A review of probability theory

Lloyd Elliott

October 6, 2010

Random variables

Moments and expectations

Conditional distributions, marginal distributions & independence

Random variables

Random variables

Let Ω be a set of events. The probability of a subset $A \subseteq \Omega$ is denoted $\Pr(A)$.

Random variables

Let Ω be a set of events. The probability of a subset $A \subseteq \Omega$ is denoted $\Pr(A)$.

Example

Let $\Omega = \{\mathsf{H}, \mathsf{T}\}$. Let $X(\omega) = \begin{cases} 1, & \omega = \mathsf{H}. \\ -1, & \text{otherwise.} \end{cases}$

A random variable on Ω is a function $X : \Omega \to \mathbb{R}$.

A random variable on Ω is a function $X : \Omega \to \mathbb{R}$.

Example

Let $\Omega_n = \{(r_1, \ldots, r_n)\}, r_i \in \{1, 2, \ldots, 6\}$. Let $X_n(r_1, \ldots, r_n) = \sum_{i=1}^{n} r_i$. Here, X can model the sum of the outcomes of n die rolls. Let $Y_i(r_1, \ldots, r_n) = r_1$. Here, Y_i is the outcome of the i-th roll. Let $Z = \#\{i \ s.t. \ t_i = 1, 1 \leq i \leq n\}$. Here, Z is the number of 1s rolled.

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \rightarrow [0, 1]$ by $F(t) = \Pr(X \leq t)$.

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \to [0, 1]$ by $F(t) = \Pr(X \leq t)$.
The cdf of X has the following properties:

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \rightarrow [0, 1]$ by $F(t) = \Pr(X \leq t)$.
The cdf of X has the following properties:

- $\lim_{t \rightarrow -\infty} F(t) = 0$

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \to [0, 1]$ by $F(t) = \Pr(X \leq t)$.
The cdf of X has the following properties:

- $\lim_{t \to -\infty} F(t) = 0$
- $\lim_{t \to \infty} F(t) = 1$

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \to [0,1]$ by $F(t) = \Pr(X \leq t)$.
The cdf of X has the following properties:

- $\lim_{t \to -\infty} F(t) = 0$
- $\lim_{t \to \infty} F(t) = 1$
- $\lim_{t \to t_0^+} F(t) = F(t_0)$

The *cumulative distribution function* (or *cdf*) of X is the function $F : x \rightarrow [0, 1]$ by $F(t) = \Pr(X \leq t)$.
The cdf of X has the following properties:

- $\lim_{t \rightarrow -\infty} F(t) = 0$
- $\lim_{t \rightarrow \infty} F(t) = 1$
- $\lim_{t \rightarrow t_0^+} F(t) = F(t_0)$
- F is non decreasing.

If F is continuous and differentiable everywhere then the function $f(t) = F'(t)$ is called the *probability density function* (or the *pdf*) of X.

If F is continuous and differentiable everywhere then the function $f(t) = F'(t)$ is called the *probability density function* (or the *pdf*) of X.

If Ω is finite then for each $\omega \in \Omega$ the *probability distribution function* (which is also referred to as the *pdf*) of X is the function $p : X(\Omega) \to [0,1]$ such that $p(x) = \Pr(X^{-1}(x))$. Sometimes $p(x)$ is written $P(X = x)$ or $p_X(x)$ to emphasise which random variable the pdf is from.

If F is continuous and differentiable everywhere then the function $f(t) = F'(t)$ is called the *probability density function* (or the *pdf*) of X.
If Ω is finite then for each $\omega \in \Omega$ the *probability distribution function* (which is also referred to as the *pdf*) of X is the function $p : X(\Omega) \to [0, 1]$ such that $p(x) = \Pr(X^{-1}(x))$. Sometimes $p(x)$ is written $P(X = x)$ or $p_X(x)$ to emphasise which random variable the pdf is from. If Ω is identified with $X(\Omega)$ then $\Pr(x) = p(x)$.

Example

Let $\Omega = \{(r_1, r_2) \textit{ s.t. } 1 \leq r_1 \leq 6, 1 \leq r_2 \leq 6\}$ be the outcomes of two successive 6-sided die rolls. Let $X = r_1 + r_2$.

$$\begin{aligned}
p(7) &= \Pr(X^{-1}(7)), \\
&= \Pr(\{r_1, r_2 \textit{ s.t. } r_1 + r_2 = 7\}), \\
&= \Pr(\{(1,6),(2,5),(3,4),(4,3),(2,5),(1,6)\}), \\
&= 2(a_1 a_6 + a_2 a_5 + a_3 a_4)
\end{aligned}$$

Moments and expectations

Moments and expectations

If Ω is finite then the expected value of a random variable X on Ω is defined as follows:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\Omega) \Pr(\Omega) \\ &= \sum_{x \in X(\Omega)} x p(x).\end{aligned}$$

Moments and expectations

If Ω is finite then the expected value of a random variable X on Ω is defined as follows:

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

If Ω is not finite and if x has a pdf then the expected value of X is defined as follows:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t p(t) \mathrm{d}t.$$

Example

Roll n 6-sided die and let X_n be their sum. Assume the dice are independent. Assume they have biases $\Pr(r_i = j) = a_j$ $(a_1 + \ldots + a_6 = 1)$

$$
\begin{aligned}
\mathbb{E}[X_n] &= \sum_{r_1=1}^{6} \dots \sum_{r_n=1}^{6} (r_1 + \dots + r_n) \Pr(r_1) \dots \Pr(r_n), \\
&= \sum_{r_1=1}^{6} \dots \sum_{r_{n-1}=1}^{6} \left(\sum_{n=1}^{6} \Pr(r_n) \right) (r_1 + \dots + r_{n-1}) \Pr(r_1) \dots \Pr(r_{n-1}) \\
&\quad + \sum_{r_1=1}^{6} \dots \sum_{r_n=1}^{6} r_n \Pr(r_1) \dots \Pr(r_n), \\
&= \sum_{r_1=1}^{6} \dots \sum_{r_{n-1}=1}^{6} (r_1 + \dots + r_{n-1}) \Pr(r_1) \dots \Pr(r_{n-1}) \\
&\quad + \sum_{r_1=1}^{6} \Pr(r_1) \sum_{r_2=1}^{6} \Pr(r_2) \dots \sum_{r_n} r_n \Pr(r_n), \\
&= \mathbb{E}[X_{n-1}] + \sum_{r_n} r_n \Pr(n) \\
&= \mathbb{E}[X_{n-1}] + a_1 + 2a_2 + \dots + 6a_6 = n(a_1 + 2a_2 + \dots + 6a_6)
\end{aligned}
$$

Suppose $a_j = \frac{1}{6}$ for $1 \leq j \leq 6$. Then $\mathbb{E}[X_n] = \frac{7}{2}n$. Note: later we'll see an easier way of solving this. Also, you don't need to show this much work on your assignments.

Often Ω is implicit and $\sum_{x \in X(\omega)} xp(x)$ is written as $\sum_{X=x} xp(x)$ or even as $\sum_{x} xp(x)$.

Often Ω is implicit and $\sum_{x \in X(\omega)} xp(x)$ is written as $\sum_{X=x} xp(x)$ or even as $\sum_{x} xp(x)$.

The expected value $\mathbb{E}[X]$ is sometimes written $\langle X \rangle_{p(X)}$ or μ_X. It is also called the mean.

Theorem (Law of the unconscious statistician)

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function. If Ω is finite and X is a random variable on Ω with pdf p:

$$\mathbb{E}[\varphi(X)] = \sum_{x} \varphi(x) p(x).$$

Theorem (Law of the unconscious statistician)

Let $\varphi : \mathbb{R} \to \mathbb{R}$ *be a function. If* Ω *is finite and* X *is a random variable on* Ω *with pdf* p*:*

$$\mathbb{E}[\varphi(X)] = \sum_{x} \varphi(x)p(x).$$

If Ω *is infinite and* X *has a pdf* f *then:*

$$\mathbb{E}[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(t)f(t)\mathrm{d}t.$$

If X is a random variable and φ is a function, then $\varphi(X)$ is also a random variable.

If X is a random variable and φ is a function, then $\varphi(X)$ is also a random variable.

Theorem

$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$

The variance of X (denoted $\mathrm{Var}[X]$ or σ_X^2) is:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The variance of X (denoted $\text{Var}[X]$ or σ_X^2) is:

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Theorem

$$Var[aX + b] = a^2\mathbb{E}[X] + b.$$

The variance of X (denoted $\mathrm{Var}[X]$ or σ_X^2) is:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Theorem

$$Var[aX + b] = a^2\mathbb{E}[X] + b.$$

Corollary

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Definition

The n-th moment of a random variable X is $\mathbb{E}[X^n]$. The n-th central moment is $\mathbb{E}[(X - \mu_X)^n]$. So the mean is the first moment and the variance is the second central moment.

Theorem (Jensen's inequality)

If φ is a convex function then $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.

Theorem (Jensen's inequality)

If φ is a convex function then $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.

You can easily remember in which direction Jensen's inequality goes with this short memory trick: $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ is certainly sometimes true because $\mathbb{E}[X]^2 = 0$ if $\mathbb{E}[X] = 0$ and $\mathbb{E}[X]^2 \geq 0$. And $\varphi(t) = t^2$ is convex. So Jensen's inequality must be in the direction $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.

Conditional distributions, marginal distributions & independence

Definition

Suppose $(X_i)_{i=1}^n$ are random vectors. Their *joint* comulative distribution function is $F : \mathbb{R}^n \to \infty$ by:

$$F(t_1, \ldots, t_n) = \Pr(X_1 \leq t_1, \ldots, X_n \leq t_n).$$

Conditional distributions, marginal distributions & independence

Definition

Suppose $(X_i)_{i=1}^n$ are random vectors. Their *joint* comulative distribution function is $F : \mathbb{R}^n \to \infty$ by:

$$F(t_1, \ldots, t_n) = \Pr(X_1 \le t_1, \ldots, X_n \le t_n).$$

Sometimes $F(t_1, \ldots, t_n)$ is denoted $F_{X_1,\ldots,X_n}(t_1, \ldots, t_n)$

Conditional distributions, marginal distributions & independence

Definition

Suppose $(X_i)_{i=1}^n$ are random vectors. Their *joint* comulative distribution function is $F : \mathbb{R}^n \to \infty$ by:

$$F(t_1, \ldots, t_n) = \Pr(X_1 \leq t_1, \ldots, X_n \leq t_n).$$

Sometimes $F(t_1, \ldots, t_n)$ is denoted $F_{X_1,\ldots,X_n}(t_1, \ldots, t_n)$ If for all subsets $X_{k_1}, \ldots, X_{k_\ell}$ of $X_1, \ldots, X_n$:

$$F_{X_{k_1},\ldots,X_{k_\ell}}(t_{k_1}, \ldots, t_{k_\ell}) = F_{X_{k_1}}(x_{k_1} \ldots F_{X_{k_\ell}}(x_{k_\ell})$$

then $X_{k_1}, \ldots, X_{k_\ell}$ are said to be independent (written $X_1, \ldots, X_n \perp$ or $X_1 \perp X_2$ if $n = 2$).

Definition

If F is continuous and has continuous partial derivatives then the *joint* probability density function is:

$$f(t_1, \ldots, t_n) = \frac{\partial^n}{\partial_1 t_1 \ldots \partial_n t_n} F(t_1, \ldots, t_n).$$

Definition

If Ω is finite the *joint* probability distribution function is $p : (x_1, \ldots, x_n) \to \mathbb{R}$ by:

$$p(x_1, \ldots, x_n) = \Pr(X_1^{-1}(x_1) \cap \ldots \cap X_n^{-1}(x_n)).$$

Definition

If Ω is finite the *joint* probability distribution function is $p : (x_1, \ldots, x_n) \rightarrow \mathbb{R}$ by:

$$p(x_1, \ldots, x_n) = \Pr(X_1^{-1}(x_1) \cap \ldots \cap X_n^{-1}(x_n)).$$

Sometimes $p(x_1, \ldots, x_n)$ is written $p(\mathcal{X})$ if $\mathcal{X}$ is the vector $\mathcal{X} = (x_1, \ldots, x_n)$.

$$f(t_1, \ldots, t_n) = \frac{\partial^n}{\partial_1 t_1 \ldots \partial_n t_n} F(t_1, \ldots, t_n).$$

Suppose $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^m$ are random variables. The *conditional* cumulative distribution function $F(X_1, \ldots, X_n | Y_1, \ldots, Y_m)$ is:

$$F(x_1, \ldots, x_n | y_1, \ldots, y_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m).$$

Suppose $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^m$ are random variables. The *conditional* cumulative distribution function $F(X_1, \ldots, X_n | Y_1, \ldots, Y_m)$ is:

$$F(x_1, \ldots, x_n | y_1, \ldots, y_n) = \\ \Pr(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m).$$

If for all subsets $X_{k_1}, \ldots, X_{k_\ell}$ of $X_{k_1}, \ldots, X_{k_\ell}$,

$$F_{X_{k_1},\ldots,X_{k_\ell}|Y_1,\ldots,Y_m}(x_{k_1}, \ldots, x_{k_\ell}) = \\ F_{X_{k_1}|Y_1,\ldots,Y_m}(x_{k_1}) \ldots F_{X_{k_\ell}|Y_1,\ldots,Y_m}(x_{k_\ell}).$$

then $X_{k_1}, \ldots, X_{k_\ell}$ are conditionally independent given $Y_1, \ldots, Y_m$ (written $X_1, \ldots, X_n \perp | Y_1, \ldots, Y_m$).

If F has partial derivatives everywhere then the *conditional* probability density function is:

$$f(x_1, \ldots, x_n | y_1, \ldots, y_n)$$
$$= \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, \ldots, x_n | Y_1 = y_1, \ldots, Y_m = y_m).$$

If F has partial derivatives everywhere then the *conditional* probability density function is:

$$f(x_1, \ldots, x_n | y_1, \ldots, y_n) \\ = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, \ldots, x_n | Y_1 = y_1, \ldots, Y_m = y_m).$$

The conditional probability distribution function for finite Ω is by analogy with the joint case.

Let $\varphi : \mathbb{R}^n \to \mathbb{R}^m$. If Ω is finite, and $X_1, \ldots, X_n$ are random variables on Ω then the expected value of $\varphi(X_1, \ldots, X_n)$ is:

$$\begin{aligned}
\mathbb{E}_{X_1,\ldots,X_n}[\varphi(X_1, \ldots, X_n)] \\
&= \sum_{x_1 \in X_1(\Omega)} \cdots \sum_{x_n \in X_n(\Omega)} \varphi(x_1, \ldots, x_n) p(x_1, \ldots, x_n) \\
&= \sum_{x_1,\ldots,x_n} \varphi(x_1, \ldots, x_n) p(x_1, \ldots, x_n), \\
&= \sum_{\mathcal{X}} \varphi(\mathcal{X}) p(\mathcal{X}).
\end{aligned}$$

Sometimes this is denoted $\langle \varphi(\mathcal{X}) \rangle_{\mathcal{X}}$ or $\mathbb{E}_{\mathcal{X}}(\varphi(\mathcal{X}))$.

Suppose $\mathcal{X} = (X_i)_{i=1}^n, \mathcal{Y} = (Y_i)_{i=1}^n$ are random variables. For Ω finite the expected value of $\varphi(\mathcal{X})$ conditioned on $\mathcal{Y} = (y_1, \ldots, y_n)$ is:

$$\mathbb{E}[\varphi(\mathcal{X})|\mathcal{Y}] = \sum_{\mathcal{X}} \varphi(\mathcal{X}) p(\mathcal{X}|\mathcal{Y}).$$

Suppose $\mathcal{X} = (X_i)_{i=1}^n, \mathcal{Y} = (Y_i)_{i=1}^n$ are random variables. For Ω finite the expected value of $\varphi(\mathcal{X})$ conditioned on $\mathcal{Y} = (y_1, \ldots, y_n)$ is:

$$\mathbb{E}[\varphi(\mathcal{X})|\mathcal{Y}] = \sum_{\mathcal{X}} \varphi(\mathcal{X}) p(\mathcal{X}|\mathcal{Y}).$$

Sometimes this is written as $\langle \varphi(\mathcal{X}) \rangle_{p(\mathcal{X}|\mathcal{Y})}$ or $\mathbb{E}_{\mathcal{X}|\mathcal{Y}}[\varphi(\mathcal{X})]$.

Suppose $\mathcal{X} = (X_i)_{i=1}^n, \mathcal{Y} = (Y_i)_{i=1}^n$ are random variables. For Ω finite the expected value of $\varphi(\mathcal{X})$ conditioned on $\mathcal{Y} = (y_1, \ldots, y_n)$ is:

$$\mathbb{E}[\varphi(\mathcal{X})|\mathcal{Y}] = \sum_{\mathcal{X}} \varphi(\mathcal{X}) p(\mathcal{X}|\mathcal{Y}).$$

Sometimes this is written as $\langle \varphi(\mathcal{X}) \rangle_{p(\mathcal{X}|\mathcal{Y})}$ or $\mathbb{E}_{\mathcal{X}|\mathcal{Y}}[\varphi(\mathcal{X})]$.

$$\mathbb{E}_{\mathcal{Y}}[\mathbb{E}_{\mathcal{X}}[\varphi(\mathcal{X})|\mathcal{Y}]] = \mathbb{E}_{\mathcal{X}}[\varphi(\mathcal{X})].$$

If $\mathcal{X}, \mathcal{Y}$ are random variables with probability distribution function $p(\mathcal{X}, \mathcal{Y})$ then the *marginal* distribution function of $\mathcal{X}$ is:

$$
\begin{aligned}
p(\mathcal{X}) &= \sum \mathcal{Y} p(\mathcal{X}|\mathcal{Y}) p(\mathcal{Y}), \\
&= \mathbb{E}_{\mathcal{Y}}[p(\mathcal{X}|\mathcal{Y})].
\end{aligned}
$$

Example

Let Y_1, Y_2 be the outcome of two successive die rolls. Suppose they are independent ($Y_1 \perp Y_2$). Let $X = Y_1 + Y_2$. Then,

$$Y_1 \not\perp Y_2 | X. \tag{1}$$

Theorem (linearity of expected value)

Let $X_1, \ldots, X_n$ be any random variables.

$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n].$$

Throw n dice and let X_n again be the sum of their spots. Don't assume independence. But assume each die has the same distribution: $\Pr(r_i = j) = a_j$. The linearity of expected value provides a much simpler way to compute $E[X_n]$. Suppose r_i is the number of spots on the i-th die. Then, $X_n = r_1 + \ldots + r_n$ has expected value:

Throw n dice and let X_n again be the sum of their spots. Don't assume independence. But assume each die has the same distribution: $\Pr(r_i = j) = a_j$. The linearity of expected value provides a much simpler way to compute $E[X_n]$. Suppose r_i is the number of spots on the i-th die. Then, $X_n = r_1 + \ldots + r_n$ has expected value:

$$\begin{aligned}\mathbb{E}[X_n] &= \mathbb{E}[r_1] + \ldots + \mathbb{E}[r_n] \\ &= n(a_1 + 2a_2 + \ldots + na_n).\end{aligned}$$

If $X_1, \ldots, X_n$ are random variables then the covariance between X_i and X_j is:

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

If $X_1, \ldots, X_n$ are random variables then the covariance between X_i and X_j is:

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

Some properties of covariance:

1. $\text{Cov}[X_i, X_i] = \text{Var}[X_i]$
2. $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$
3. $\text{Cov}[aX_i, bX_j] = ab\text{Cov}[X_i, X_j]$
4. $\text{Cov}[X_1 + X_2, X_3] = \text{Cov}[X_1, X_3] + \text{Cov}[X_2, X_3]$

If $X_1, \ldots, X_n$ are random variables then the covariance between X_i and X_j is:

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

Some properties of covariance:

1. $\text{Cov}[X_i, X_i] = \text{Var}[X_i]$
2. $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$
3. $\text{Cov}[aX_i, bX_j] = ab\text{Cov}[X_i, X_j]$
4. $\text{Cov}[X_1 + X_2, X_3] = \text{Cov}[X_1, X_3] + \text{Cov}[X_2, X_3]$

These properties mean Cov is a bilinear form (like an inner product).

Let X_n be the sum of the outcomes of n successive die rolls.
Define $S_n = \dfrac{X_n - n\mathbb{E}[X]}{\sqrt{n\mathrm{Var}[X]}}$.

Let X_n be the sum of the outcomes of n successive die rolls.
Define $S_n = \dfrac{X_n - n\mathbb{E}[X]}{\sqrt{n\mathrm{Var}[X]}}$.

$$f_{S_n}(x) \longrightarrow \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \text{ as } n \to \infty.$$

These questions don't matter for this course (nor do they matter for much of probability theory) and so they're conscious omissions from this tutorial:

These questions don't matter for this course (nor do they matter for much of probability theory) and so they're conscious omissions from this tutorial:

1. What about discrete distributions (where Ω is infinite but the cdf of a random variable on Ω is not differentiable)?

These questions don't matter for this course (nor do they matter for much of probability theory) and so they're conscious omissions from this tutorial:

1. What about discrete distributions (where Ω is infinite but the cdf of a random variable on Ω is not differentiable)?
2. How is the expected value defined for random variables that don't have pdfs?

These questions don't matter for this course (nor do they matter for much of probability theory) and so they're conscious omissions from this tutorial:

1. What about discrete distributions (where Ω is infinite but the cdf of a random variable on Ω is not differentiable)?
2. How is the expected value defined for random variables that don't have pdfs?
3. What properties do $\Omega, \Pr(A)$ have to satisfy?