Introduction to Probability Theory

Fabio G. Cozman - fgcozman@usp.br

September 11, 2019

Part I: Basic Concepts for Finite Spaces

1. Possibility/sample space, outcomes, events
2. Variables and indicator functions
3. Probabilities, expectations
4. Properties of probabilities

Possibility/sample space

1. Possibility/sample space: set Ω.
2. Elements ω of Ω are *outcomes*.
3. Subsets of Ω are *events* (no fuzziness!).

Example

Two coins are tossed; each coin can be heads (H) or tail (T). Then $\Omega = \{HH, HT, TH, TT\}$. Consider three events. Event $A = \{HH\}$ is the event that both tosses produce heads. Event $B = \{HH, TT\}$ is the event that both tosses produce identical outcomes. Event $C = \{HH, TH\}$ is the event that the second toss yields heads. Note that $A = B \cap C$.

Probability measure (finite spaces!)

A *probability measure* is a function that assigns a probability value to each event.

PU1 For any event A, $\mathbb{P}(A) \geq 0$.

PU2 The space Ω has probability one: $\mathbb{P}(\Omega) = 1$.

PU3 If events A and B are disjoint (that is, $A \cap B = \emptyset$), then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Easy example

Example

A six-sided die is rolled. Suppose all outcomes of Ω are assigned precise and identical probability values.

- We must have $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = \mathbb{P}(\Omega) = 1$, thus we have $\mathbb{P}(\omega) = 1/6$ for all outcomes.
- The event $A = \{1, 3, 5\}$ (outcome is odd) has probability $\mathbb{P}(A) = 1/2$.
- The event $B = \{1, 2, 3, 5\}$ (outcome is prime) has probability $\mathbb{P}(B) = 2/3$ and $\mathbb{P}(A \cap B) = \mathbb{P}(\{1, 3, 5\}) = 1/2$.

Properties of probabilities

1. As A and A^c are disjoint and $A \cup A^c = \Omega$, we have $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$,
$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)\,.$$
2. $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\emptyset^c) = 1 - \mathbb{P}(\Omega) = 0$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
4. For n mutually disjoint events B_i,
$$\mathbb{P}(\cup_{i=1}^n B_i) = \sum_{i=1}^n \mathbb{P}(B_i)\,.$$
5. If events $\{B_i\}$ form a partition of Ω,
$$\mathbb{P}(A) = \mathbb{P}(\cup_{i=1}^n A \cap B_i) = \sum_i \mathbb{P}(A \cap B_i)\,.$$

Random variables

1. Function $X : \Omega \to \Re$ is usually called a *random variable*.
2. If X is a variable, then any function $f : \Re \to \Re$ defines a random variable $f(X)$.

Example

The age in months of a person ω selected from a population Ω is a variable X. The same population can be used to define a different variable Y where $Y(\omega)$ is the weight (rounded to the next kilogram) of a person ω selected from Ω. We can also have a random variable $Z = X + Y$.

Distributions

- Possibility space Ω, variable $X : \Omega \to \Re$.
- Then: possibility space Ω_X containing every possible value of X.
- Probability measure on Ω induces a measure over subsets of Ω_X:

$$\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

- Induced measure on Ω_X is usually called the *distribution* of X.

Expectations

- Given variable X, its *expectation* is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_{x} x\mathbb{P}(X = x)\,.$$

- An expectation functional yields a real number for each variable.
- Properties:
  - For constants α and β, if $\alpha \leq X \leq \beta$, then $\alpha \leq \mathbb{E}[X] \leq \beta$.
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- Variance is $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

Part II: A Bit of History

1. Old times.
2. Classical probabilities.
3. Frequentist and Bayesian schemes.

Brief Look: History of Probabilities

- **Classical**
  - Leibnitz, Fermat, Pascal, De Moivre (1600)
  - Bayes (1700)
  - Laplace (1800)
  - Modifications: Keynes, Jeffreys, Jaynes (1940)
- **Frequentist**
  - Venn, Boole, De Morgan (1850)
  - Fisher, Neyman/Pearson (1900)
- **Bayesian**
  - Ramsey, De Finetti (1930)
  - Savage (1950)

The Classical Theory: Ancient Time

- First thoughts appeared in Philosophy:
  - Aristotle: "the probable is that which for the most part happens"
  - Bishop Butler: "to us probability is the very guide of life"
- Also, many philosophers have used probability to prove the existence of God (e.g., the proof of the ecliptic)

The Classical Theory: Evolution

- Pascal, De Moivre, Bernoulli: Central limit theorem, law of large numbers.
- Thomas Bayes: What you believe depends on what you believed before; we need prior distributions.

  Bayes' rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$.

The Classical Theory: Laplace

- Probability is the ratio of the number of favorable cases to that of all the cases possible
- The Principle of Non-Sufficient Reason: two possible cases are equally probable if there is no reason to prefer one to the other

The Classical Theory: Difficulties

The great problem: the Principle of Non-Sufficient Reason

- Too many proofs from too little knowledge.
- The problem of reparameterizations:
  If you are not sure about x, you are not sure about x^2. How to express that?

Now Come the Frequentists

- Basic Idea: instead of using ignorance, let's use knowledge
- Let's define probability as the limiting relative frequency of observations

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

- Venn, Boole and De Morgan proposed it around 1850; Statistics was built upon this conception of probabilities

The Frequentist Theory: Difficulties

- The definition is too poor compared to what we want
  - It is impossible to talk about probabilities for things that will happen only once!
- More mathematically, how to use the limit in the definition ($lim_{n\to\infty} n_A/n$)?
  - Many deterministic sequences have limits
  - Do random sequences have limits?

A Brief Summary So Far

- Classical Theory:
  - Probability is is the ratio of favorable cases to the number of cases (Principle of Non-Sufficient Reason)
  - Problem: Principle of Non-Sufficient Reason is untenable
- Frequentist Theory:
  - Probability is a limiting relative frequency
  - Problems: too narrow a concept; hard to define mathematically

The Emergence of Subjectivism

- Since everything else seems to fail, why don't we admit that there is a component of subjectivism in probability?
- Ramsey/De Finetti groundbreaking idea: let's define probability as a "fair" betting strategy:
  - I'll give you 1 unit of currency if President X is re-elected
  - How much would you pay to bet "fairly" that X will not be re-elected?
  - The amount you pay is your probability for *X re-elected*

The Bayesian Theory: Savage's Idea

- Axiomatize preferences over "gambles"
- From preferences, obtain "money" (utility) and probabilities
- Result: If $\mathbf{f} \preceq \mathbf{g}$ then there is a probability measure P and a utility function U such that $E[U(\mathbf{f})] < E[U[(\mathbf{g})]]$.

The Bayesian Theory: Basics

- All forms of uncertainty are reduced to probability
- Judgements of uncertainty are reduced to preferences
- All forms of updating knowledge are equivalent to application of Bayes' rule

Frequentists Versus Bayesians

Bayesians:

- Induction is a solved problem: you define your prior, you collect data and then you apply Bayes rule, always following decision theory
- Challenges: basically subjective (annoying priors).

Frequentists:

- Induction is an *ad hoc* activity; Statistics furnishes useful tools for induction
- Some tools: significance testing, hypothesis testing, least-squares...
- Challenges: based on shaky foundations; piecemeal and ad hoc approach to problems.

Part III

1. Moments, variance, covariance.
2. Weak laws of large numbers.

Moments

Definition

The ith *moment* of X is the expectation $\mathbb{E}\left[X^i\right]$.

Definition

The ith *central moment* of X is the expectation $\mathbb{E}\left[(X - \mathbb{E}[X])^i\right]$.

Definition

The *variance* $V[X]$ of X is second central moment of X.

Note:

$$V[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Markov inequality

1. Suppose $X \geq 0$ and $t > 0$.

   If $X(\omega) < t$, $I_{\{X \geq t\}}(\omega) = 0$. Then $X(\omega)/t \geq I_{\{X \geq t\}}(\omega)$. If $X(\omega) \geq t$, then $X(\omega)/t \geq 1 = I_{\{X \geq t\}}(\omega)$. Consequently, $X/t \geq I_{\{X \geq t\}}$ and then $\mathbb{E}[X]/t \geq \mathbb{E}\left[I_{\{X \geq t\}}\right]$, so:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

2. *Chebyshev inequality*:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{V[X]}{t^2}.$$

Digression: Covariance

Definition

The *covariance* of variables X and Y is $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

If two variables X and Y are such that $\mathrm{Cov}(X, Y) = 0$, then X and Y are *uncorrelated*.

Very weak law of large numbers

Theorem

If variables $X_1, X_2, \ldots, X_n$ have expectations $\mathbb{E}[X_i] \in [\underline{\mu}, \overline{\mu}]$ and variances $V[X_i] \in [\underline{\sigma}^2, \overline{\sigma}^2]$, and X_i and X_j are uncorrelated for every $i \neq j$, then for any $\epsilon > 0$,

$$\underline{\mathbb{P}}\left(\underline{\mu} - \epsilon < \frac{\sum_i X_i}{n} < \overline{\mu} + \epsilon\right) \geq 1 - \frac{\overline{\sigma}^2}{n\epsilon^2}.$$

Weak law of large numbers

Theorem

If variables $X_1, X_2, \ldots$ *have expectations* $\mathbb{E}[X_i] = \mu$ *and variances* $V[X_i] = \sigma^2$, *and* X_i *and* X_j *are uncorrelated for every* $i \neq j$, *then for any* $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{\sum_i X_i}{n} - \mu \right| < \epsilon \right) = 1.$$

Philosophy behing the "law"

- Idea: irregularities observed in X_i do not affect the average of these variables.
- We should have regularity out of apparent chaos: even though the random variables behave randomly, their average does approach some meaningful number (the probability...).
  - Suggests the "definition": $\mathbb{P}(A) = \lim_{n\to\infty} \#A/n$.

Part IV: Conditioning

1. Bayes rule.
2. Theorem of total probabilities.

Conditioning: Bayes rule

Definition

If $\mathbb{P}(B) > 0$, then

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

Definition

The conditional expectation of X given B, denoted by $\mathbb{E}[X|B]$, is defined only if $\mathbb{P}(B) > 0$ as

$$\mathbb{E}[X|B] = \sum_x x\mathbb{P}(X = x|B)\,.$$

Basic facts

For any C such that $\mathbb{P}(C) > 0$:

- For any A, $\mathbb{P}(A|C) \geq 0$.
- $\mathbb{P}(\Omega|C) = 1$.
- If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B|C) = \mathbb{P}(A|C) + \mathbb{P}(B|C)$.

Note that $\mathbb{P}(A|A) = 1$ whenever $\mathbb{P}(A) > 0$.

Properties

1. $\mathbb{E}[X|B] = \sum_{\omega \in B} X(\omega)\mathbb{P}(\omega|B)$.
2. For events $\{B_i\}_{i=1}^n$,

$$\mathbb{P}(B_1 \cap B_2 \cap \cdots \cap B_n) = \mathbb{P}(B_1) \prod_{i=2}^{n} \mathbb{P}\left(B_i | \cap_{j=1}^{i-1} B_j\right).$$

More properties

1. *Total probability theorem*: If events $\{B_i\}$ form a partition of Ω such that all $\mathbb{P}(B_i) > 0$,

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A|B_i)\,\mathbb{P}(B_i)\,.$$

2. Then:

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\,\mathbb{P}(B_i)}{\sum_i \mathbb{P}(A|B_i)\,\mathbb{P}(B_i)}.$$

Example

1. Individuals in an office have a disease D. Test to detect the disease (R or R^c).
2. $\mathbb{P}(R|D) = 9/10$ (sensitivity of the test).
3. $\mathbb{P}(R^c|D^c) = 4/5$ (specificity of the test).
4. $\mathbb{P}(D) = 1/9$.
5. Then:

$$\mathbb{P}(D|R) = \frac{9/10 \times 1/9}{9/10 \times 1/9 + 1/5 \times 8/9} = 9/25.$$

Three-prisoners problem

1. Three prisoners, Teddy, Jay, and Mark, waiting to be executed.
2. Governor will select one to be freed (equal probability)
3. Warden knows the governor's decision
4. Teddy convinces the warden to say the name of one of his fellow inmates who will be executed (useless information...)
5. Warden is honest
6. Warden says that Jay is to be executed: Teddy is happy (1/3 to 1/2)!
7. But if warden said Mark, Teddy would be happy??

Analysis

1. Possibility space:

$$\Omega = \left\{ \begin{array}{c} \text{Teddy freed} \cap \text{warden says Jay}, \\ \text{Teddy freed} \cap \text{warden says Mark}, \\ \text{Jay freed} \cap \text{warden says Mark}, \\ \text{Mark freed} \cap \text{warden says Jay} \end{array} \right\}.$$

2. We know that

$$\mathbb{P}(\text{Teddy freed}) = \mathbb{P}(\text{Jay freed}) = \mathbb{P}(\text{Mark freed}) = 1/3.$$

3. How would the warden behave if Teddy is to be freed?

$$\mathbb{P}(\text{warden says Jay} | \text{Teddy freed}).$$

Possible conclusion...

If

$$\mathbb{P}(\text{warden says Jay}|\text{Teddy freed}) = 1/2,$$

then:

$$\mathbb{P}(\text{Teddy freed} \cap \text{warden says Jay}) = 1/6,$$

$$\mathbb{P}(\text{Teddy freed} \cap \text{warden says Mark}) = 1/6,$$

$$\mathbb{P}(\text{Jay freed} \cap \text{warden says Mark}) = 1/3,$$

$$\mathbb{P}(\text{Mark freed} \cap \text{warden says Jay}) = 1/3.$$

Hence

$$\mathbb{P}(\text{Teddy freed}|\text{warden names Jay}) = \frac{1/6}{1/3 + 1/6} = 1/3.$$

Complete analysis...

1. Statement does not say anything about the behaviour of the warden.
2. All that is really known is

$$\mathbb{P}(\text{warden names Jay}|\text{Teddy freed}) \in [0, 1].$$

3. Consequently

$$\mathbb{P}(\text{Teddy freed}|\text{warden names Jay}) \in \left[\frac{0}{0+1/3}, \frac{1/3}{1/3+1/3}\right].$$

Part V: Probability mass functions

1. Mass functions.
2. Marginal probability mass functions.
3. Conditional probability mass functions.
4. Multivariate models.

Probability mass function

1. Probability mass function is simply $p_X(x) = \mathbb{P}(\{X = x\})$.
2. Then $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$.

Example (Uniform distribution)

Uniform distribution for X assigns $p_X(x) = 1/k$ for every value x of X.

Example (Bernoulli distribution)

Binary variable X with values 0 and 1. *Bernoulli distribution with parameter p* for X takes two values: $p_X(0) = (1 - p)$ and $p_X(1) = p$.
$\mathbb{E}[X] = 0(1 - p) + 1p = p$; $V[X] = p(1 - p)$.

Functions...

- For $Y = f(X)$,

$$p_Y(y) = \mathbb{P}(\{Y = y\}) = \sum_{x \in \Omega_X, Y(x) = y} p_X(x),$$

$$p_Y(y) = \mathbb{P}(\{Y = y\}) = \sum_{\omega \in \Omega, Y(X(\omega)) = y} \mathbb{P}(\omega)\,.$$

Conditional/joint mass functions

- *Conditional probability mass function* $p_{X|B}(x|B) = \mathbb{P}(\{X = x\}|B)$.
- *Joint probability mass function* $p(X, Y)$:

$$p_{X,Y}(x, y) = \mathbb{P}(\{X = x\} \cap \{Y = y\}) .$$

Marginal probability mass functions

$$\begin{aligned} p_X(x) &= \mathbb{P}(\{X = x\}) \\ &= \sum_{y \in \Omega_Y} \mathbb{P}(\{X = x\} \cap \{Y = y\}) \\ &= \sum_{y \in \Omega_Y} p_{X,Y}(x, y). \end{aligned}$$

Example

X and Y with three values each and

$p_{X,Y}(x, y)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	$1/10$	$1/25$	$1/20$
$x = 2$	$1/20$	$1/5$	$1/25$
$x = 3$	$1/10$	$1/50$	$2/5$

Expectations and mass functions

Finite spaces:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) \\
&= \sum_{x\in\Omega_X} \sum_{\omega:X(\omega)=x} x\mathbb{P}(\omega) \\
&= \sum_{x\in\Omega_X} x \sum_{\omega:X(\omega)=x} \mathbb{P}(\omega) \\
&= \sum_{x\in\Omega_X} x\mathbb{P}(X=x) \\
&= \sum_{x\in\Omega_X} x p_X(x).
\end{aligned}$$

Conditional probability mass

- For variable X and event A such that $\mathbb{P}(A) > 0$,

$$p_{X|A}(x|A) = \mathbb{P}(X = x|A)\,.$$

- For variables X and Y,

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|\,Y = y)\,.$$

Iterated expectations

- Denote $\mathbb{E}[X|Y=y]$ by $\mathbb{E}[X|y]$.
- Then $\mathbb{E}[X|Y]$ is a function of Y.
- For finite spaces:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x\in\Omega_X} x p_X(x) \\
&= \sum_{x\in\Omega_X} x \sum_{y\in\Omega_Y} p_{X,Y}(x,y) \\
&= \sum_{x\in\Omega_X} \sum_{y\in\Omega_Y} x p_{X|Y}(x|y) p_Y(y) \\
&= \sum_{x\in\Omega_X} \mathbb{E}[X|Y=y]\, p_Y(y) \\
&= \mathbb{E}[\mathbb{E}[X|Y]]\,.
\end{aligned}$$

For sets of variables

1. Probability mass function, conditional, joint, marginal, etc.
2. Vectors:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = [X_1, \ldots, X_n]^T,$$

$$p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\{\mathbf{X} = \mathbf{x}\}),$$

$$\mathbb{E}[\mathbf{X}] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^T,$$

Part VI: Independence

1. Independence for two events, for many events.
2. Independence for random variables.
3. Conditional independence.

Independence for events

1. A and B are independent

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \text{whenever } \mathbb{P}(B) > 0;$$

or, equivalently,

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

2. Many events are independent: for all subsets of events $\{A_i\}_{i=1}^n$,

$$\mathbb{P}(\cap_i A_i) = \prod_i \mathbb{P}(A_i)\,.$$

(Pairwise independence is not enough!)

Independence for random variables

- For all events such that the conditional probabilities are defined,

$$\mathbb{P}(\{X_i = x_i\} | \cap_{j \neq i} \{X_j = x_i\}) = \mathbb{P}(\{X_i = x_i\});$$

that is,

$$p(x_i | \cap_{j \neq i} \{X_j = x_i\}) = p(x_i).$$

- Or, more concisely:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i).$$

Conditional independence

1. $(X \perp\!\!\!\perp Y \,|\, A)$ if

$$\mathbb{E}[f(X)g(Y)|A] = \mathbb{E}[f(X)|A]\,\mathbb{E}[g(Y)|A]$$

for all functions f, g, whenever $\mathbb{P}(A) > 0$.

2. $(X \perp\!\!\!\perp Y \,|\, Z)$ if

$$(X \perp\!\!\!\perp Y \,|\, \{Z = z\})$$

for every category z of Z such that $\mathbb{P}(\{Z = z\}) > 0$.

Part VII: Laws of Large Numbers

1. Weak law.
2. Strong law.

Weak law of large numbers again

1. Independence implies uncorrelation:

$$\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] =$$

$$\mathbb{E}[X_i - \mathbb{E}[X_i]]\,\mathbb{E}[X_j - \mathbb{E}[X_j]] = 0.$$

2. If independent variables $X_1, X_2, \ldots$ have expectations $\mathbb{E}[X_i] = \mu$ and variances $V[X_i] = \sigma^2$, then for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\sum_i X_i}{n} - \mu\right| < \epsilon\right) = 1.$$

3. There are variants: assuming no variance, assuming expectations change, etc.

Advanced: strong law of large numbers

In a sequence of variables $X_1, \ldots, X_n$, the mean converges to the expectation with probability one:

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mu \right) = 1.$$

1. It requires the theory of infinite spaces.
2. It is hard to prove and requires several assumptions.
3. It is really a *strong* result.

Part VIII: General Spaces

1. Infinities.
2. General axioms.
3. Cumulative distribution functions and densities.
4. A summary.

Infinite spaces

- So, far Ω has been a finite set.
  1. Random variables have finitely many values.
  2. A probability mass function specifies a distributions through finitely many values.
- Now suppose Ω is an infinite set: Ω may be
  1. *countable* (natural, odd, integer, rational numbers) or
  2. *uncountable* (real numbers).

Kolmogorov's axioms

P1 For any event A, $\mathbb{P}(A) \geq 0$.

P2 The space Ω has probability one: $\mathbb{P}(\Omega) = 1$.

P3 If countably many events $\{A_i\}_{i=1}^{\infty}$ are disjoint, then $\mathbb{P}(\cup_i A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

- The last axiom introduces *countable additivity*.

Example with discrete variable

- Suppose X has integer values $0, 1, 2, \ldots$.
- Then X has a Poisson distribution with parameter $\lambda > 0$ when

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

for $x \geq 0$.

Summarizing a number of points:

A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

- Ω is a set (possibility space).
- $\mathcal{F}$ is a σ-algebra on Ω.
- $\mathbb{P}$ is a probability measure on $\mathcal{F}$; that is, a non-negative normalized (to unity) and countable additive set-function.

Note: in *almost all* books on probability theory, the probability space takes the real numbers and their Borel algebra.

Extending the previous theory

- A variable with finitely many values is called *simple*. We know how to relate expectation and probability for those.
- In general take:

$$\mathbb{E}[X] = \sup\left(\mathbb{E}[Y] \quad : \quad Y \leq X, Y \text{ is simple.}\right).$$

The Lebesgue integral

- So we have

$$\mathbb{E}[X] = \sup\left(\mathbb{E}[Y] \quad : \quad Y \leq X, Y \text{ is simple.}\right).$$

The Lebesgue integral

- So we have

$$\mathbb{E}[X] = \sup\left(\mathbb{E}[Y] \quad : \quad Y \leq X, Y \text{ is simple.}\right).$$

- This quantity is the Lebesgue integral with respect to the probability measure $\mathbb{P}$.
- Notation:

$$\mathbb{E}[X] = \int X d\mathbb{P}.$$

The Riemann integral

- Under quite general conditions, the Lebesgue integral can be computed using the Riemann integral (the "usual" integral).

- The key idea is to define *densities* and then to integrate with respect to densities.

Cumulative distribution function

- The function

$$F_X(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\})$$

  is the *cumulative distribution function* of X.
- Note:
  - F_X is a non-negative non-decreasing function, with $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
  - $\mathbb{P}([a, b]) = F_X(b) - F_X(a) = \int_a^b p(x)\, dx$.

Densities

- For a measurable variable X, the *density* of X is, when it exists:

$$p_X(x) = \frac{dF_X(x)}{dx} = \left.\frac{d\mathbb{P}(\{\omega : X(\omega) \leq \tau\})}{d\tau}\right|_x.$$

- Then:

$$\mathbb{E}[X] = \int_{\Omega_X} x p_X(x)\, dx,$$

where Ω_X is the set of values of X, and the integral is the Riemann integral.

Summary

1. Kolmogorov's theory: probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a general possibility space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a non-negative, normalized to unity and countably additive set-function (a normalized measure).
   - The most common σ-algebra for the real numbers is the Borel algebra (intervals).
   - Random variables are $\mathcal{F}$-measurable functions.
   - Expectations of measurable functions are Lebesgue integrals.
2. The distribution of X is entirely captured by $F_X(x)$, the *cumulative distribution function*.
3. If F_X is continuous, the variable X is *continuous*, and we can differentiate $F_X(x)$ to obtain the *density* $p_X(x)$.
4. If the distribution of X has a density $p_X(x)$, then expectation $\mathbb{E}[X]$ is a Riemann integral $\int x p_X(x)\, dx$.

Part IX: Catalog of Distributions

1. Common densities.
2. De Moivre - Laplace's theorem.

Uniform distribution

- Suppose X is a real-valued variable.
- The distribution of X is uniform if its density is

$$p_X(x) = \frac{1}{b-a} \text{ if } x \in [a, b].$$

and $p_X(x) = 0$ otherwise.

Gaussian distribution

- X has a Gaussian distribution when

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- $\mathbb{E}[X] = \mu$ and $V[X] = \sigma^2$.

De Moivre - Laplace's theorem

- Take n, $p \in [0, 1]$, such that $n \times p \times (1 - p) >> 1$; then

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\exp\left(-(k - np)^2 / 2np(1-p)\right)}{\sqrt{2\pi np(1-p)}}$$

  (That is, the ratio of two sides goes to 1.)
- That is, the probability that k among n trials are positive, when n grows without bound, can be approximated by a Gaussian density.

Gamma distribution

- X has a gamma distribution with parameters α and β when

$$p_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

for $x > 0$, and $p_X(x) = 0$ otherwise.

- Gamma function:

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz.$$

Note: For any positive integer k, $\Gamma(k) = (k-1)!$.

- We have $\mathbb{E}[X] = \alpha/\beta$ and the variance of X is α/β^2.

Gamma and exponential distributions

- Important: If $\{X_i\}_{i=1}^n$ are independent and have a Gamma distribution with parameters α_i and β, then $X = X_1 + \cdots + X_n$ has a Gamma distribution with parameter $\alpha_1 + \cdots + \alpha_n$ and β.
- If $\alpha = 1$ and $\beta > 0$, then X has an *exponential* distribution with parameter β,

$$p_X(x) = \beta e^{-\beta x},$$

for $x > 0$, and $p_X(x) = 0$ otherwise.

Chi-square distribution

- X has chi-square distribution when

$$p_X(x) = \frac{1}{\sqrt{2}\Gamma(1/2)} \frac{\exp(-x/2)}{\sqrt{x}}$$

  when $x > 0$, and $p_X(x) = 0$ otherwise.
- The χ^2 with n degrees of freedom:

$$p_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp(-x/2)$$

  when $x > 0$, and $p_X(x) = 0$ otherwise.
  (Gamma distribution, $\alpha = n/2$ and $\beta = 1/2$).
- If $\{X_i\}_{i=1}^n$ are Gaussian variables with $\mu = 0$ and $\sigma^2 = 1$, then $X_1^2 + \cdots + X_n^2$ has a χ^2 distribution with n degrees of freedom.

Beta distribution

- Often used to model random variables that are limited to an interval.
- X has a beta distribution with parameters α and β when

$$p_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

  for $x \in [0,1]$ and 0 otherwise.
- Note: a beta distribution is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$. If $\alpha = \beta = 1$, then we obtain the uniform distribution.
- For a beta distribution $p_X(\cdot)$ with parameters α and β, the expected value of X is $\alpha/(\alpha+\beta)$ and the variance is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

Dirichlet distribution

- A column vector of dimension n has a Dirichlet distribution when:

$$p_{X_1,\ldots,X_2}(x_1,\ldots,x_n) = \frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}x_i^{\alpha_i-1},$$

  when $\sum_i x_i = 1$, and 0 otherwise.
- Distribution is defined in a simplex of dimension $n-1$.
- The values $\{\alpha_i\}_{i=1}^{n}$, where $\alpha_i > 0$, are the parameters of the Dirichlet distribution.
- This is a direct generalization of the beta distribution.

t distribution

- X has a t distribution with n degrees of freedom (for $n > 0$) when

$$p_X(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}(1 + x^2/n)^{-(n+1)/2}.$$

- When $n = 1$, the distribution is called the *Cauchy* distribution — it is a distribution with undefined expected value and variance!
- Important: if X has a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$, and Y has a χ^2 distribution with n degrees of freedom, then $X/\sqrt{Y/n}$ has a t distribution with n degrees of freedom.

Part X

1. Multivariate densities.

Multivariate densities

- For two variables X and Y,

$$F_{X,Y}(x,y) = \mathbb{P}(\{X \leq x, Y \leq y\})$$

and

$$p_{X,Y}(x,y) = \frac{\partial F_{X,Y}(x,y)}{\partial X \partial Y}.$$

Then

$$\mathbb{P}((X,Y) \in D) = \int\int_D p_{X,Y}(x,y)\, dxdy.$$

- ...and similarly for any number of variables.

Gaussian vector

- A column vector $\mathbf{X}$ of dimension n has a Gaussian *joint* probability density when:

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det P}} \exp\left(-\frac{(\mathbf{x}-\mu)^T P^{-1}(\mathbf{x}-\mu)}{2}\right),$$

  where μ is a vector and P is a square matrix of appropriate dimensions.
- For a Gaussian vector, we have:
  - $\mathbb{E}[\mathbf{X}] = \mu$;
  - $V[\mathbf{X}] = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right] = P$.

Part XI: Functions

1. Functions of random variables.
2. Expected values of functions.

Functions of a random variable

- Example: If $Y = aX + b$, with $a > 0$, then:

$$f(X) \leq y \Rightarrow X \leq \frac{y-b}{a},$$

and then:

$$F_Y(y) = \mathbb{P}(\{f(X) \leq y\}) = \mathbb{P}\left(\left\{X \leq \frac{y-b}{a}\right\}\right),$$

so $F_Y(y) = F_X\left(\frac{y-b}{a}\right)$.

Functions of a random variable

- Example: If $Z = X + Y$, then:

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) \\ &= \mathbb{P}(X + Y \leq z) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} p_{X,Y}(x, y)\, dxdy. \end{aligned}$$

Linear combinations and conditioning

- For a vector $\mathbf{Y}$ of dimension m such that $\mathbf{Y} = A\mathbf{X}$, where A is a matrix, we have:

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^m \det Q}} \exp\left(-\frac{(\mathbf{y}-\nu)^T Q^{-1}(\mathbf{y}-\nu)}{2}\right),$$

  where $\nu = A\mu$ and $Q = APA^T$.

- If $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian vectors, $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ is also Gaussian with

$$\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathrm{Cov}[\mathbf{X},\mathbf{Y}]V[\mathbf{Y}]^{-1}(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]),$$

$$V[\mathbf{X}|\mathbf{Y}] = V[\mathbf{X}] - \mathrm{Cov}[\mathbf{X},\mathbf{Y}]V[\mathbf{Y}]^{-1}\mathrm{Cov}[\mathbf{Y},\mathbf{X}].$$

Also, we have:

- If $Y = f(X)$ and X has density $p_X(x)$, then

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \int f(x) p_X(x)\, dx.$$

- Concepts of covariance, correlation, etc, defined just as for discrete random variables.

Part XII: Important concepts

1. Conditioning.
2. Independence.

Usual solution for conditioning

- Introduce:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{X,Y}(x,y)}{\int p_{X,Y}(x,y)\,dx}.$$

- Then:

$$\mathbb{E}[X|Y] = \int x p_{X|Y}(x|Y)\,dx.$$

Independence: basic definition

- Consider random variables $X_1, X_2, \ldots, X_n$.
- These random variables are independent:
  - When all distributions have densities,

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i)\,.$$

Part XIII: Some advanced concepts

1. Measurability and fields.
2. Kolmogorovian conditioning.
3. Other definitions of independence.
4. Convergence.
5. Laws of large numbers.
6. Exchangeability.
7. Central limit theorem.
8. Bayesian consensus.

Digression: Measurability

- Given an infinite Ω, we have to specify probability values for its subsets.
- All subsets?
  1. It is *impossible* to define a countably additive probability measure over all subsets of the real numbers, for which the probability of an interval $[a, b]$ is $b - a$.
  2. In fact, there are subsets of $\Re$ that are *unmeasurable*: a countable additive set-function cannot be defined on them such that $[a, b]$ maps to $b - a$.
  3. Ulam's theorem: if a countably additive measure is defined over all subsets of the real numbers and vanishes on all singletons, it is identically zero.

Kolmogorov's solution: fields

- Consider first a *finite* set Ω.
- A *field* $\mathcal{F}$ is a nonempty set of subsets of Ω such that:
  1. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
  2. if $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$.
- Note: if A is in a field, then A^c, $\emptyset$ and Ω are automatically in the field.
- Example:
  $\{\emptyset, A, B, A^c, B^c, A \cup B, A^c \cup B, A \cup B^c, A^c \cup B^c, A \cap B, A^c \cap B, A \cap B^c, A^c \cap B^c, (A^c \cap B) \cup (A \cap B^c), (A \cup B^c) \cap (A^c \cup B), \Omega\}$.

σ-fields

- Now consider an infinite set Ω.
- A σ-field is a set of subsets of Ω such that
  1. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
  2. if $A_i \in \mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$.
- Note that σ-fields are closed under *countable* unions.

Fields and algebras

- Fields are also called *algebras*.
- σ fields are also called *σ-algebras*.
- In fact, "algebra" seems to be a better term (there are other meanings for the word "field" that do not apply here...).
- Terminology is confusing!

Borel algebras

- "Minimal σ-algebra containing the open sets/compact sets of a topological set Ω."
- The Borel algebra for the real numbers:
  - The smallest σ-algebra on $\Re$ that contains the intervals.
- The elements of a Borel algebra are the *Borel sets*.

Consequences of countable additivity

- No way to extend arbitrary assessments over arbitrary spaces.
- No uniform distribution on the integers.

BUT: countable additivity basically allows us to use integrals to compute expectations!

VERY important: there is a unique probability measure that corresponds to an expectation (and vice-versa)!

Conditioning, again

- Conditional expectation:

$$\mathbb{E}[X|A] = \frac{\mathbb{E}[XA]}{\mathbb{P}(A)}$$

  whenever $\mathbb{P}(A) > 0$.
- Now consider $\mathbb{E}[X|Y]$.
  - If Ω is uncountable, $\mathbb{E}[X|Y = y]$ may face the difficulty that one can have $\mathbb{P}(Y = y) = 0$ for *all* values of Y.
  - So, it is hard to define $\mathbb{P}(X|Y)$ as a function of X and Y.

Real thing: Kolmogorovian conditioning

- Definition: $\mathbb{E}[X|Y]$ is a *random variable* that is "Y-measurable" and such that

$$\mathbb{E}[f(Y)(X - \mathbb{E}[X|Y])] = 0$$

  for any function $f(Y)$.
- This is *not* simple!
- Usually, a proper density $p_{X|Y}(x|y)$ exists such that "probabilities" $\mathbb{P}(A(X)|B(Y))$ can be calculated.

Independence, again

- Variables $\{X_i\}_{i=1}^n$ are *independent* if

$$\mathbb{E}[f_i(X_i) | \cap_{j \neq i} \{X_j \in A_j\}] = \mathbb{E}[f_i(X_i)]\,,$$

  - for all functions $f_i(X_i)$ and
  - all events $\cap_{j \neq i} \{X_j \in A_j\}$ with positive probability.
- For all functions $f_i(X_i)$,

$$\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[f_i(X_i)]\,.$$

- For all sets of events $\{A_i\}_{i=1}^n$,

$$\mathbb{P}(\cap_{i=1}^n \{X_i \in A_i\}) = \prod_{i=1}^n \mathbb{P}(\{X_i \in A_i\})\,.$$

Convergence

There are many notions of convergence for random variables.
Consider a sequence of random variables $\{X_i\}$ defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Convergence in distribution (function): $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$
2. Convergence in probability: $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$.
3. Almost sure convergence (with probability one): $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

Review: Law of Large Numbers

Consider an infinite sequence of independent variables with expectation μ, variance σ^2. Define $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

- Weak law of large numbers:

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X} - \mu| \geq \epsilon\right) = 0.$$

- Strong law of large numbers:

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1.$$

Central limit theorem

- Take sequence of n independent random variables X_i with mean μ_i and variance σ_i^2.
- Consider the random variable $X = \sum_i X_i$; then $\mathbb{E}[X] = \sum_i \mu_i$ and variance $\sigma^2 = \sum_i \sigma_i^2$.
- If we define

$$Z = \frac{X - \mu}{\sigma},$$

then the distribution of Z tends to a Gaussian distribution with expectation 0 and variance 1 as $n \to \infty$.

Exchangeability

1. Binary variables $X_1, X_2, \ldots$ are exchangeable if $\mathbb{P}(X_1, X_2, \ldots)$ does not change if we just change the order of variables.
2. If $X_1, X_2, \ldots$ are exchangeable, then

$$\mathbb{P}(k \text{ ones in } n \text{ selected variables})$$

can always be written as

$$\int \binom{n}{k} \theta^k (1-\theta)^{n-k} p(\theta)\, d\theta.$$

(This is De Finetti's representation theorem.)

3. Note the deep implications of exchangeability!

Bayesian consensus

1. Suppose that n Bayesians have different priors $P_i(\theta)$.
2. Suppose they all observe $X_1, X_2, \ldots$.
3. Suppose all X_i are independent and identically distributed $\mathbb{P}(X_i|\theta^*)$.
4. Then all Bayesians will reach an identical posterior $\mathbb{P}(\theta|X_1, X_2, \ldots)$ that is infinitely concentrated around θ^*.

Summary

1. Kolmogorov's theory: probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a general possibility space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a non-negative, normalized to unity and countably additive set-function (a normalized measure).
2. Random variables are $\mathcal{F}$-measurable functions, expectations are Lebesgue integrals.
3. Random variables are $\mathcal{F}$-measurable functions, expectations are Lebesgue integrals (there are many *univariate* and *multivariate* densities!).
4. Conditional density $p(X|Y)$ is given by $p(X, Y)/p(Y)$; independence means "conditional equal to unconditional" or "factorization" (actually only the latter).
5. Convergence concepts are important, with many results: laws of large numbers, central limit theorems...