18.600: Lecture 31

Lectures 19-30 Review

Scott Sheffield

MIT

Outline

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a, b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a, b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.
- Probability of any single point is zero.

Continuous random variables

- Say X is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that $P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx$.
- We may assume $\int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$ and f is non-negative.
- Probability of interval $[a, b]$ is given by $\int_a^b f(x)dx$, the area under f between a and b.
- Probability of any single point is zero.
- Define **cumulative distribution function** $F(a) = F_X(a) := P\{X < a\} = P\{X \leq a\} = \int_{-\infty}^{a} f(x)dx$.

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

- How should we define $E[X]$ when X is a continuous random variable?

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.
- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[g(X)] = \sum_{x:p(x)>0} p(x)g(x).$$

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.
- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[g(X)] = \sum_{x:p(x)>0} p(x)g(x).$$

- What is the analog when X is a continuous random variable?

Expectations of continuous random variables

- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[X] = \sum_{x:p(x)>0} p(x)x.$$

- How should we define $E[X]$ when X is a continuous random variable?
- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)xdx$.
- Recall that when X was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote

$$E[g(X)] = \sum_{x:p(x)>0} p(x)g(x).$$

- What is the analog when X is a continuous random variable?
- Answer: we will write $E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x)dx$.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big) f(x)dx = E[g_1(X)] + E[g_2(X)]$.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big)f(x)dx = E[g_1(X)] + E[g_2(X)]$.
- Furthermore, $E[ag(X)] = aE[g(X)]$ when a is a constant.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big) f(x)dx = E[g_1(X)] + E[g_2(X)]$.
- Furthermore, $E[ag(X)] = aE[g(X)]$ when a is a constant.
- Just as in the discrete case, we can expand the variance expression as $\mathrm{Var}[X] = E[X^2 - 2\mu X + \mu^2]$ and use additivity of expectation to say that $\mathrm{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.

Variance of continuous random variables

- Suppose X is a continuous random variable with mean μ.
- We can write $\mathrm{Var}[X] = E[(X-\mu)^2]$, same as in the discrete case.
- Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int \big(g_1(x) + g_2(x)\big)f(x)dx = E[g_1(X)] + E[g_2(X)]$.
- Furthermore, $E[ag(X)] = aE[g(X)]$ when a is a constant.
- Just as in the discrete case, we can expand the variance expression as $\mathrm{Var}[X] = E[X^2 - 2\mu X + \mu^2]$ and use additivity of expectation to say that $\mathrm{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.
- This formula is often useful for calculations.

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^{n} X_n$.

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^{n} X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.
- **Poisson point process**: toss one λ/n coin during each length $1/n$ time increment, take $n \to \infty$ limit.

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.
- **Poisson point process**: toss one λ/n coin during each length $1/n$ time increment, take $n \to \infty$ limit.
- **Exponential**: time till first event in λ Poisson point process.

It's the coins, stupid

- Much of what we have done in this course can be motivated by the i.i.d. sequence X_i where each X_i is 1 with probability p and 0 otherwise. Write $S_n = \sum_{i=1}^n X_n$.
- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.
- **Poisson point process**: toss one λ/n coin during each length $1/n$ time increment, take $n \to \infty$ limit.
- **Exponential**: time till first event in λ Poisson point process.
- **Gamma distribution**: time till nth event in λ Poisson point process.

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).
- **Expectation of geometric random variable** with parameter p is $1/p$.

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).
- **Expectation of geometric random variable** with parameter p is $1/p$.
- **Expectation of binomial random variable** with parameters (n, p) is np.

Discrete random variable properties derivable from coin toss intuition

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).
- **Expectation of geometric random variable** with parameter p is $1/p$.
- **Expectation of binomial random variable** with parameters (n, p) is np.
- **Variance of binomial random variable** with parameters (n, p) is $np(1 - p) = npq$.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.
- **Sum of λ_1 Poisson and independent λ_2 Poisson** is a $\lambda_1 + \lambda_2$ Poisson.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.
- **Sum of λ_1 Poisson and independent λ_2 Poisson** is a $\lambda_1 + \lambda_2$ Poisson.
- **Times between successive events** in λ Poisson process are independent exponentials with parameter λ.

Continuous random variable properties derivable from coin toss intuition

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.
- **Sum of λ_1 Poisson and independent λ_2 Poisson** is a $\lambda_1 + \lambda_2$ Poisson.
- **Times between successive events** in λ Poisson process are independent exponentials with parameter λ.
- **Minimum of independent exponentials** with parameters λ_1 and λ_2 is itself exponential with parameter $\lambda_1 + \lambda_2$.

DeMoivre-Laplace Limit Theorem

- **DeMoivre-Laplace limit theorem (special case of central limit theorem):**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

DeMoivre-Laplace Limit Theorem

- **DeMoivre-Laplace limit theorem (special case of central limit theorem):**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- This is $\Phi(b) - \Phi(a) = P\{a \le X \le b\}$ when X is a standard normal random variable.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?
- Here $\sqrt{npq} = \sqrt{60000 \times \frac{1}{6} \times \frac{5}{6}} \approx 91.28$.

Problems

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?
- Here $\sqrt{npq} = \sqrt{60000 \times \frac{1}{6} \times \frac{5}{6}} \approx 91.28$.
- And $200/91.28 \approx 2.19$. Answer is about $1 - \Phi(-2.19)$.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- Function $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$ can't be computed explicitly.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- Function $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$ can't be computed explicitly.
- Values: $\Phi(-3) \approx .0013$, $\Phi(-2) \approx .023$ and $\Phi(-1) \approx .159$.

Properties of normal random variables

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- Function $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$ can't be computed explicitly.
- Values: $\Phi(-3) \approx .0013$, $\Phi(-2) \approx .023$ and $\Phi(-1) \approx .159$.
- Rule of thumb: "two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean."

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.
- Repeated integration by parts gives $E[X^n] = n!/\lambda^n$.

Properties of exponential random variables

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.
- Repeated integration by parts gives $E[X^n] = n!/\lambda^n$.
- If $\lambda = 1$, then $E[X^n] = n!$. Value $\Gamma(n) := E[X^{n-1}]$ defined for real $n > 0$ and $\Gamma(n) = (n-1)!$.

Defining Γ distribution

- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1}e^{-\lambda x}\lambda}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$.

Defining Γ distribution

- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1} e^{-\lambda x} \lambda}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$.
- Same as exponential distribution when $\alpha = 1$. Otherwise, multiply by $x^{\alpha-1}$ and divide by $\Gamma(\alpha)$. The fact that $\Gamma(\alpha)$ is what you need to divide by to make the total integral one just follows from the definition of Γ.

Defining Γ distribution

- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1} e^{-\lambda x} \lambda}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$.
- Same as exponential distribution when $\alpha = 1$. Otherwise, multiply by $x^{\alpha-1}$ and divide by $\Gamma(\alpha)$. The fact that $\Gamma(\alpha)$ is what you need to divide by to make the total integral one just follows from the definition of Γ.
- Waiting time interpretation makes sense only for integer α, but distribution is defined for general positive α.

Outline

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

Properties of uniform random variables

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta]. \end{cases}$

Properties of uniform random variables

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta]. \end{cases}$
- Then $E[X] = \frac{\alpha+\beta}{2}$.

Properties of uniform random variables

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta]. \end{cases}$
- Then $E[X] = \frac{\alpha+\beta}{2}$.
- And $\mathrm{Var}[X] = \mathrm{Var}[(\beta-\alpha)Y + \alpha] = \mathrm{Var}[(\beta-\alpha)Y] = (\beta-\alpha)^2\mathrm{Var}[Y] = (\beta-\alpha)^2/12$.

Distribution of function of random variable

- Suppose $P\{X \leq a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \leq 27\}$?

Distribution of function of random variable

- Suppose $P\{X \leq a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \leq 27\}$?
- Answer: note that $Y \leq 27$ if and only if $X \leq 3$. Hence $P\{Y \leq 27\} = P\{X \leq 3\} = F_X(3)$.

Distribution of function of random variable

- Suppose $P\{X \leq a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \leq 27\}$?
- Answer: note that $Y \leq 27$ if and only if $X \leq 3$. Hence $P\{Y \leq 27\} = P\{X \leq 3\} = F_X(3)$.
- Generally $F_Y(a) = P\{Y \leq a\} = P\{X \leq a^{1/3}\} = F_X(a^{1/3})$

Distribution of function of random variable

- Suppose $P\{X \leq a\} = F_X(a)$ is known for all a. Write $Y = X^3$. What is $P\{Y \leq 27\}$?
- Answer: note that $Y \leq 27$ if and only if $X \leq 3$. Hence $P\{Y \leq 27\} = P\{X \leq 3\} = F_X(3)$.
- Generally $F_Y(a) = P\{Y \leq a\} = P\{X \leq a^{1/3}\} = F_X(a^{1/3})$
- This is a general principle. If X is a continuous random variable and g is a strictly increasing function of x and $Y = g(X)$, then $F_Y(a) = F_X(g^{-1}(a))$.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \dots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \dots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.
- In other words, the probability mass functions for X and Y are the row and columns sums of $A_{i,j}$.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.
- In other words, the probability mass functions for X and Y are the row and columns sums of $A_{i,j}$.
- Given the joint distribution of X and Y, we sometimes call distribution of X (ignoring Y) and distribution of Y (ignoring X) the **marginal** distributions.

Joint probability mass functions: discrete random variables

- If X and Y assume values in $\{1, 2, \ldots, n\}$ then we can view $A_{i,j} = P\{X = i, Y = j\}$ as the entries of an $n \times n$ matrix.
- Let's say I don't care about Y. I just want to know $P\{X = i\}$. How do I figure that out from the matrix?
- Answer: $P\{X = i\} = \sum_{j=1}^{n} A_{i,j}$.
- Similarly, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$.
- In other words, the probability mass functions for X and Y are the row and columns sums of $A_{i,j}$.
- Given the joint distribution of X and Y, we sometimes call distribution of X (ignoring Y) and distribution of Y (ignoring X) the **marginal** distributions.
- In general, when X and Y are jointly defined discrete random variables, we write $p(x, y) = p_{X,Y}(x, y) = P\{X = x, Y = y\}$.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a,b) = P\{X \le a, Y \le b\}$.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \leq a, Y \leq b\}$.
- The region $\{(x, y) : x \leq a, y \leq b\}$ is the lower left "quadrant" centered at (a, b).

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \leq a, Y \leq b\}$.
- The region $\{(x, y) : x \leq a, y \leq b\}$ is the lower left "quadrant" centered at (a, b).
- Refer to $F_X(a) = P\{X \leq a\}$ and $F_Y(b) = P\{Y \leq b\}$ as **marginal** cumulative distribution functions.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \leq a, Y \leq b\}$.
- The region $\{(x, y) : x \leq a, y \leq b\}$ is the lower left "quadrant" centered at (a, b).
- Refer to $F_X(a) = P\{X \leq a\}$ and $F_Y(b) = P\{Y \leq b\}$ as **marginal** cumulative distribution functions.
- Question: if I tell you the two parameter function F, can you use it to determine the marginals F_X and F_Y?

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a,b) = P\{X \le a, Y \le b\}$.
- The region $\{(x,y) : x \le a, y \le b\}$ is the lower left "quadrant" centered at (a,b).
- Refer to $F_X(a) = P\{X \le a\}$ and $F_Y(b) = P\{Y \le b\}$ as **marginal** cumulative distribution functions.
- Question: if I tell you the two parameter function F, can you use it to determine the marginals F_X and F_Y?
- Answer: Yes. $F_X(a) = \lim_{b\to\infty} F(a,b)$ and $F_Y(b) = \lim_{a\to\infty} F(a,b)$.

Joint distribution functions: continuous random variables

- Given random variables X and Y, define $F(a, b) = P\{X \leq a, Y \leq b\}$.
- The region $\{(x, y) : x \leq a, y \leq b\}$ is the lower left "quadrant" centered at (a, b).
- Refer to $F_X(a) = P\{X \leq a\}$ and $F_Y(b) = P\{Y \leq b\}$ as **marginal** cumulative distribution functions.
- Question: if I tell you the two parameter function F, can you use it to determine the marginals F_X and F_Y?
- Answer: Yes. $F_X(a) = \lim_{b\to\infty} F(a, b)$ and $F_Y(b) = \lim_{a\to\infty} F(a, b)$.
- Density: $f(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x, y)$.

Independent random variables

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

Independent random variables

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

- When X and Y are discrete random variables, they are independent if $P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y for which $P\{X = x\}$ and $P\{Y = y\}$ are non-zero.

Independent random variables

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

- When X and Y are discrete random variables, they are independent if $P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y for which $P\{X = x\}$ and $P\{Y = y\}$ are non-zero.
- When X and Y are continuous, they are independent if $f(x, y) = f_X(x)f_Y(y)$.

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \leq a\} = \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$
  $$= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy.$$

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$
  $$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy.$$
- Differentiating both sides gives $f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.

Summing two random variables

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$ $$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy.$$
- Differentiating both sides gives $f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- Latter formula makes some intuitive sense. We're integrating over the set of x, y pairs that add up to a.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.
- We can *define* the conditional probability density of X given that $Y = y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.

Conditional distributions

- Let's say X and Y have joint probability density function $f(x, y)$.
- We can *define* the conditional probability density of X given that $Y = y$ by $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.
- This amounts to restricting $f(x, y)$ to the line corresponding to the given y value (and dividing by the constant that makes the integral along that line equal to 1).

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0, 1]$, independently of each other.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \dots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \dots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
- The n-tuple $(X_1, X_2, \dots, X_n)$ has a constant density function on the n-dimensional cube $[0, 1]^n$.
- What is the probability that the *largest* of the X_i is less than a?

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0, 1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0, 1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \ldots, X_n\}$, then what is the probability density function of X?

Maxima: pick five job candidates at random, choose best

- Suppose I choose n random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0,1]$, independently of each other.
- The n-tuple $(X_1, X_2, \ldots, X_n)$ has a constant density function on the n-dimensional cube $[0,1]^n$.
- What is the probability that the *largest* of the X_i is less than a?
- ANSWER: a^n.
- So if $X = \max\{X_1, \ldots, X_n\}$, then what is the probability density function of X?
- Answer: $F_X(a) = \begin{cases} 0 & a < 0 \\ a^n & a \in [0,1] \\ 1 & a > 1 \end{cases}$. And $f_x(a) = F'_X(a) = na^{n-1}$.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$

General order statistics

- Consider i.i.d random variables $X_1, X_2, \dots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \dots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \dots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \dots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \dots, Y_n)$ independent of each other?

General order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^n f(x_i)$ if $x_1 < x_2 \ldots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \ldots, Y_n)$ independent of each other?
- Yes.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x, y)$ then $E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$.

Properties of expectation

- Several properties we derived for discrete expectations continue to hold in the continuum.
- If X is discrete with mass function $p(x)$ then $E[X] = \sum_x p(x)x$.
- Similarly, if X is continuous with density function $f(x)$ then $E[X] = \int f(x)xdx$.
- If X is discrete with mass function $p(x)$ then $E[g(x)] = \sum_x p(x)g(x)$.
- Similarly, X if is continuous with density function $f(x)$ then $E[g(X)] = \int f(x)g(x)dx$.
- If X and Y have joint mass function $p(x,y)$ then $E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y)$.
- If X and Y have joint probability density function $f(x,y)$ then $E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X + Y] = E[X] + E[Y]$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y]=E[X]+E[Y]$.
- In both discrete and continuous settings, $E[aX]=aE[X]$ when a is a constant. And $E[\sum a_i X_i]=\sum a_i E[X_i]$.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X + Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0, 1]$ and note that $g(Y)$ has the same probability distribution as X.

Properties of expectation

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.
- So $E[X] = E[g(Y)] = \int_0^1 g(y)dy$, which is indeed the area under the graph of $1 - F_X$.

A property of independence

- If X and Y are independent then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

A property of independence

- If X and Y are independent then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- Just write $E[g(X)h(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy$.

A property of independence

- If X and Y are independent then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.
- Just write $E[g(X)h(Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy$.
- Since $f(x,y) = f_X(x)f_Y(y)$ this factors as $\int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx = E[h(Y)]E[g(X)]$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.
- Covariance formula $E[XY] - E[X]E[Y]$, or "expectation of product minus product of expectations" is frequently useful.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.
- Covariance formula $E[XY]-E[X]E[Y]$, or "expectation of product minus product of expectations" is frequently useful.
- If X and Y are independent then $\mathrm{Cov}(X,Y) = 0$.

Defining covariance and correlation

- Now define covariance of X and Y by $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])$.
- Note: by definition $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.
- Covariance formula $E[XY] - E[X]E[Y]$, or "expectation of product minus product of expectations" is frequently useful.
- If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.
- Converse is not true.

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$.

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$.

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$.
- **General statement of bilinearity of covariance:**

$$\mathrm{Cov}(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

Basic covariance facts

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$.
- **General statement of bilinearity of covariance:**

$$\mathrm{Cov}(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

- Special case:

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{(i,j):i<j} \mathrm{Cov}(X_i, X_j).$$

Defining correlation

- Again, by definition $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

Defining correlation

- Again, by definition $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Defining correlation

- Again, by definition $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX+b, cY+d) = \rho(X,Y)$.

Defining correlation

- Again, by definition $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX + b, cY + d) = \rho(X, Y)$.
- Satisfies $-1 \le \rho(X, Y) \le 1$.

Defining correlation

- Again, by definition $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX + b, cY + d) = \rho(X, Y)$.
- Satisfies $-1 \le \rho(X, Y) \le 1$.
- If a and b are positive constants and $a > 0$ then $\rho(aX + b, X) = 1$.

Defining correlation

- Again, by definition $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.
- **Correlation** of X and Y defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- Correlation doesn't care what units you use for X and Y. If $a > 0$ and $c > 0$ then $\rho(aX + b, cY + d) = \rho(X, Y)$.
- Satisfies $-1 \leq \rho(X, Y) \leq 1$.
- If a and b are positive constants and $a > 0$ then $\rho(aX + b, X) = 1$.
- If a and b are positive constants and $a < 0$ then $\rho(aX + b, X) = -1$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.
- In words: first restrict sample space to pairs (x, y) with given y value. Then divide the original mass function by $p_Y(y)$ to obtain a probability mass function on the restricted space.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.
- In words: first restrict sample space to pairs (x, y) with given y value. Then divide the original mass function by $p_Y(y)$ to obtain a probability mass function on the restricted space.
- We do something similar when X and Y are continuous random variables. In that case we write $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$.

Conditional probability distributions

- It all starts with the definition of conditional probability: $P(A|B) = P(AB)/P(B)$.
- If X and Y are jointly discrete random variables, we can use this to define a probability mass function for X *given* $Y = y$.
- That is, we write $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}$.
- In words: first restrict sample space to pairs (x, y) with given y value. Then divide the original mass function by $p_Y(y)$ to obtain a probability mass function on the restricted space.
- We do something similar when X and Y are continuous random variables. In that case we write $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$.
- Often useful to think of sampling (X, Y) as a two-stage process. First sample Y from its marginal distribution, obtain $Y = y$ for some particular y. Then sample X from its probability distribution *given* $Y = y$.

Conditional expectation

- Now, what do we mean by $E[X|Y=y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y=y$.

Conditional expectation

- Now, what do we mean by $E[X|Y=y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y=y$.
- Can write this as $E[X|Y=y] = \sum_x xP\{X=x|Y=y\} = \sum_x xp_{X|Y}(x|y)$.

Conditional expectation

- Now, what do we mean by $E[X|Y=y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y=y$.
- Can write this as $E[X|Y=y]=\sum_x xP\{X=x|Y=y\}=\sum_x xp_{X|Y}(x|y)$.
- Can make sense of this in the continuum setting as well.

Conditional expectation

- Now, what do we mean by $E[X|Y=y]$? This should just be the expectation of X in the conditional probability measure for X given that $Y=y$.
- Can write this as $E[X|Y=y] = \sum_x xP\{X=x|Y=y\} = \sum_x xp_{X|Y}(x|y)$.
- Can make sense of this in the continuum setting as well.
- In continuum setting we had $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$. So $E[X|Y=y] = \int_{-\infty}^{\infty} x\frac{f(x,y)}{f_Y(y)}dx$

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case: $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case: $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.
- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.

Conditional expectation as a random variable

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case: $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.
- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.
- $E[E[X|Y = y]] = \sum_y p_Y(y)\sum_x x\frac{p(x,y)}{p_Y(y)} = \sum_x \sum_y p(x,y)x = E[X]$.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] =$
  $E[X^2] - E[E[X|Y]^2]$.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\left[(X - E[X|Y])^2 | Y\right] = E\left[X^2 - E[X|Y]^2 | Y\right]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.
- One can discover X in two stages: first sample Y from marginal and compute $E[X|Y]$, then sample X from distribution given Y value.

Conditional variance

- Definition:
  $\mathrm{Var}(X|Y) = E\big[(X - E[X|Y])^2|Y\big] = E\big[X^2 - E[X|Y]^2|Y\big]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.
- One can discover X in two stages: first sample Y from marginal and compute $E[X|Y]$, then sample X from distribution given Y value.
- Above fact breaks variance into two parts, corresponding to these two stages.

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?
- What is $E[Z|X]$? And how about $\mathrm{Var}(Z|X)$?

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?
- What is $E[Z|X]$? And how about $\mathrm{Var}(Z|X)$?
- Both of these values are functions of X. Former is just X. Latter happens to be a constant-valued function of X, i.e., happens not to actually depend on X. We have $\mathrm{Var}(Z|X) = \sigma_Y^2$.

Example

- Let X be a random variable of variance σ_X^2 and Y an independent random variable of variance σ_Y^2 and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.
- What are the covariances $\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X, Z)$?
- How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?
- What is $E[Z|X]$? And how about $\mathrm{Var}(Z|X)$?
- Both of these values are functions of X. Former is just X. Latter happens to be a constant-valued function of X, i.e., happens not to actually depend on X. We have $\mathrm{Var}(Z|X) = \sigma_Y^2$.
- Can we check the formula $\mathrm{Var}(Z) = \mathrm{Var}(E[Z|X]) + E[\mathrm{Var}(Z|X)]$ in this case?

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?
- By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all t.

Moment generating functions

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?
- By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all t.
- In other words, adding independent random variables corresponds to multiplying moment generating functions.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \dots X_n$ are i.i.d. copies of X and $Z = X_1 + \dots + X_n$ then what is M_Z?

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \dots X_n$ are i.i.d. copies of X and $Z = X_1 + \dots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \dots X_n$ are i.i.d. copies of X and $Z = X_1 + \dots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.
- If $Z = aX$ then $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$.

Moment generating functions for sums of i.i.d. random variables

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.
- If $Z = aX$ then $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$.
- If $Z = X + b$ then $M_Z(t) = E[e^{tZ}] = E[e^{tX+bt}] = e^{bt}M_X(t)$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.
- If X is normal with mean 0, variance 1, then $M_X(t) = e^{t^2/2}$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.
- If X is normal with mean 0, variance 1, then $M_X(t) = e^{t^2/2}$.
- If X is normal with mean μ, variance σ^2, then $M_X(t) = e^{\sigma^2 t^2/2 + \mu t}$.

Examples

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.
- If X is normal with mean 0, variance 1, then $M_X(t) = e^{t^2/2}$.
- If X is normal with mean μ, variance σ^2, then $M_X(t) = e^{\sigma^2 t^2/2 + \mu t}$.
- If X is exponential with parameter $\lambda > 0$ then $M_X(t) = \frac{\lambda}{\lambda - t}$.

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0, 1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0, 1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.
- $F_X(x) = P\{X \leq x\} = P\{\tan\theta \leq x\} = P\{\theta \leq tan^{-1}x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}x$.

Cauchy distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0,1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.
- $F_X(x) = P\{X \leq x\} = P\{\tan\theta \leq x\} = P\{\theta \leq tan^{-1}x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}x$.
- Find $f_X(x) = \frac{d}{dx}F(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.

Beta distribution

- Two part experiment: first let p be uniform random variable $[0, 1]$, then let X be binomial (n, p) (number of heads when we toss n p-coins).

Beta distribution

- Two part experiment: first let p be uniform random variable $[0,1]$, then let X be binomial (n,p) (number of heads when we toss n p-coins).
- **Given** that $X = a - 1$ and $n - X = b - 1$ the conditional law of p is called the β distribution.

Beta distribution

- Two part experiment: first let p be uniform random variable $[0,1]$, then let X be binomial (n,p) (number of heads when we toss n p-coins).
- **Given** that $X = a-1$ and $n-X = b-1$ the conditional law of p is called the β distribution.
- The density function is a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.

Beta distribution

- Two part experiment: first let p be uniform random variable $[0,1]$, then let X be binomial (n,p) (number of heads when we toss n p-coins).
- **Given** that $X = a-1$ and $n - X = b-1$ the conditional law of p is called the β distribution.
- The density function is a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.
- That is $f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ on $[0,1]$, where $B(a,b)$ is constant chosen to make integral one. Can show $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Beta distribution

- Two part experiment: first let p be uniform random variable $[0, 1]$, then let X be binomial (n, p) (number of heads when we toss n p-coins).
- **Given** that $X = a - 1$ and $n - X = b - 1$ the conditional law of p is called the β distribution.
- The density function is a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.
- That is $f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ on $[0, 1]$, where $B(a, b)$ is constant chosen to make integral one. Can show $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
- Turns out that $E[X] = \frac{a}{a+b}$ and the mode of X is $\frac{(a-1)}{(a-1)+(b-1)}$.

Outline

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

Central limit theorem

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.

Central limit theorem

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.
- Write $S_n = \sum_{i=1}^n X_i$. So $E[S_n] = n\mu$ and $\mathrm{Var}[S_n] = n\sigma^2$ and $\mathrm{SD}[S_n] = \sigma\sqrt{n}$.

Central limit theorem

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.
- Write $S_n = \sum_{i=1}^{n} X_i$. So $E[S_n] = n\mu$ and $\mathrm{Var}[S_n] = n\sigma^2$ and $\mathrm{SD}[S_n] = \sigma\sqrt{n}$.
- Write $B_n = \frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$. Then B_n is the difference between S_n and its expectation, measured in standard deviation units.

Central limit theorem

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.
- Write $S_n = \sum_{i=1}^n X_i$. So $E[S_n] = n\mu$ and $\mathrm{Var}[S_n] = n\sigma^2$ and $\mathrm{SD}[S_n] = \sigma\sqrt{n}$.
- Write $B_n = \frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$. Then B_n is the difference between S_n and its expectation, measured in standard deviation units.
- **Central limit theorem:**

$$\lim_{n\to\infty} P\{a \le B_n \le b\} \to \Phi(b) - \Phi(a).$$

Weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.

Weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+...+X_n}{n}$ is called the *empirical average* of the first n trials.

Weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- We'd guess that when n is large, A_n is typically close to μ.

Weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+...+X_n}{n}$ is called the *empirical average* of the first n trials.
- We'd guess that when n is large, A_n is typically close to μ.
- Indeed, **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.

Weak law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- We'd guess that when n is large, A_n is typically close to μ.
- Indeed, **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- Example: as n tends to infinity, the probability of seeing more than $.50001n$ heads in n fair coin tosses tends to zero.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+...+X_n}{n}$ is called the *empirical average* of the first n trials.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+...+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.
- Recall: **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+...+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.
- Recall: **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- The **strong law of large numbers** states that with probability one $\lim_{n\to\infty} A_n = \mu$.

Strong law of large numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.
- Recall: **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- The **strong law of large numbers** states that with probability one $\lim_{n\to\infty} A_n = \mu$.
- It is called "strong" because it implies the weak law of large numbers. But it takes a bit of thought to see why this is the case.

Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties

CLE plus weak/strong laws

Markov chains

Outline

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.
- Interpret X_n as state of the system at time n.

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.
- Interpret X_n as state of the system at time n.
- Sequence is called a **Markov chain** if we have a fixed collection of numbers P_{ij} (one for each pair $i, j \in \{0, 1, \ldots, M\}$) such that whenever the system is in state i, there is probability P_{ij} that system will next be in state j.

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.
- Interpret X_n as state of the system at time n.
- Sequence is called a **Markov chain** if we have a fixed collection of numbers P_{ij} (one for each pair $i, j \in \{0, 1, \ldots, M\}$) such that whenever the system is in state i, there is probability P_{ij} that system will next be in state j.
- Precisely,
  $P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij}$.

Markov chains

- Consider a sequence of random variables $X_0, X_1, X_2, \dots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \dots, M\}$.
- Interpret X_n as state of the system at time n.
- Sequence is called a **Markov chain** if we have a fixed collection of numbers P_{ij} (one for each pair $i, j \in \{0, 1, \dots, M\}$) such that whenever the system is in state i, there is probability P_{ij} that system will next be in state j.
- Precisely,
  $P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P_{ij}$.
- Kind of an "almost memoryless" property. Probability distribution for next state depends only on the current state (and not on the rest of the state history).

Matrix representation

- To describe a Markov chain, we need to define P_{ij} for any $i, j \in \{0, 1, \ldots, M\}$.

Matrix representation

- To describe a Markov chain, we need to define P_{ij} for any $i, j \in \{0, 1, \ldots, M\}$.
- It is convenient to represent the collection of transition probabilities P_{ij} as a matrix:

$$A = \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0M} \\ P_{10} & P_{11} & \ldots & P_{1M} \\ . & & & \\ . & & & \\ . & & & \\ P_{M0} & P_{M1} & \ldots & P_{MM} \end{pmatrix}$$

Matrix representation

- To describe a Markov chain, we need to define P_{ij} for any $i, j \in \{0, 1, \ldots, M\}$.
- It is convenient to represent the collection of transition probabilities P_{ij} as a matrix:

$$A = \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0M} \\ P_{10} & P_{11} & \ldots & P_{1M} \\ . & & & \\ . & & & \\ . & & & \\ P_{M0} & P_{M1} & \ldots & P_{MM} \end{pmatrix}$$

- For this to make sense, we require $P_{ij} \geq 0$ for all i, j and $\sum_{j=0}^{M} P_{ij} = 1$ for each i. That is, the rows sum to one.

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.
- This means that the row vector

$$\pi = \begin{pmatrix} \pi_0 & \pi_1 & \dots & \pi_M \end{pmatrix}$$

is a left eigenvector of A with eigenvalue 1, i.e., $\pi A = \pi$.

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.
- This means that the row vector

$$\pi = \left(\begin{array}{cccc} \pi_0 & \pi_1 & \dots & \pi_M \end{array} \right)$$

  is a left eigenvector of A with eigenvalue 1, i.e., $\pi A = \pi$.
- We call π the *stationary distribution* of the Markov chain.

Ergodic Markov chains

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.
- This means that the row vector

$$\pi = \begin{pmatrix} \pi_0 & \pi_1 & \dots & \pi_M \end{pmatrix}$$

  is a left eigenvector of A with eigenvalue 1, i.e., $\pi A = \pi$.
- We call π the *stationary distribution* of the Markov chain.
- One can solve the system of linear equations $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ to compute the values π_j. Equivalent to considering A fixed and solving $\pi A = \pi$. Or solving $(A - I)\pi = 0$. This determines π up to a multiplicative constant, and fact that $\sum \pi_j = 1$ determines the constant.