|X = x) = \sum_y (y - \psi(x))^2 f_{Y|X}(y|x)$$

This can also be written as

$$\mathbb{V}(Y|X) = \mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2$$

and can be manipulated into (try it!)

$$\mathbb{V}Y = \mathbb{E}\mathbb{V}(Y|X) + \mathbb{V}\mathbb{E}(Y|X)$$

which **partitions** the variance of Y.

Random vectors

When a single stochastic experiment defines the value of several stochastic variables.
Example: Throw a dart. Record both vertical and horizontal distance to center.

$$X = (X_1, \ldots, X_n) : \Omega \mapsto \mathbf{R}^n$$

Also random vectors are characterised by the distribution function $F : \mathbf{R}^n \mapsto [0, 1]$:

$$F(x) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

where $x = (x_1, \ldots, x_n)$.

Example

In one experiment, we toss two fair coins and assign the results to V and X.

In another experiment, we toss *one* fair coin and assign the result to *both* Y and Z.

V, X, Y and Z are all identically distributed.

But (V, X) and (Y, Z) are *not* identically distributed.

E.g. $\mathbb{P}(V = X = \text{heads}) = 1/4$ while $\mathbb{P}(Y = Z = \text{heads}) = 1/2$.

The Bernoulli process

Start with working out one single **Bernoulli experiment**.

Then consider a finite number of Bernoulli experiments: **The binomial distribution**

Next, a sequence of Bernoulli experiments: **The Bernoulli process.**

Waiting times in the Bernoulli process: **The negative binomial distribution**.

The Bernoulli experiment

A Bernoulli experiment models e.g. tossing a coin.

The sample space is $\Omega = \{H, T\}$.

Events are $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\} = 2^\Omega = \{0,1\}^\Omega$.

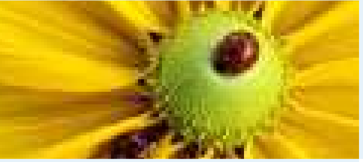
The probability $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ is defined by (!)

$$\mathbb{P}(\{H\}) = p \qquad .$$

The stochastic variable $X : \Omega \mapsto \mathbf{R}$ with

$$X(H) = 1 \qquad X(T) = 0$$

is *Bernoulli distributed* with parameter p.



A finite number of Bernoulli variables

We toss a coin n times.

The sample space is $\Omega = \{H, T\}^n$.

For the case $n = 2$, this is $\{TT, TH, HT, HH\}$.

Events are $\mathcal{F} = 2^\Omega = \{0,1\}^\Omega$.

How many events are there? $|\mathcal{F}| = 2^{|\Omega|} = 2^{(2^n)}$ (a lot).

Introduce A_i for the event "the i'th toss showed heads".

The probability $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ is defined by

$$\mathbb{P}(A_i) = p$$

and by requiring that

the events $\{A_i : i = 1, \ldots, n\}$ are independent.

From this we derive

$$\mathbb{P}(\{\omega\}) = p^k (1-p)^{n-k} \text{ if } \omega \text{ has } k \text{ heads and } n-k \text{ tails.}$$

and from that the probability of *any* event.

Define the stochastic variable X as number of heads

$$X = \sum_{i=1}^{n} \mathbf{1}(A_i)$$

To find its **probability mass function**, consider the events

$$\mathbb{P}(X = x) \qquad \text{which is shorthand for } \mathbb{P}(\{\omega : X(\omega) = x\})$$

This event $\{X = x\}$ has $\binom{n}{x}$ elements. Each $\omega \in \{X = x\}$ has probability

$$\mathbb{P}(\omega) = p^x (1-p)^{n-x}$$

so the probability is

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Properties of $B(n,p)$

Probability mass function $$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Cumulated distribution function $$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i=0}^{x} f_X(i)$$

Mean value $$\mathbb{E}X = \mathbb{E}\sum_{i=1}^{n} \mathbf{1}(A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i) = np$$

Variance $$\mathbb{V}X = \sum_{i=1}^{n} \mathbb{V}\mathbf{1}(A_i) = np(1-p)$$

because $\{A_i : i = 1, \ldots n\}$ are (pairwise) independent.

Problem 3.11.8

Let $X \sim B(n, p)$ and $Y \sim B(m, p)$ be independent.
Show that $Z = X + Y \sim B(n + m, p)$

Problem 3.11.8

Let $X \sim B(n,p)$ and $Y \sim B(m,p)$ be independent.
Show that $Z = X + Y \sim B(n+m,p)$

Solution:

Consider $m+n$ independent Bernoulli trials, each w.p. p.

Set $X = \sum_{i=1}^{n} \mathbf{1}(A_i)$ and $Y = \sum_{i=n+1}^{n+m} \mathbf{1}(A_i)$.

Then X and Y are as in the problem, and

$$Z = \sum_{i=1}^{n+m} \mathbf{1}(A_i) \sim B(n,p)$$

The Bernoulli process

A sequence of Bernoulli experiments.

The sample space Ω is the set of functions $\mathbf{N} \mapsto \{0, 1\}$.

Introduce events A_i for "the ith toss showed heads".

Strictly: $A_i = \{\omega : \omega(i) = 1\}$

Let $\mathcal{F}$ be the smallest σ-field that contains all A_i.

Probabilities in the Bernoulli process

Define (!) $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ by

$$\mathbb{P}(A_i) = p$$

and

$$\{A_i : i \in \mathbf{N}\} \text{ are independent.}$$

Probabilities in the Bernoulli process

Define (!) $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ by

$$\mathbb{P}(A_i) = p$$

and

$$\{A_i : i \in \mathbf{N}\} \text{ are independent.}$$

Waiting times in the Bernoulli process

Let W_r be the waiting time for the rth succes:

$$W_t = \min\{i : \sum_{j=1}^{i} \mathbf{1}(A_j) = r\}$$

To find the probability mass function of W_r, note that $W_r = k$ is the same event as

$$(\sum_{i=1}^{k-1} \mathbf{1}(A_i) = r-1) \cap A_k$$

Since the two events involved here are independent, we get

$$f_W(k) = \mathbb{P}(W_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

The geometric distribution

The waiting time W to the *first* success

$$\mathbb{P}(W = k) = (1 - p)^{k-1} p$$

(First $k - 1$ failures and then one success)

The *survival function* is

$$G_W(k) = \mathbb{P}(W > k) = (1 - p)^k$$

Summary

We need be precise in our use of probability theory, at least until we have developed intuition.

When in doubt, ask: What is the stochastic experiment? What is the probability triple? Which event am I considering?

Venn diagrams a very useful. This holds particularly for conditioning, which is central to stochastic processes.

Indicator functions are powerful tools, once mastered.

You need to know the distributions that can be derived from the Bernoulli process: The binomial, geometric, and negative binomial distribution.