Continuous random variables

Problems motivated by coin tossing

Random variable properties
Continuous random variables

Problems motivated by coin tossing

Random variable properties
Say $X$ is a **continuous random variable** if there exists a **probability density function** $f = f_X$ on $\mathbb{R}$ such that

$$P\{X \in B\} = \int_B f(x)dx := \int 1_B(x)f(x)dx.$$
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.
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

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

$$E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x) \, dx.$$
Expectations of continuous random variables

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

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

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

- Answer: \( E[X] = \int_{-\infty}^{\infty} f(x)x\,dx \).
Expectations of continuous random variables

- Recall that when $X$ was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
  \[ E[X] = \sum_{x : p(x) > 0} p(x)x. \]

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

- Answer: $E[X] = \int_{-\infty}^{\infty} f(x)x\,dx$.

- Recall that when $X$ was a discrete random variable, with $p(x) = P\{X = x\}$, we wrote
  \[ E[g(X)] = \sum_{x : p(x) > 0} p(x)g(x). \]
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: $E[X] = \int_{-\infty}^{\infty} f(x)x \, dx$.

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: $E[X] = \int_{-\infty}^{\infty} f(x)x dx$.  

- 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 $E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x) dx$.  

Suppose $X$ is a continuous random variable with mean $\mu$. 

We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.

Next, if $g = g_1 + g_2$ then $E[g(X)] = \int g_1(x)f(x)\,dx + \int g_2(x)f(x)\,dx = \int (g_1(x) + g_2(x))f(x)\,dx = E[g_1(X)] + E[g_2(X)]$.

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 $\text{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2]$ and use additivity of expectation to say that $\text{Var}[X] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.

This formula is often useful for calculations.
Suppose $X$ is a continuous random variable with mean $\mu$.
We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.
Suppose $X$ is a continuous random variable with mean $\mu$.

We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.

Next, if $g = g_1 + g_2$ then

$$E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int (g_1(x) + g_2(x))f(x)dx = E[g_1(X)] + E[g_2(X)].$$
Suppose $X$ is a continuous random variable with mean $\mu$.

We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.

Next, if $g = g_1 + g_2$ then
\[ E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int (g_1(x) + g_2(x))f(x)dx = E[g_1(X)] + E[g_2(X)]. \]

Furthermore, $E[ag(X)] = aE[g(X)]$ when $a$ is a constant.
Suppose $X$ is a continuous random variable with mean $\mu$.

We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.

Next, if $g = g_1 + g_2$ then

$$E[g(X)] = \int g_1(x)f(x)dx + \int g_2(x)f(x)dx = \int (g_1(x) + g_2(x))f(x)dx = E[g_1(X)] + E[g_2(X)].$$

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 $\text{Var}[X] = E[X^2] - 2\mu X + \mu^2$ and use additivity of expectation to say that

Suppose $X$ is a continuous random variable with mean $\mu$.

We can write $\text{Var}[X] = E[(X - \mu)^2]$, same as in the discrete case.

Next, if $g = g_1 + g_2$ then
\[
E[g(X)] = \int g_1(x)f(x)\,dx + \int g_2(x)f(x)\,dx = \\
\int (g_1(x) + g_2(x))f(x)\,dx = E[g_1(X)] + E[g_2(X)].
\]

Furthermore, $E[a\,g(X)] = aE[g(X)]$ when $a$ is a constant.

Just as in the discrete case, we can expand the variance expression as $\text{Var}[X] = E[X^2] - 2\mu X + \mu^2$ and use additivity of expectation to say that $\text{Var}[X] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$.

This formula is often useful for calculations.
Continuous random variables

Problems motivated by coin tossing

Random variable properties
Continuous random variables

Problems motivated by coin tossing

Random variable properties
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 $S_n = \sum_{i=1}^{n} X_i$. 

- Binomial: 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 $S_n - \mathbb{E}[S_n]$.

Here $\mathbb{E}[S_n] = np$ and $\text{SD}(S_n) = \sqrt{\text{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

- Poisson is limit of binomial as $n \to \infty$ when $p = \lambda/n$.

- Poisson point process: toss one $\lambda/n$ coin during each length $1/n$ time increment, take $n \to \infty$ limit.

- Exponential: time till first event in $\lambda$ Poisson point process.

- Gamma distribution: time till $n$th event in $\lambda$ Poisson point process.
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 $S_n = \sum_{i=1}^{n} X_i$.

- **Binomial** ($S_n$ — number of heads in $n$ tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain $n$ heads).
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 $S_n = \sum_{i=1}^{n} X_i$.

- **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 $\frac{S_n - E[S_n]}{\text{SD}(S_n)}$. Here $E[S_n] = np$ and $\text{SD}(S_n) = \sqrt{\text{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$. 
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 $S_n = \sum_{i=1}^{n} X_i$.

**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 $\frac{S_n - E[S_n]}{\text{SD}(S_n)}$. Here $E[S_n] = np$ and $\text{SD}(S_n) = \sqrt{\text{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

**Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$. 
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 $S_n = \sum_{i=1}^{n} X_i$.

**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 $\frac{S_n - E[S_n]}{SD(S_n)}$. Here $E[S_n] = np$ and $SD(S_n) = \sqrt{Var(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

**Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.

**Poisson point process**: toss one $\lambda/n$ coin during each length $1/n$ time increment, take $n \to \infty$ limit.
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 $S_n = \sum_{i=1}^{n} X_i$.

**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 $\frac{S_n - E[S_n]}{SD(S_n)}$. Here $E[S_n] = np$ and $SD(S_n) = \sqrt{Var(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

**Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.

**Poisson point process**: toss one $\lambda/n$ coin during each length $1/n$ time increment, take $n \to \infty$ limit.

**Exponential**: time till first event in $\lambda$ Poisson point process.
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 $\frac{S_n - E[S_n]}{SD(S_n)}$. Here $E[S_n] = np$ and $SD(S_n) = \sqrt{Var(S_n)} = \sqrt{npq}$ where $q = 1 - p$.

- **Poisson** is limit of binomial as $n \rightarrow \infty$ when $p = \lambda/n$.

- **Poisson point process**: toss one $\lambda/n$ coin during each length $1/n$ time increment, take $n \rightarrow \infty$ limit.

- **Exponential**: time till first event in $\lambda$ Poisson point process.

- **Gamma distribution**: time till $n$th event in $\lambda$ 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 $\lambda$ is gamma with parameters $(n, \lambda)$.
Continuous random variable properties derivable from coin toss intuition

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

- **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 $X$. 

- **Poisson random variable expectation** is $\lim_{n \to \infty} np = \lim_{n \to \infty} n\lambda = \lambda$.

- **Variance** is $\lim_{n \to \infty} np(1-p) = \lim_{n \to \infty} n(1-\lambda/n)\lambda/n = \lambda$.

- **Sum of $\lambda_1$ Poisson and independent $\lambda_2$ Poisson** is a $\lambda_1 + \lambda_2$ Poisson.

- **Times between successive events** in $\lambda$ Poisson process are independent exponentials with parameter $\lambda$.

- **Minimum of independent exponentials** with parameters $\lambda_1$ and $\lambda_2$ is itself exponential with parameter $\lambda_1 + \lambda_2$. 
Continuous random variable properties derivable from coin toss intuition

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

- **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 \( \lambda \) is gamma with parameters \((n, \lambda)\).

- **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 \( \lambda_1 \) Poisson and independent \( \lambda_2 \) 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 \( \lambda \) is gamma with parameters \((n, \lambda)\).

- **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** \( \lambda_1 \) **Poisson and independent** \( \lambda_2 \) **Poisson** is a \( \lambda_1 + \lambda_2 \) Poisson.

- **Times between successive events** in \( \lambda \) Poisson process are independent exponentials with parameter \( \lambda \).
Continuous random variable properties derivable from coin toss intuition

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

- **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** $\lambda_1$ Poisson and independent $\lambda_2$ Poisson is a $\lambda_1 + \lambda_2$ Poisson.

- **Times between successive events** in $\lambda$ Poisson process are independent exponentials with parameter $\lambda$.

- **Minimum of independent exponentials** with parameters $\lambda_1$ and $\lambda_2$ is itself exponential with parameter $\lambda_1 + \lambda_2$. 
DeMoivre-Laplace limit theorem (special case of central limit theorem):

$$\lim_{n \to \infty} P\{a \leq \frac{S_n - np}{\sqrt{npq}} \leq b\} \to \Phi(b) - \Phi(a).$$
DeMoivre-Laplace Limit Theorem

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

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

- This is \(\Phi(b) - \Phi(a) = P\{a \leq X \leq b\}\) 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.
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 \( 1 - \Phi(2) = \Phi(-2) \).
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?
Toss a million fair coins. Approximate the probability that I get more than 501,000 heads.

Answer: well, \(\sqrt{npq} = \sqrt{10^6 \times 0.5 \times 0.5} = 500\). 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, \( \sqrt{npq} = \sqrt{10^6 \times 0.5 \times 0.5} = 500 \). 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

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$
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.
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 $\sigma^2$ and expectation $\mu$. 

Values: $\Phi(-3) \approx 0.0013$, $\Phi(-2) \approx 0.023$ and $\Phi(-1) \approx 0.159$. 

Rule of thumb: “two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean.”
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 $\sigma^2$ and expectation $\mu$.

$Y$ is said to be normal with parameters $\mu$ and $\sigma^2$. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$.
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 $\sigma^2$ and expectation $\mu$.

$Y$ is said to be normal with parameters $\mu$ and $\sigma^2$. 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: $\Phi(-3) \approx 0.0013$, $\Phi(-2) \approx 0.023$ and $\Phi(-1) \approx 0.159$. Rule of thumb: “two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean.”
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 $\sigma^2$ and expectation $\mu$.
- $Y$ is said to be normal with parameters $\mu$ and $\sigma^2$. 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.”
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 \( \sigma^2 \) and expectation \( \mu \).
- \( Y \) is said to be normal with parameters \( \mu \) and \( \sigma^2 \). 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.”
Say \( X \) is an exponential random variable of parameter \( \lambda \) 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** $\lambda$ 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

$$F_X(a) = \int_0^a f(x)\,dx = \int_0^a \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\bigg|_0^a = 1 - e^{-\lambda a}.$$
Say $X$ is an **exponential random variable of parameter** $\lambda$ 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

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

Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$. 
Say $X$ is an exponential random variable of parameter $\lambda$ 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

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

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.
Say $X$ is an **exponential random variable of parameter** $\lambda$ 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

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

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 $E[X^n] = n!/\lambda^n$. 

Say $X$ is an **exponential random variable** of parameter $\lambda$ 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

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

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

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 $\Gamma$. 

Waiting time interpretation makes sense only for integer $\alpha$, but distribution is defined for general positive $\alpha$. 

---

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

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 $\Gamma$. 

Waiting time interpretation makes sense only for integer $\alpha$, but distribution is defined for general positive $\alpha$. 
Say that random variable $X$ has gamma distribution with parameters $(\alpha, \lambda)$ 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 $\Gamma$.

Waiting time interpretation makes sense only for integer $\alpha$, but distribution is defined for general positive $\alpha$. 
Outline

Continuous random variables

Problems motivated by coin tossing

Random variable properties
Continuous random variables

Problems motivated by coin tossing

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

And \( \text{Var}[X] = \text{Var}[(\beta - \alpha)Y + \alpha] = \text{Var}[(\beta - \alpha)Y] = (\beta - \alpha)^2 \text{Var}[Y] = (\beta - \alpha)^2 / 12. \)
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 \( P\{Y \leq 27\} \)?
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 \( P\{Y \leq 27\} = P\{X \leq 3\} = F_X(3) \).
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 $F_Y(a) = P\{Y \leq a\} = P\{X \leq a^{1/3}\} = F_X(a^{1/3})$.
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 $F_Y(a) = F_X(g^{-1}(a))$. 

Distribution of function of random variable
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.
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?
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 the distribution of $X$ (ignoring $Y$) and the distribution of $Y$ (ignoring $X$) the marginal distributions.
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, $P\{Y = j\} = \sum_{i=1}^{n} A_{i,j}$. 
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.
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 column 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 $p(x, y) = p_{X,Y}(x, y) = P\{X = x, Y = y\}$. 
Given random variables $X$ and $Y$, define

$$F(a, b) = P\{X \leq a, Y \leq b\}.$$
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 $(a, b)$. 

---

Density:

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$
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.
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 \( F_Y \)?
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 \( F_Y(b) = \lim_{a \to \infty} F(a, b) \).
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: $f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y)$.
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.
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
\[
f(x, y) = f_X(x)f_Y(y).
\]
Say we have independent random variables $X$ and $Y$ and we know their density functions $f_X$ and $f_Y$. 

This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus, 

$$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y) \, dx \, dy = \int_{-\infty}^{\infty} F_X(a-y)f_Y(y) \, dy.$$ 

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 $a$. 


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,

$$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y) \, dx \, dy = \int_{-\infty}^{\infty} F_X(a-y)f_Y(y) \, dy.$$

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 $a$. 
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,

$$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y) \, dx \, dy$$

$$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y) \, dy.$$
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,

$$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\,dy.$$

- Differentiating both sides gives

$$f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)\,dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)\,dy.$$
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 $f(x, y)$. 

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).
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).
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).
Suppose I choose $n$ random variables $X_1, X_2, \ldots, X_n$ uniformly at random on $[0, 1]$, independently of each other.
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 $[0, 1]^n$. 
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 $a$?
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 \textit{largest} of the \( X_i \) is less than \( a \)?

\textbf{ANSWER:} \( a^n \).
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 \( X \)?
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: \( F_X(a) = \begin{cases} 
0 & a < 0 \\
Na^n & a \in [0, 1]. \\
1 & a > 1
\end{cases} \)

\[ f_X(a) = F_X'(a) = na^{n-1}. \]
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 a 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: 

$$f(x_1, x_2, \ldots, x_n) = \frac{n!}{n^n} \prod_{i=1}^{n} f(x_i) \text{ if } x_1 < x_2 < \ldots < x_n,$$

zero otherwise.

Let $\sigma: \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$.

Are $\sigma$ and the vector $(Y_1, \ldots, Y_n)$ independent of each other? Yes.
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 $X_j$. 

- What is the joint probability density of the $Y_i$?
  
  **Answer:** $f(x_1, x_2, \ldots, x_n) = \frac{n!}{\prod_{i=1}^{n} f(x_i)}$ if $x_1 < x_2 \ldots < x_n$, zero otherwise.

- Let $\sigma: \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$.

- Are $\sigma$ and the vector $(Y_1, \ldots, Y_n)$ independent of each other?
  
  **Yes.**
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.
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 $X_1, X_2, \ldots, X_n$ 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.
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 $\sigma : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
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.
- 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 $\sigma : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are $\sigma$ and the vector $(Y_1, \ldots, Y_n)$ 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 $\sigma$ and the vector $(Y_1, \ldots, Y_n)$ independent of each other?
- Yes.
Several properties we derived for discrete expectations continue to hold in the continuum.
Several properties we derived for discrete expectations continue to hold in the continuum.

If $X$ is discrete with mass function $p(x)$ then

$$E[X] = \sum_x p(x)x.$$
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
$$E[X] = \int f(x)x\,dx.$$
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

$$E[g(x)] = \sum_x p(x)g(x).$$
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, if $X$ 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
  
  \[ E[X] = \int f(x)x\,dx. \]

▶ If $X$ is discrete with mass function $p(x)$ then
  
  \[ E[g(X)] = \sum_x p(x)g(x). \]

▶ Similarly, if $X$ 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, if $X$ 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 
  
  $$E[g(X, Y)] = \int_\infty \int_\infty g(x, y)f(x, y)dxdy.$$  

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 $E[\sum a_iX_i] = \sum a_iE[X_i]$.
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 $E[\sum a_iX_i] = \sum a_iE[X_i]$.
- 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 $E[\sum a_iX_i] = \sum a_iE[X_i]$.
- 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 $E[\sum a_iX_i] = \sum a_iE[X_i]$.

- 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 $E[\sum a_iX_i] = \sum a_iE[X_i]$.
- 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 $1 - F_X$. 
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
  \[
  E[g(X)h(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x, y) \, dx \, dy.
  \]
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
  \[ \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx = E[h(Y)]E[g(X)]. \]
Now define covariance of $X$ and $Y$ by
\[
\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])].
\]
Now define covariance of $X$ and $Y$ by
\[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \]
Note: by definition $\text{Var}(X) = \text{Cov}(X, X)$. 
Now define covariance of $X$ and $Y$ by
\[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \]

Note: by definition $\text{Var}(X) = \text{Cov}(X, X)$.

Now define covariance of $X$ and $Y$ by
\[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \]
Note: by definition $\text{Var}(X) = \text{Cov}(X, X)$.
If $X$ and $Y$ are independent then $\text{Cov}(X, Y) = 0$. 
Now define covariance of $X$ and $Y$ by
\[ \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \]
Note: by definition \( \text{Var}(X) = \text{Cov}(X, X) \).
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 \( \text{Cov}(X, Y) = 0 \).
Converse is not true.
Basic covariance facts

- \( \text{Cov}(X, Y) = \text{Cov}(Y, X) \)
Basic covariance facts

▶ $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
▶ $\text{Cov}(X, X) = \text{Var}(X)$
Basic covariance facts

- \( \text{Cov}(X, Y) = \text{Cov}(Y, X) \)
- \( \text{Cov}(X, X) = \text{Var}(X) \)
- \( \text{Cov}(aX, Y) = a\text{Cov}(X, Y) \).
Basic covariance facts

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

Basic covariance facts

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

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j).$$
Basic covariance facts

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

\[
\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j).
\]

- Special case:

\[
\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{(i, j): i < j} \text{Cov}(X_i, X_j).
\]

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 $\rho(aX + b, X) = -1$. 

Defining correlation

- Correlation of $X$ and $Y$ defined by

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

**Correlation** of \( X \) and \( Y \) defined by

\[
\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.
\]

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

- **Correlation** of $X$ and $Y$ defined by
  $$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$ 

- 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 $-1 \leq \rho(X, Y) \leq 1$. 
Defining correlation

- **Correlation** of $X$ and $Y$ defined by

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

- 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 $\rho(aX + b, X) = 1$. 
Defining correlation

- **Correlation** of $X$ and $Y$ defined by

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

- 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 $\rho(aX + b, X) = -1$. 

It all starts with the definition of conditional probability:
\[ P(A|B) = \frac{P(AB)}{P(B)}. \]
Conditional probability distributions

- It all starts with the definition of conditional probability:
  \[ P(A|B) = \frac{P(AB)}{P(B)} \]

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

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
\[ p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x,y)}{p_Y(y)}. \]
It all starts with the definition of conditional probability:
\[ P(A|B) = \frac{P(AB)}{P(B)}. \]

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: \( P(A|B) = \frac{P(AB)}{P(B)} \).
- 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 \( f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \).
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 \( Y = y \).
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 $Y = y$. 

In continuum setting, we have $f_{X|Y}(x|y) = f(x, y) / f_Y(y)$. So $E[X|Y = y] = \int_{-\infty}^{\infty} x f(x, y) f_Y(y) dx$. 

Can make sense of this in the continuum setting as well.
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

$$E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x xP_{X|Y}(x|y).$$
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.
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

$$E[X|Y = y] = \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} dx$$
Can think of $E[X|Y]$ as a function of the random variable $Y$. When $Y = y$ it takes the value $E[X|Y = y]$. 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: $E[X|Y = y] = \sum x x P\{X = x|Y = y\} = \sum x x p(x, y) p_Y(y)$. Recall that, in general, $E[g(Y)] = \sum y p_Y(y) g(y)$. $E[E[X|Y]] = \sum y p_Y(y) \sum x x p(x, y) p_Y(y) = \sum x \sum y p(x, y) x = E[X]$. 
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 $Y$. 

Proof in discrete case:

$E[X|Y = y] = \sum_x x p_{X|Y}(x|y) = \sum_x x p(x,y) p_Y(y)$.

Recall that, in general, $E[g(Y)] = \sum_y g(y) p_Y(y)$. 

$E[E[X|Y = y]] = \sum_y p_Y(y) \sum_x x p(x,y) p_Y(y) = \sum_x x \sum_y p(x,y) p_Y(y) = E[X]$.
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 $E[E[X|Y]]$?
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.
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, $E[g(Y)] = \sum_y p_Y(y)g(y)$. 

$E[E[X|Y]] = \sum_y p_Y(y)E[X|Y = y] = \sum_y p_Y(y)\sum_x x\frac{p(x,y)}{p_Y(y)} = \sum_x xE[X]$. 

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:

$E[X|Y = y] = \sum_x x P\{X = x|Y = y\} = \sum_x x \frac{p(x,y)}{p_Y(y)}$.

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:
  
  $E[X|Y = y] = \sum_x x P\{X = x|Y = y\} = \sum_x x \frac{p(x,y)}{p_Y(y)}$.

- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.

- $E[E[X|Y = y]] = \sum_y p_Y(y) \sum_x x \frac{p(x,y)}{p_Y(y)} = \sum_x \sum_y p(x, y) x = E[X]$. 
Definition:
\[
\text{Var}(X \mid Y) = E[(X - E[X \mid Y])^2 \mid Y] = E[X^2 - E[X \mid Y]^2 \mid Y].
\]
Conditional variance

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

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

\[ \text{Var}(X \mid Y) \] is a random variable that depends on \( Y \). It is the variance of \( X \) in the conditional distribution for \( X \) given \( Y \).

Note \( E[\text{Var}(X \mid Y)] = E[E[X^2 \mid Y]] - E[E[X \mid Y]^2 \mid Y] = E[X^2] - E[E[X \mid Y]^2]. \)
Definition: \( \text{Var}(X|Y) = E[(X - E[X|Y])^2|Y] = E[X^2 - E[X|Y]^2|Y] \).

\( \text{Var}(X|Y) \) is a random variable that depends on \( Y \). It is the variance of \( X \) in the conditional distribution for \( X \) given \( Y \).


If we subtract \( E[X]^2 \) from first term and add equivalent value \( E[E[X|Y]]^2 \) to the second, RHS becomes \( \text{Var}[X] - \text{Var}[E[X|Y]] \), which implies following:
Definition:
\[ \text{Var}(X|Y) = E[(X - E[X|Y])^2|Y] = E[X^2 - E[X|Y]^2|Y]. \]

\text{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:

If we subtract \( E[X]^2 \) from first term and add equivalent value \( E[E[X|Y]]^2 \) to the second, RHS becomes \( \text{Var}[X] - \text{Var}[E[X|Y]] \), which implies following:

**Useful fact:** \( \text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)]. \)
Definition: 
\[ \text{Var}(X|Y) = E[(X - E[X|Y])^2|Y] = E[X^2 - E[X|Y]^2|Y]. \]

\( \text{Var}(X|Y) \) is a random variable that depends on \( Y \). It is the variance of \( X \) in the conditional distribution for \( X \) given \( Y \).

Note \( E[\text{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]. \)

If we subtract \( E[X]^2 \) from first term and add equivalent value \( E[E[X|Y]]^2 \) to the second, RHS becomes \( \text{Var}[X] - \text{Var}[E[X|Y]] \), which implies following:

**Useful fact:** \( \text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)]. \)

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

\[ \text{Var}(X|Y) \] is a random variable that depends on \( Y \). It is the variance of \( X \) in the conditional distribution for \( X \) given \( Y \).

Note \( E[\text{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]. \)

If we subtract \( E[X]^2 \) from first term and add equivalent value \( E[E[X|Y]]^2 \) to the second, RHS becomes \( \text{Var}[X] - \text{Var}[E[X|Y]] \), which implies following:

**Useful fact:** \( \text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)]. \)

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

What are the covariances $\text{Cov}(X, Y)$ and $\text{Cov}(X, Z)$? 

How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$? 

What is $E[Z \mid X]$? And how about $\text{Var}(Z \mid X)$? 

Both of these values are functions of $X$. Former is just $X$. Latter happens to be a constant-valued function of $X$, i.e., it happens not to actually depend on $X$. We have $\text{Var}(Z \mid X) = \sigma_Y^2$. 

Can we check the formula $\text{Var}(Z) = \text{Var}(E[Z \mid X]) + E[\text{Var}(Z \mid X)]$ in this case?
Example

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

What are the covariances $\text{Cov}(X, Y)$ and $\text{Cov}(X, Z)$?
Example

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

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

What are the covariances $\text{Cov}(X, Y)$ and $\text{Cov}(X, Z)$?

How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?

What is $E[Z|X]$? And how about $\text{Var}(Z|X)$?
Let $X$ be a random variable of variance $\sigma_X^2$ and $Y$ an independent random variable of variance $\sigma_Y^2$ and write $Z = X + Y$. Assume $E[X] = E[Y] = 0$.

What are the covariances $\text{Cov}(X, Y)$ and $\text{Cov}(X, Z)$?

How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?

What is $E[Z|X]$? And how about $\text{Var}(Z|X)$?

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 $\text{Var}(Z|X) = \sigma_Y^2$. Can we check the formula $\text{Var}(Z) = \text{Var}(E[Z|X]) + E[\text{Var}(Z|X)]$ in this case?
Example

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

What are the covariances $\text{Cov}(X, Y)$ and $\text{Cov}(X, Z)$?

How about the correlation coefficients $\rho(X, Y)$ and $\rho(X, Z)$?

What is $E[Z|X]$? And how about $\text{Var}(Z|X)$?

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 $\text{Var}(Z|X) = \sigma_Y^2$.

Can we check the formula $\text{Var}(Z) = \text{Var}(E[Z|X]) + E[\text{Var}(Z|X)]$ in this case?
Let $X$ be a random variable and $M(t) = E[e^{tX}]$. 

By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all $t$. In other words, adding independent random variables corresponds to multiplying moment generating functions.
Let $X$ be a random variable and $M(t) = E[e^{tX}]$.

Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, $n$th 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

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

Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, $n$th 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 $M_Z(t) = E[e^{tZ}]$. In other words, adding independent random variables corresponds to multiplying 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, $n$th 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 $M_Z$?
Let $X$ be a random variable and $M(t) = E[e^{tX}]$.

Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, $n$th 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

- Let $X$ be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, $n$th 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.
We showed that if $Z = X + Y$ and $X$ and $Y$ are independent, then $M_Z(t) = M_X(t)M_Y(t)$. 

If $Z = aX$ then $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$. 

If $Z = X + b$ then $M_Z(t) = E[e^{tZ}] = E[e^{tX+b}] = e^{bt}M_X(t)$. 

This is 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 $X_1 \ldots X_n$ are i.i.d. copies of $X$ and $Z = X_1 + \ldots + X_n$ 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 $X_1 \ldots X_n$ 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.
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 is 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 $X_1 \ldots X_n$ 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 $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$. 
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: $M^n_X$. 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 $\mu$, variance $\sigma^2$, 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 $\mu$, variance $\sigma^2$, then $M_X(t) = e^{\sigma^2 t^2/2 + \mu t}$.
- If $X$ is exponential with parameter $\lambda > 0$ then $M_X(t) = \frac{\lambda}{\lambda - t}$.
A standard **Cauchy random variable** is a random real number with probability density \( f(x) = \frac{1}{\pi} \frac{1}{1+x^2} \).
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.
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.

\[ F_X(x) = P\{X \leq x\} = P\{\tan \theta \leq x\} = P\{\theta \leq \tan^{-1} x\} = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x. \]
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 \( f_X(x) = \frac{d}{dx} F(x) = \frac{1}{\pi} \frac{1}{1+x^2} \).
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).
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 $\beta$ 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 $\beta$ 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 $B(a,b) =$ constant chosen to make integral one. Can show $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. 

Turns out that $E[X] = a/(a+b)$ and the mode of $X$ is $(a-1)/(a+b)$. 
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 $\beta$ 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 $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. 
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 $\beta$ 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 $\frac{(a-1)}{(a-1)+(b-1)}$. 