STAT 210 - PROBABILITY THEORY

GREG TAM

Contents

1. September 3rd, 2013

- Unification: prob & stats
- Conditioning: discrete & continuous
- Stories, Representation, Characterization

1.1. **Representation.**

$$\underbrace{X}_{\text{Expo}}^{\beta} \sim \text{Weibull}$$

1.2. **Characterization.** Suppose that X, Y are i.i.d. random variables with $X + Y \perp\!\!\!\perp X - Y$. Then

$$\Rightarrow X, Y \sim \text{Normal}$$

Consider a Cauchy distribution with

$$f(x) = \frac{1}{1+x^2}\frac{1}{\pi} \qquad -\infty < x < \infty$$

Now consider $y = \frac{1}{x}$. This is also Cauchy. We can prove this the long way with Jacobians, but it becomes extremely tedious. Now by representation, we say that by definition, something is Cauchy if

$$Y = \frac{Z_1}{Z_2}$$

if Z_1, Z_2 are independent $\mathcal{N}(0,1)$.

Let $C \sim$Cauchy. What would the distribution be of

$$\frac{1+C}{1-C}?$$

The representation of the Cauchy distribution means that

$$C \sim \frac{Z_1}{Z_2}$$

where $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

$$\begin{aligned}\frac{1+\frac{Z_1}{Z_2}}{1-\frac{Z_1}{Z_2}} &= \frac{Z_1+Z_2}{Z_2-Z_1}\\ &\sim \frac{\sqrt{2}Z_3}{\sqrt{2}Z_4}\\ &\sim C\end{aligned}$$

1.3. **"Nonmeasurable Set".**

2. September 5th, 2013

Definition 1 (Sigma Algebra)**.** *Let Ω be a non-empty set (sample space). Let $\mathcal{F} \subseteq 2^\Omega$. $\mathcal{F}$ is a sigma algebra if*

(1) $\emptyset \in \mathcal{F}$
(2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
(3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^\infty A_n \in \mathcal{F}$

Definition 2. *$(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space if $\mathcal{F}$ is a sigma algebra of subsets of Ω and it*

$$\mathcal{P} : \mathcal{F} \to [0,1]$$

such that

(1) $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$
(2) If $A_1, A_2, \ldots$ be disjoint events in $\mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty \mathbb{P}(A_n)$$

Why not always choose $\mathcal{F} = 2^\Omega$? Consider the finite case $\Omega = \{\omega_1, \ldots, \omega_n\}$. We have

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i$$

where $p_i = \mathbb{P}(\{\omega_i\})$

Example 1. *Let $\Omega = [0,1]$. Suppose we want a Uniform distribution (i.e. probability = length) and want*

$$\mathbb{P}(A \oplus C) = \mathbb{P}(A)$$

where C is a constant and $\oplus$ is addition with wraparound. Then it is impossible to have $\mathcal{F} = 2^\Omega$ (assuming axiom of choice).

We have

$$A \times B = \{(a,b), a \in A, b \in B\}$$

and

$$\prod_{i \in I} A_i = \{(a_i, i \in I)\}$$

Definition 3 (Axiom of Choice)**.** *Cartesian product of non-empty sets is non-empty*

Let $\mathcal{A} \subseteq 2^\Omega$. Then there is a unique smallest σ-algebra containing $\mathcal{A}$ (call it $\sigma(\mathcal{A})$).

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{C}\ \sigma\text{-algebra containing } \mathcal{A}} \mathcal{C}$$

Example 2 (Examples of sigma algebras)**.**

(0) $\{\emptyset, \Omega\}$
(1) Suppose we have a partition A, B, C, D. We look at the sigma algebra generated by this partition, which would be

$$\{\emptyset, A, B, C, D, A \cup B, A \cup B \cup C, A^c \cup B^c, \ldots, \Omega\}$$

(2) Let $\Omega = \mathbb{R}$. Define the Borel σ-algebra $\mathfrak{B}$. Let $\mathfrak{B}_0 = \{$open intervals$\}$. This is clearly not a sigma algebra since the union of two disjoint intervals is not always an interval. Let $\mathfrak{B} = \sigma(\mathfrak{B}_0)$. Let $\mathfrak{B}_1 = \{$countable unions of elements of $\mathfrak{B}_1$ countable intersections of elements of $\mathfrak{B}_0$, complements and such $\}$. Let $\mathfrak{B}_n = \{$countable unions of elements of $\mathfrak{B}_n$ countable intersections of elements of $\mathfrak{B}_{n-1}$, complements and such $\}$. Finally, we have

$$\mathfrak{B}_\infty = \bigcup_{n=1}^\infty \mathfrak{B}_n$$

which is still not a σ-algebra.

$$\mathfrak{B}_{\infty+1}, \mathfrak{B}_{\infty+2}, \ldots$$

Definition 4. *$X : \Omega \to \mathbb{R}$ is a random variable if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathfrak{B}$.*

$$\begin{aligned} X^{-1}(B) &= \{\omega \in \Omega : X(\omega) \in B\} \\ &= \{X \in B\} \end{aligned}$$

Definition 5 (Distribution)**.** *The distribution (law) of a random variable X is the function $\mathcal{L}$,*

$$\mathcal{L}(B) = \mathbb{P}(X \in B) \qquad B \in \mathfrak{B}$$

Theorem 1. *If $X \sim Y$, then $g(X) \sim g(Y)$ for any measurable function g.*

Proof.

$$\begin{aligned} \mathbb{P}(g(X) \in B) &= \mathbb{P}\big(X \in g^{-1}(B)\big) \\ &= \mathbb{P}\big(Y \in g^{-1}(B)\big) \\ &= \mathbb{P}(g(Y) \in B) \end{aligned}$$

□

3. September 10th, 2013

Sections are on Wednesday 4-5 and 5-6 in room 304.
Joint Distribution of $X_1, \ldots, X_n$:
$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n)$ where $B_i \in \mathfrak{B}$.
Joint CDF:
$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = F(x_1, \ldots, x_n)$

3.1. Independence.

Definition 6. *$X_1, X_2, \ldots, X_n$ are independent if*

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdot \ldots \cdot \mathbb{P}(X_n \in B_n)$$

for all $B_1, B_2, \ldots, B_n$ Borel. For infinitely many X_j's, then the definition of independence says any finite subset are independent.

3.2. Independence Lemma.

If $X \perp\!\!\!\perp Y$, $g(X) \perp\!\!\!\perp h(X)$, then $g(X), h(X), Y$ are fully independent.

Proof.

$$\begin{aligned} &\mathbb{P}(g(X) \in A, h(X) \in B, Y \in C) \\ &= \mathbb{P}\big(X \in g^{-1}(A), X \in h^{-1}(B), Y \in C\big) \\ &= \mathbb{P}\big(X \in g^{-1}(A) \cap H^{-1}(B), Y \in C\big) \\ &= \mathbb{P}\big(X \in g^{-1}(A) \cap H^{-1}(B)\big)\, \mathbb{P}(Y \in C) \qquad \text{since } X \perp\!\!\!\perp Y \\ &= \mathbb{P}(g(X) \in A)\, \mathbb{P}(h(X) \in B)\, \mathbb{P}(Y \in C) \end{aligned}$$

□

Recall that for $X : \Omega \to \mathbb{R}$ to be a random variable, we require $\{X \in B\} = X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathfrak{B}$. It suffices to show that $\{X \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. Let

$$\mathcal{A} = \{B \in \mathbb{R} : X^{-1}(B) \in \mathcal{F}\}$$

Let $f : A \to B$ be a function. Then $f^{-1}\left(\bigcup_\alpha B_\alpha\right) = \bigcup_\alpha f^{-1}\left(B_\alpha\right)$ and it is similar for the intersections. For complements, $f^{-1}(B^c) = (f^{-1}(B))^c$. What this all implies is that $\mathcal{A}$ is a σ-algebra containing $(-\infty, x]$ for all x. This implies $\mathcal{A} = \mathfrak{B}$.

3.3. π-system.

$\mathcal{S} \subseteq 2^\Omega$ is a π-system if $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$.

Example 3. *$\{(-\infty, a] : a \in \mathbb{R}\}$ is an example of a π-system since intersecting two intervals of that form, leaves another interval of that same form.*

3.4. **λ-system.** Let $\mathcal{L} \subseteq 2^\Omega$.

- $\Omega \in sL$
- $A, B \in \mathcal{L}, A \subseteq B \Rightarrow B \setminus A \in \mathcal{L}$
- $A_1 \subseteq A_2 \subseteq \ldots \in \mathcal{L} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$

Theorem 2 (Dynkin's π-λ)**.** *If $\mathcal{S}$ is a π-system and $\mathcal{L}$ be a λ-system with $\mathcal{S} \subseteq \mathcal{L}$, then $\sigma(\mathcal{S}) \subseteq \mathcal{L}$.*

Before proving this, we must look at two facts.

- If $\mathcal{S}$ is both a π and λ system, then it is a σ-algebra.
- Any intersection of λ-systems is a λ-system.

Proof. WELOG (without essential loss of generality), assume $\mathcal{L}$ is the smallest λ-system containing $\mathcal{S}$. Then our goal is to show $\mathcal{L}$ is a π-system. We need to show

$$A \in \mathcal{L}, B \in \mathcal{L} \Rightarrow A \cap B \in \mathcal{L}$$

(this seems hard) Let us first look at a smaller example.

$$A \in \mathcal{S}, B \in \mathcal{S} \stackrel{?}{\Rightarrow} A \cap B \in \mathcal{S}$$

but this is obvious since $\mathcal{S}$ is a π-system. Now let us try and prove

$$A \in \mathcal{S}, B \in \mathcal{L} \Rightarrow A \cap B \in \mathcal{L}$$

Let

$$\mathcal{L}(A_0) = \{B \in \mathcal{L} : A_0 \cap B \in \mathcal{L}\}$$

for any $A_0 \in \mathcal{L}$. It is easy to see that this is a λ-system.
Case 1: $A_0 \in \mathcal{S}$
Then we know that $\mathcal{L}(A_0) \supseteq \mathcal{S}$, so $\mathcal{L}(A_0) = \mathcal{L}$. This proves that $A_0 \cap B \in$ for all $A_0 \in \mathcal{S}, B \in \mathcal{L}$.
Case 2: $A_0 \in \mathcal{L}$
This implies $\mathcal{L}(A_0) = \mathcal{L}$. □

Example 4. *Let P, Q be probability measures on $(\Omega, \mathcal{F})$. Then $\{A \in \mathcal{F} : P(A) = Q(A)\}$ is a λ-system. So if we can show $P(A) = Q(A)$ for all A in a π-system $\mathcal{S}$, then by π-λ, then $P(A) = Q(A)$ for all $A \in \sigma(\mathcal{S})$.*

Theorem 3. *The CDF F, $F(x) = \mathbb{P}(X \leq x)$ completely determines the distribution.*

Proof. Suppose P_1, P_2 are distributions that are compatible with F.

$$P_1(X \in (-\infty, a]) = P_2(X \in (-\infty, a])$$

for all a. That is, P_1, P_2 agree on the π-system $\mathcal{S} = \{(-\infty, a] : a \in \mathbb{R}\}$. By the π-λ theorem, then they agree on $\sigma(\mathcal{S})$. □

Exercise. *Assume*

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

Show

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

Hint: start by fixing $A = (-\infty, a]$.

4. September 12th, 2013

Office Hours:
Carl Morris: Thursday, 2:35-3:55pm
Peng: Wednesday, 10:00-11:00am

4.1. **Representation.** X is a real-valued random variable.

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

is the CDF (cumulative distribution function).

- $F(x) \uparrow$ that is F is an increasing function.
- $0 \leq F(X) \leq 1$
- Right continuous

4.2. **PIT (Probability Integral Transform).** Let $F(x)$ be continuous.

Theorem 4 (3.4.6, 3.4.7)**.** *If F is continuous and $X \sim CDF = F$, then*

$$U = F(X) \sim \mathcal{U}[0,1]$$

Proof. For $0 \le u \le 1$

$$\begin{aligned}
\mathbb{P}(U \le u) &= \mathbb{P}(F(X) \le u) \\
&= \mathbb{P}\big(F^{-1}(F(X)) \le F^{-1}(u)\big) \quad \text{since } F^{-1} \text{ is increasing} \\
&= \mathbb{P}\big(X \le F^{-1}(u)\big) \\
&= F(F^{-1}(u)) \\
&= u
\end{aligned}$$

□

Is

$$F(x) = 1 - e^{-x} = U$$

a CDF?

$$X = -\log(1-U)$$

is Exponential or

$$X = -\log(U)$$

is Exponential (since $1-U$ is also uniformly distributed)

The gamma function is defined as

$$\Gamma(a) \equiv \int_0^\infty x^a e^{-x} \left(\frac{\mathrm{d}x}{x}\right)$$

if $a > 0$. One interesting property is that

$$\Gamma(a+1) = a\Gamma(a)$$

We can prove this by integrating by parts.

$$\begin{aligned}
\Gamma(a+1) &= -\int_0^\infty x^a \,\mathrm{d}e^{-x} \\
&= 0 + \int_0^\infty e^x a x^{a-1} \,\mathrm{d}x
\end{aligned}$$

This is just a factorial, since

$$\begin{aligned}
\Gamma(1) &= 1 \\
\Gamma(n+1) &= n\Gamma(n) \\
&= n! \\
\Gamma\left(\frac{1}{2}\right) &= \sqrt{n}
\end{aligned}$$

Definition 7. *$\Gamma(a)$ has density*

$$f(x) = \begin{cases} \frac{x^{a-1}e^{-x}}{\Gamma(a)} \,\mathrm{d}x & x > 0 \\ 0 & x \le 0 \end{cases}$$

with $a > 0$ is a valid distribution. This integrates to 1.

Let $G_1 \sim \Gamma(a_1), G_2 \sim \Gamma(a_2)$ be independent. Let $T = G_1 + G_2$ and $B = \frac{G_1}{T}$. Then,

$$\begin{aligned}
T &\sim \Gamma(a_1 + a_2) \\
B &\sim \text{Beta}(a_1, a_2)
\end{aligned}$$

and are independent.

To go from $X, Y \to V, W$, where we want to go from $f(x,y)dxdy$ to $g(v,w)dvdw$, we need to write

$$\begin{aligned}
x &= x(v,w) \\
y &= y(v,w)
\end{aligned}$$

$$dxdy = \left|\frac{dxdy}{dvdw}\right| dvdw$$

Now note that

$$G_1 = TB$$
$$G_2 = T(1-B)$$

and we see that $T > 0$ and $0 \leq B \leq 1$.

$$\begin{aligned} f(g_1, g_2)dg_1, dg_2 &= \frac{g_1^{a_1-1} g_2^{a_2-1} e^{-g_1-g_2}}{\Gamma(a_1)\Gamma(a_2)} \left| \left(\frac{dg_1 dg_2}{dtdb} \right) \right| dtdb \\ &= \frac{g_1^{a_1-1} g_2^{a_2-1} e^{-t}}{\Gamma(a_1)\Gamma(a_2)} \left| \left(\frac{dg_1 dg_2}{dtdb} \right) \right| dtdb \\ &= \frac{b^{a_1-1} g_2^{a_2-1} e^{-t} t^{a_1-1}}{\Gamma(a_1)\Gamma(a_2)} \left| \left(\frac{dg_1 dg_2}{dtdb} \right) \right| dtdb \\ &= \frac{b^{a_1-1}(1-b)^{a_2-1} e^{-t} t^{a_1-1+a_2-1}}{\Gamma(a_1)\Gamma(a_2)} \left| \left(\frac{dg_1 dg_2}{dtdb} \right) \right| dtdb \end{aligned}$$

Now to compute the Jacobian, we have

$$\begin{aligned} J &= \begin{vmatrix} B & 1-B \\ T & -T \end{vmatrix} \\ &= T \end{aligned}$$

which gives us

$$\begin{aligned} f(g_1, g_2)dg_1, dg_2 &= \frac{b^{a_1-1}(1-b)^{a_2-1} e^{-t} t^{a_1-1+a_2-1+1}}{\Gamma(a_1)\Gamma(a_2)} dtdb \\ &= \Gamma(a_1+a_2) \frac{b^{a_1-1}(1-b)^{a_2-1}}{\Gamma(a_1)\Gamma(a_2)} \frac{e^{-t} t^{a_1-1+a_2}}{\Gamma(a_1+a_2)} dtdb \end{aligned}$$

which is valid for $0 < b < 1$, $0 < t$. Looking at this, we can see that $T \sim \Gamma(a_1 + a_2)$ and also we have that

$$\beta(a_1, a_2) \equiv \frac{\Gamma(a_1)\Gamma(a_2)}{\Gamma(a_1+a_2)}$$

which is the beta function, which is equivalently written as

$$\beta(a_1, a_2) = \int_0^1 b^{a_1-1}(1-b)^{a_2-1} \, \mathrm{d}b$$

So the beta distribution is defined has density

$$\frac{1}{\beta(a_1, a_2)} b^{a_1-1}(1-b)^{a_2-1} \, \mathrm{d}b$$

4.3. **Special Cases.** Suppose $a_1 = a_2 = 1$. Then

- $\text{Beta}(1, 1) \sim \mathcal{U}[0, 1]$
- $X \sim \Gamma(1)$ is Exponential
- If X_1, X_2 are independent Exponential, then

  $$U = \frac{X_1}{X_1 + X_2}$$

  is uniformly distributed and independent of $X_1 + X_2 \sim \Gamma(2)$.

4.4. **Location-Scale Parameters.**

- If we have $Z \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = \mu + \sigma Z$.
- For uniform distributions, we have

  $$\mathcal{U}[a, b] = a + (b - a)U$$

  where $U \sim \mathcal{U}[0, 1]$
- Let $G \sim \Gamma(a)$ and $G' = bG$.

$$\begin{aligned} f(x) &= \frac{d}{dx} \mathbb{P}(G' \leq x) \\ &= \frac{d}{dx} \mathbb{P}\Big(G \leq \frac{x}{b}\Big) \\ &= \frac{1}{b} f\left(\frac{x}{b}\right) \end{aligned}$$

So the density is

$$\left(\frac{x}{b}\right)^a e^{-x/b}\left(\frac{dx}{x}\right)$$

so $G \sim b\Gamma(a)$, which is a scaled density.

5. September 17th, 2013

If we take two gammas G_a,G_b which are independent with

$$G_a \sim \Gamma(a)$$
$$G_b \sim \Gamma(b)$$

We have

$$G_a + G_b \sim \Gamma(a+b)$$
$$\frac{G_a}{G_a+G_b} = B \sim \text{Beta}(a,b) \perp\!\!\!\perp G_a + G_b$$

We also have the property that $\Gamma(r), r > 0$ is "infinitely divisible"

Definition 8 (Infinitely divisible)**.** *A distribution is infinitely divisible if for all $n > 0$,*

$$X = X_1 + \ldots + X_n$$

with X_i i.i.d.

If $X \sim \Gamma(r)$, then $X_i \sim \Gamma\left(\frac{r}{n}\right)$.

Theorem 5. *We have $G_j \sim \Gamma(a_j)$ which are independent and $j = 1, 2, 3$. Let*

$$T_1 = G_1$$
$$T_2 = G_1 + G_2$$
$$T_3 = G_1 + G_2 + G_3$$

and

$$B_1 = \frac{T_1}{T_2} = \frac{G_1}{G_1+G_2} \perp\!\!\!\perp G_1 + G_2$$
$$B_2 = \frac{T_2}{T_3} = \frac{G_1+G_2}{G_1+G_2+G_3} \perp\!\!\!\perp T_3$$

Then (B_1, B_2, T_3) are independent and

$$B_1 \sim \text{Beta}(a_1, a_2)$$
$$B_2 \sim \text{Beta}(a_1+a_2, a_3)$$
$$T_3 \sim \Gamma(a_1+a_2+a_3)$$

Proof. We show this using the independence lemma. If $X \perp\!\!\!\perp Y$ and $g_1(X) \perp\!\!\!\perp g_2(X)$ then $g_1(X), g_2(x), Y$ are fully independent. Let $X = (G_1, G_2)$ and $Y = G_3$ and

$$g_1(X) = \frac{G_1}{G_1+G_2} \perp\!\!\!\perp g_2(X) = G_1 + G_2$$

This gives us that (B_1, G_1+G_2, G_3) are fully independent and we apply this again to B_1 and (G_2, G_3). □

$$\frac{T_1}{T_2} = B_1 \sim \text{Beta}(a_1, a_2)$$
$$\frac{T_2}{T_3} = B_2 \sim \text{Beta}(a_1+a_2, a_3)$$

which are independent. Now we have

$$B_1 B_2 = \frac{T_1}{T_3} \sim \text{Beta}(a_1, a_2+a_3)$$

What is $\mathbb{E}[B_1]$?

$$\mathbb{E}[B_1] = \mathbb{E}\left[\frac{G_1}{G_1+G_2}\right] \tag{1}$$
$$= \frac{\mathbb{E}[G_1]}{\mathbb{E}[G_1]+\mathbb{E}[G_2]} \tag{2}$$

Step (2) is **NOT** true in general. Weird right? To get this, we look at

$$\begin{aligned}\mathbb{E}[G_1] &= \mathbb{E}\left[\frac{G_1}{G_1+G_2}(G_1+G_2)\right] \\ &= \mathbb{E}[B_1]\,\mathbb{E}[G_1+G_2]\end{aligned}$$

and rearranging we get

$$\mathbb{E}[B_1] = \frac{\mathbb{E}[G_1]}{\mathbb{E}[G_1+G_2]} = \frac{\mathbb{E}[G_1]}{\mathbb{E}[G_1]+\mathbb{E}[G_2]}$$

$Z \sim \mathcal{N}(0,1)$ if the density function if

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sigma^2} \qquad -\infty < z < \infty$$

with

$$\text{CDF} = \Phi(z) = \int_{-\infty}^{z} \phi(z')\,\mathrm{d}z'$$

$$\begin{aligned}\text{logit}(\Phi(z)) &= 1.6z\sqrt{1+\frac{z^2}{10}} \\ &= \log\left(\frac{\Phi(z)}{\Phi(-z)}\right) \\ &\approx \frac{1}{2}z^2 \quad \text{as } z \to \infty\end{aligned}$$

Representation: $Z = \Phi^{-1}(U)$, but we cannot really get this inverse.

5.1. **Box-Muller Representation.** We cannot get one normal distribution from one uniform distribution

$$BOXMULLERDIAGRAM$$

Changing to polar coordinates, we let $R = \sqrt{Z_1^2 + Z_2^2}$, so our density would be

$$\frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2+z_2^2)}$$

To get Θ, we have $\Theta = 2\pi U_1$ and for R, we have

$$\begin{aligned}Z_1 &= R\cos(\Theta) \\ Z_2 &= R\sin(\Theta) \\ Z_1^2 + Z_2^2 &= R^2 = -2\log(U_2)\end{aligned}$$

which gives us

$$R = \sqrt{-2\log(U_2)}$$

Theorem 6. *Given $U_1 \perp\!\!\!\perp U_2$ and*

$$\begin{aligned}Z_1 &= R\cos(\Theta) \\ Z_2 &= R\sin(\Theta) \\ \Theta &= 2\pi U_1 \\ R &= \sqrt{-2\log(U_2)}\end{aligned}$$

we have Z_1, Z_2 are two i.i.d. $\mathcal{N}(0,1)$ variables

Now let $Z \sim \mathcal{N}(0,1)$ and $t > 0$. Consider

$$\begin{aligned}\mathbb{P}(Z^2 \le t) &= \mathbb{P}\left(-\sqrt{t} < Z < \sqrt{t}\right) \\ &= \Phi(\sqrt{t}) - \Phi(-\sqrt{t})\end{aligned}$$

We want the density function so we differentiate both sides.

$$\begin{aligned}\frac{d}{dt}\mathbb{P}(Z^2 \le t) &= \frac{d}{dt}\left[\Phi(\sqrt{t}) - \Phi(-\sqrt{t})\right] \\ &= \frac{1}{\sqrt{t}}\Phi(\sqrt{t}) \\ &= \frac{t}{\sqrt{2\pi t}} = e^{-\frac{1}{2}t} \\ &= \frac{1}{\Gamma(\frac{1}{2})}\left(\frac{t}{2}\right)^{1/2} e^{-t/2}\end{aligned}$$

which is the density of $2\Gamma(\frac{1}{2})$. So

$$Z^2 \sim 2\Gamma\left(\frac{1}{2}\right)$$

Definition 9 (Chi-Square Distribution)**.** *The definition by representation of a chi-square distribution with degrees of freedom n is*

$$\chi_n^2 \equiv Z_1^2 + \ldots + Z_n^2$$

where Z_i are i.i.d. $\mathcal{N}(0,1)$.

Theorem 7.

$$\chi_n^2 \sim 2\Gamma\left(\frac{n}{2}\right)$$

$$\frac{\chi_m^2}{\chi_m^2 + \chi_n^2} \sim \text{Beta}\left(\frac{m}{2}, \frac{n}{2}\right)$$

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n}$$

By representation, t-distribution is

$$t_n = \frac{Z}{\sqrt{\chi_n^2/n}}$$

where Z and χ_n^2 are independent. Equivalently,

$$2\frac{t_n^2}{n} = 2\frac{Z^2}{\chi_n^2}$$

6. September 19th, 2013

If $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$.

$$\begin{aligned}\bar{Y} &= \frac{1}{n}\sum_{i=1}^n Y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\ Z &= \frac{\bar{Y} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1) \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \frac{\chi_{n-1}^2}{n-1} \\ &= \hat{\sigma^2}\end{aligned}$$

and Z and S^2 are independent. We have that

$$\frac{\bar{Y} - \mu}{\hat{\sigma^2}} \approx \mathcal{N}(0,1)$$

We have exactly that

$$t_n = \frac{\bar{Y} - \mu}{\hat{\sigma^2}}$$

is the t distribution with $n-1$ degrees of freedom.

$$t_m = \frac{Z}{\sqrt{\chi_m^2/m}}$$

$Z \perp\!\!\!\perp \chi_m^2$ where $Z \sim \mathcal{N}(0,1)$.

Theorem 8. *t_m has density*

$$f(t) = \frac{c_n}{(1 + t^2/n)^{\frac{1+m}{2}}}$$

Definition 10. *X is symmetric about 0 if*

$$X \sim S \cdot X$$

or

$$X \sim S \cdot A$$

where (S, A) is independent and

$$\begin{aligned} S &= \begin{cases} +1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{cases} \\ &= 2\text{Bern}(1/2) - 1 \end{aligned}$$

Now let

$$t_n^2 = W$$

We have

$$\begin{aligned} W &= m\frac{Z^2}{\chi_m^2} \\ &= m\frac{G_{(1/2)}}{G_{(m/2)}} \\ \frac{W}{W+m} &= \frac{G_{(1/2)}}{G_{(1/2)} + G_{(m/2)}} \\ &\sim \text{Beta}\left(\frac{1}{2}, \frac{m}{2}\right) \end{aligned}$$

$$B \sim \text{Beta}(a, b)$$

at $x \in (0, 1)$. The density is given by

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{\beta(a, b)}$$

$$\begin{aligned} \frac{d}{dx}\,\text{logit}(x) &= \frac{d}{dx}\log\left(\frac{x}{1-x}\right) \\ &= \frac{1}{x} + \frac{1}{1-x} \\ &= \frac{1}{x(1-x)} \end{aligned}$$

Now, in our example we have

$$\frac{W}{W+m} = X \qquad \text{logit}(x) = \log\left(\frac{W}{m}\right)$$

and density

$$f(x) = \frac{x^{1/2}(1-x)^{m/2}}{\beta(1/2, m/2)}\mathrm{d}\,\text{logit}(x) = \frac{w^{1/2}m^{m/2}}{\beta(1/2, m/2)(w+m)^{\frac{1+m}{2}}}\frac{\mathrm{d}W}{W}$$

If we substitute $w = t^2$, this gives us

$$\frac{tm^{m/2}}{(y^2+m)^{\frac{1+m}{2}}}2\frac{\mathrm{d}t}{t} = \frac{m^{m/2}}{(y^2+m)^{\frac{1+m}{2}}}2\mathrm{d}t =$$

is the density of $|t_m|$. If $-\infty < t < \infty$, then

$$f_m(t) = \frac{m^{m/2}}{\beta(1/2, m/2)}\frac{1}{(t^2+m)^{\frac{1+m}{2}}}(\mathrm{d}t)$$

Example 5. *Set $m = 1$ and*

$$\begin{aligned} f_1(t) &= \frac{1}{\beta(1/2, m/2)(1+t^2)} \\ &= \frac{1}{\pi(1+t^2)} \qquad \textit{for } -\infty < t < \infty \\ \beta(1/2, 1/2) &= \frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} \\ &= \pi \\ F(t) &= \int_{-\infty}^{t} \frac{1}{\pi} \frac{1}{1+t'^2} \, \mathrm{d}t' \\ &= \frac{1}{\pi} \arctan(t) + \frac{1}{2} \\ &= \textit{Cauchy CDF} \end{aligned}$$

But this is slightly weird? We defined the Cauchy as the ratio between two normals.

$$C = \frac{Z_1}{Z_2} \overset{?}{\sim} \frac{Z_1}{|Z_2|}$$

We have that

$$\begin{aligned} \frac{Z_1}{Z_2} &= \frac{Z_1}{S_2|Z_2|} \\ &= \frac{S_2 Z_1}{|Z_2|} \\ &= \frac{S_2 S_1 |Z_1|}{|Z_2|} \\ &\sim \frac{S_2 |Z_1|}{|Z_2|} \\ &\sim \frac{Z_1}{|Z_2|} \end{aligned}$$

Now, if we have a uniform random variable U, we can get a Cauchy random variable from

$$\tan(\pi(U - 1/2)) \overset{PIT}{\sim} \text{Cauchy}$$

When we look at the expectation, we have

$$\mathbb{E}\left[\frac{Z_1}{|Z_2|}\right] = \mathbb{E}[Z_1] \, \mathbb{E}\left[\frac{1}{|Z_2|}\right]$$

and the expectation on the right is

$$\int \frac{1}{z} e^{-1/2z^2} \, \mathrm{d}z \approx \int \frac{1}{z} \, \mathrm{d}z$$

when near 0, so this is going to blow up to infinity. Now consider

$$F^*_{a,b} \sim \frac{\text{Gamma}(a)}{\text{Gamma}(b)}$$

where the Gamma random variables are independent. The expectation is given by

$$\begin{aligned} \mathbb{E}\big[F^*_{a,b}\big] &= \mathbb{E}[\text{Gamma}(a)] \, \mathbb{E}\left[\frac{1}{\text{Gamma}(b)}\right] \\ &= a \times \frac{1}{b-1} \end{aligned}$$

We get this since

$$\begin{aligned} \int_0^\infty \frac{1}{x} \frac{x^b e^{-x}}{\Gamma(b)} \frac{\mathrm{d}x}{x} &= \frac{\Gamma(b-1)}{\Gamma(b)} \\ &= \frac{\Gamma(b-1)}{(b-1)(\Gamma(b-1)} \\ &= \frac{1}{b-1} \end{aligned}$$

6.1. **Famous Continuous Distributions on** $(0,\infty)$**.**

(1) Gamma(a) for $a>0$ (includes Exponential)
(2) χ^2_n where $n=1,2,\ldots$
(3) F, F^*.

$$F_{n,m}=\frac{\chi^2_m/m}{\chi^2_n/n}=\frac{n}{m}F^*_{m/2,n/2}$$

(4) Lognormal$(\mu,\sigma)\overset{\text{Rep}}{=}\exp(\mu+\sigma Z)$ where $Z\sim\mathcal{N}(0,1)$.
(5) Weibull. $X\sim\exp$

$$W=\alpha X^\beta$$

where α is a scale parameter and β is a power parameter.

$$\log W=\log\alpha+\beta\log X$$

Here, $\log\alpha$ is a location parameter and β is a scale parameter.

Definition 11 (Stable Laws)**.** *F is a stable CDF if given independent $Y_1,Y_2\sim F$, then $a_1Y_2+a_2Y_2+a_0=T$ has the property such that*

$$T\sim c_0+c_1Y$$

where $Y\sim F$.

Example 6. *Let C_1,C_2 be two independent Cauchy random variables.*

$$\bar{C}=\frac{C_1+C_2}{2}\sim \textit{Cauchy}$$

We can find this distribution using calculus, but why would you want to do something crazy like that? Let

$$\begin{aligned}C_1&=\frac{Z_1}{Z_2}\\&=\frac{R_1\sin(\Theta_1)}{R_1\cos(\Theta_1)}\\&=\tan(\Theta_1)\\&=\tan(2\pi U_1)\\C_2&=\frac{R_2\sin(\Theta_2)}{R_2\cos(\Theta_2)}\\&=\tan(2\pi U_2)\end{aligned}$$

7. September 24th, 2013

Let $X_1,X_2\overset{iid}{\sim}$ Expo. What is $X_1|X_1+X_2$, the conditional distribution?

$$\begin{aligned}X_1|X_1+X_2=&\ (X_1+X_2)\frac{X_1}{X_1+X_2}\,\Big|\,X_1+X_2\\&\sim(X_1+X_2)U\end{aligned}$$

where $U\sim$ Unif. We get this since we know $\frac{X_1}{X_1+X_2}\sim\text{Beta}(1,1)\sim U$.

7.1. **Poisson Process.** Interarrival times i.i.d. λ^{-1}Expo. Pois(λt) is the distribution of the number of arrivals in an interval of length t.

Properties of the Poisson distribution:

(1) $N_1\sim\text{Pois}(\lambda_1), X_2\sim\text{Pois}(\lambda_2)$, which are independent implies

$$N_1+N_2\sim\text{Pois}(\lambda_1+\lambda_2)$$

We can derive this using a convolution, MGF,
(2) $N_1|N_1+N_2=n\sim\text{Bin}\left(n,\frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
(3) "Chicken-egg problem". Suppose $N\sim\text{Pois}(\lambda)$, $X|N\sim\text{Bin}(N,p)$ and $X+Y=N$. (N is number of eggs, X is number of eggs that hatch, and Y is the number of eggs that don't hatch.) At a glance we may think X and Y are independent, but we actually have

$$X\sim\text{Pois}(\lambda p)\perp\!\!\!\perp Y\sim\text{Pois}(\lambda q)$$

7.2. **Generating a Poisson Process, Rate λ on $[0,t]$.**

(1) Draw $N_t \sim \text{Pois}(\lambda t)$.
(2) Draw N_t i.i.d. points in $[0,t]$.

7.3. **Count-time Duality.**

$$N_t = \text{number of arrivals up to } t \sim \text{Pois}(\lambda t)$$
$$T_n = \text{time of the nth arrival} \sim \lambda^{-1}\text{Gamma}(n)$$

$$\{N_t \geq n\} = \{T_n \leq t\}$$

It follows that

$$\mathbb{P}(N_t \geq n) = \underbrace{\mathbb{P}(T_n \leq t)}_{\text{CDF of } \lambda^{-1}\text{Gamma}(n)}$$

So then

$$\begin{aligned}\mathbb{P}(N_t = k) &= \mathbb{P}(T_k \leq t < T_{k+1}) \\ &= \mathbb{P}(T_k \leq t) - \mathbb{P}(T_{k+1} \leq t)\end{aligned}$$

as

$$\mathbb{P}(T_k \leq t) = \mathbb{P}(T_{k+1} \leq t) + \mathbb{P}(T_k \leq t < T_{k+1})$$

Definition 12 (Poisson Process in 2D)**.**

(1) The number of X's in a region of area a is $\text{Pois}(\lambda a)$.
(2) The number of X's in disjoint regions are independent.

7.3.1. *Superposition and Thinning.*
Superpositioning: Suppose we have Joe's emails are either STAT110 or STAT210. STAT110 emails are modeled by a Poisson process with rate parameter λ_1 and STAT220 with parameter λ_2, Then the superposed process is a Poisson process with rate $\lambda_1 + \lambda_2$.
Thinning: Now let p be the probability that the email is a 210 email. With this probability, whenever we get a new email, we will keep it with probability p. Now the two thinned processes (one of where we had emails and one where we discarded emails) is a Poisson process.

7.4. **Order Statistics, Rényi Representation.** If we have random variables $X_1, \ldots, X_n$, then the order statistics are

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

Let $X_1, \ldots X_n \overset{iid}{\sim}$ Expo. Suppose we have n independent Poisson processes, each with rate 1. The first arrival for each of these processes is an Exponential. It follows from super position that

$$\begin{aligned}X_{(1)} &\sim \frac{1}{n}X_1 \quad \text{by superposition} \\ X_{(2)} - X_{(1)} &\sim \frac{1}{n-1}X_2 \\ X_{(3)} - X_{(2)} &\sim \frac{1}{n-2}X_3 \\ &\vdots\end{aligned}$$

which are independent. This gives us that

$$(X_{(1)}, X_{(2)}, \ldots X_{(n)}) \sim \left(\frac{1}{n}X_1, \frac{1}{n}X_1 + \frac{1}{n-1}X_2, \ldots, \frac{1}{n}X_1 + \frac{1}{n-1}X_2 + \ldots + X_n\right)$$

Example 7.

$$\begin{aligned}\text{Cov}(X_{(1)}, X_{(2)}) &= \frac{1}{n^2}\,\text{Var}(X_1) \\ &= \frac{1}{n^2}\end{aligned}$$

8. September 26th, 2013

8.1. **Lebesgue Decomposition.** Any random variable X can be represented as

$$X \sim J_0 X_0 + J_1 X_1 + J_2 X_2$$

where J_0, J_1, J_2 are indicators with exactly one equal to 1.

X_0	purely discrete
X_1	absolutely continuous
X_2	singular continuous

$$m_1 \overbrace{<<}^{\text{abs. continuous}} m_2$$

is absolutely continuous if $m_2(A) = 0 \Rightarrow m_1(A) = 0$.

We have that something is singular continuous if

$$\mathbb{P}(X_2 = x) = 0 \text{ for all } x$$

but the support has measure 0. We have

$$(J_0, J_1, J_2) \perp\!\!\!\perp (X_0, X_1, X_2)$$

8.1.1. *Cantor Distribution.*

$$X_2 \sim \sum_{j=0}^{\infty} \frac{2B_j}{3^j}$$

where $B_j \overset{iid}{\sim} \text{Bern}(1/2)$. The CDF F is the "Cantor function", which is continuous, but $F'(x) = 0$. This distribution is an example of a singular continuous function.

8.2. **Expectation.**

$$\int g(x)\,\mathrm{d}F(x)$$

Riemann sum:

$$\sum_{i=1}^{n} g(x_i^*)\Delta x_i$$

Riemann-Stieltjes:

$$\sum_{i=1}^{n} g(x_i^*)\Delta F(x_i)$$

This is bad if F, g are discontinuous at the same point so for our purposes, we will assume that this isn't the case.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\,\mathrm{d}F(x)$$

where F is the CDF of X.
Special cases: Absolutely continuous,

$$\mathrm{d}F(x) = f(x)\,\mathrm{d}x$$

where f is the PDF. In the discrete case,

$$\sum_x x\mathbb{P}(X = x)$$

since $\Delta F(x_i) \neq 0$ only at the jumps. Otherwise it's flat and equals 0.

$$F = p_0 F_0 + p_1 F_1$$
$$\mathrm{d}F(x) = p_0\,\mathrm{d}F_0(x) + p_1\,\mathrm{d}F_1(x)$$

Linearity: This is harder to do with regular integration

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Do we have

$$\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y \overset{?}{=} \int_{-\infty}^{\infty} t f_T(t)\,\mathrm{d}t$$

where $T = X + Y$. Instead, we look at

$$\int_\Omega X(\omega)P(\mathrm{d}\omega) + \int_\Omega Y(\omega)P(\mathrm{d}\omega) = \int_\Omega (X + Y)(\omega)P(\mathrm{d}\omega)$$

but we haven't defined this yet.

What is an example of something that is not Riemann integral but is Lebesgue integrable??

$$I_{\mathbb{Q}} = \begin{cases} 1 & \text{on } \mathbb{Q} \\ 0 & \text{on } \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

Definition 13 (Lebesgue Integral)**.** *We define this in 4 steps*

(1) $X = I_A$*, the indicator function.* $\mathbb{E}[X] = A$ *which we call the fundamental bridge*

(2) Simple Random variables

$$X = \sum_{i=1}^{n} a_j I_{A_j}$$

and so

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_j P(A_j)$$

(3) $X \geq 0$

$$\mathbb{E}[X] = \sup\{\mathbb{E}[X^*] : X^* \geq 0, X^* \text{ simple } X^* \leq X\}$$

This is always defined in $[0, \infty]$*.*

(4) For general X*,*

$$X = X^+ - X^-$$

where

$$X^+ = \max(X, 0)$$
$$X^- = \max(-X, 0)$$

Then

$$\mathbb{E}[X] = \mathbb{E}\big[X^+\big] - \mathbb{E}\big[X^-\big]$$

unless we have $\infty - \infty$*. This makes sense since we need this to be linear.*

This is called the "InSiPod" process where I stands for indicator, S stands for simple, P stands for positive, and D stands for difference.

Proof of Linearity in Bounded Case. $|X| \leq c$, $|Y| \leq c$.

(1) if $X = I_A$, $Y = I_B$,

$$\begin{aligned} \mathbb{E}[X + Y] &= \mathbb{E}[I_A + I_B] \\ &= \mathbb{P}(A) + \mathbb{P}(B) \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

(2) Consider the case where

$$X = \sum_{i=1}^{n} a_i I_{A_i} \qquad Y = \sum_{j=1}^{m} b_j I_{B_j}$$

Then intersect A_i and B_j partitions to have an intersecting partition.

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} a_i I_{A_i} + \sum_{j=1}^{m} b_j I_{B_j}\right] &= \mathbb{E}\left[\sum_{i=1}^{k} c_i I_{C_i}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{k} c_i \mathbb{P}(C_i)\right] \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

(3) Suppose $X \geq 0, Y \geq 0$. Fix $\varepsilon > 0$. Find $X^*, Y^* \geq 0$ which are simple and $X^* \leq X, Y^* \leq Y$ such that

$$\mathbb{E}[X^*] \geq \mathbb{E}[X] - \varepsilon \qquad \mathbb{E}[Y^*] \geq \mathbb{E}[Y] - \varepsilon$$

$$\begin{aligned} \mathbb{E}[X + Y] &\geq \mathbb{E}[X^* + Y^*] \\ &= \mathbb{E}[X^*] + \mathbb{E}[Y^*] \\ &\geq \mathbb{E}[X] + \mathbb{E}[Y] - 2\varepsilon \end{aligned}$$

It follows that

$$\mathbb{E}[X + Y] \geq \mathbb{E}[X] + \mathbb{E}[Y]$$

For the other direction, replace X by $c - X$, and Y by $c - Y$.

(4) $X = X^+ - X^-$, $Y = Y^+ - Y^-$. Add big constant to X and Y.

□

$$X_n \overset{a.s.}{\to} X$$

means $X_n(\omega) \to X(\omega)$ for all ω except on $\omega \in N_0$ where $\mathbb{P}(N_0) = 0$. A weaker form of convergence is convergence in probability whereby

$$X + n \overset{P}{\to} X$$

means that for all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0$$

9.1. **Bounded Convergence Theorem.**

Theorem 9 (Bounded Convergence Theorem)**.** *Let $X_n \overset{P}{\to} X$, let $|X_n| \leq c$ almost surely. Then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.*

Proof. Show $\mathbb{E}[|X_n - X|] = 0$. Check that X is bounded.

$$\begin{aligned}\mathbb{P}(|X| > 2c) &\leq \mathbb{P}(|X - X_n| > c) \\ &\to 0 \quad \text{as } n \to \infty\end{aligned}$$

so $|X| \leq 2c$ almost surely.

Now fix $\varepsilon > 0$.

$$\begin{aligned}\mathbb{E}[|X_n - X|] &= \mathbb{E}[|X_n - X|]\, I(|X_n - X| > \varepsilon) + \mathbb{E}[|X_n - X|]\, I(|X_n - X| \leq \varepsilon) \\ &\leq 3c\mathbb{P}(|X_n - X| > \varepsilon) + \varepsilon \\ &\to \varepsilon\end{aligned}$$

and so

$$\lim_{n\to\infty} |X_n - X| = 0$$

□

9.2. **Covariance.**

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

Properties: (Bilinear form)

(1) $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
(2) $\operatorname{Cov}(X, Y + a) = \operatorname{Cov}(X, Y)$ where a is a constant
(3) $\operatorname{Cov}(X, Y + Z) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z)$
(4) $\operatorname{Cov}\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_{i,j} a_i b_j \operatorname{Cov}(X_i, Y_j)$
(5) $\operatorname{Cov}(X, X) = \operatorname{Var}(X) \geq 0$
(6) $\operatorname{Cov}(g(X), h(X)) \geq 0$ where g, h are increasing.

Let $X_1, X_2 \overset{iid}{\sim} X$. Consider

$$(g(X_1) - g(X_2))(h(X_1) - h(X_2)) \geq 0$$

This is non-negative since both g and h are monotone functions. Now take expectations.

$$\begin{aligned}&\mathbb{E}[(g(X_1) - g(X_2))(h(X_1) - h(X_2))] \\ &= \mathbb{E}[g(X_1)h(X_1)] - \mathbb{E}[g(X_2)h(X_1)] - \mathbb{E}[g(X_1)h(X_2)] + \mathbb{E}[g(X_2)h(X_2)] \\ &= 2\left(\mathbb{E}[g(X)h(X)] - \mathbb{E}[g(X)]\,\mathbb{E}[h(X)]\right) \quad \text{since } X_1 \perp\!\!\!\perp X_2 \\ &= 2\operatorname{Cov}(g(X), h(X)) \\ &\geq 0\end{aligned}$$

$$\operatorname{Corr}(X, Y) = \operatorname{Cov}\left(\frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}(X)}}, \frac{Y - \mathbb{E}[Y]}{\sqrt{\operatorname{Var}(Y)}}\right)$$

9.3. **Conditional Probability.** If we define the conditional expectation as $\mathbb{E}[Y|X=X] \equiv g(X)$, then what happens if $\mathbb{P}(X=x)=0$? There is a definition that conditions on that σ-algebra .$\mathbb{E}[Y|\mathcal{G}]$.

This is how we do it in this course. $\mathbb{E}[Y|X]$ is a random variable $g(X)$ such that

$$\mathbb{E}[(Y-g(X))h(X)]=0$$

for all bounded measurable h. $y-g(X)$ is uncorrelated with all $h(X)$. If we plugged in $h(X)=1$, then $\mathbb{E}[Y-g(X)]=0$, i.e.

$$\mathbb{E}[\mathbb{E}[Y|X]]=\mathbb{E}[Y]$$

We call this Adam's Law.

9.3.1. *Uniqueness.* Let $g_1(X), g_2(X)$ both satisfy the definition of $\mathbb{E}[Y|X]$. Then

$$g_1(X)=g_2(X) \quad \text{almost surely}$$

So we have

$$\mathbb{E}[(Y-g_1(X))h(X)]=0$$
$$\mathbb{E}[(Y-g_2(X))h(X)]=0$$

Taking the difference, we get

$$\mathbb{E}[(g_1(X)-g_2(X))h(X)]=0$$

for all functions h. The most obvious choice is to let $h(X)=\operatorname{sgn}(g_1(X)-g_2(X))$. Then

$$\mathbb{E}[|g_1(X)-g_2(X)|]=0$$

Lemma (Quadratic Function Lemma)**.** *If*

$$Q(x)=q_2(x^2)+q_1x+q_0$$

then

$$\mathbb{E}[Q(X)]=Q(\mathbb{E}[X])+q_2\operatorname{Var}(X)$$

9.3.2. *Eve's Law.*

$$\operatorname{Var}(Y)=\mathbb{E}[\operatorname{Var}(Y|X)]+\operatorname{Var}(\mathbb{E}[Y|X])$$

Proof. WELOG assume $\mathbb{E}[Y]=0$.

$$\begin{aligned}\operatorname{Var}(Y)&=\mathbb{E}\left[Y^2\right]\\&=\mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right]\\&=\mathbb{E}\left[(\mathbb{E}[Y|X])^2+\operatorname{Var}(Y|X)\right]\\&=\mathbb{E}[\operatorname{Var}(Y|X)]+\operatorname{Var}(\mathbb{E}[Y|X])\end{aligned}$$

□

9.3.3. *Ecce!*

$$\operatorname{Cov}(X,Y)=\mathbb{E}[\operatorname{Cov}(X,Y|Z)]+\operatorname{Cov}(\mathbb{E}[X|Z],\mathbb{E}[Y|Z])$$

Proof. Pencil Problem (First assume WELOG $\mathbb{E}[X]=0=\mathbb{E}[Y]$.) □

10. October 3rd, 2013

LOTEC: Law of the Extended Conversation

$$\mathbb{E}[Y]=\mathbb{E}\left[\sum_{k=1}^{\infty}\mathbb{E}[Y|X_i]\right]$$

Let

$$X=\begin{cases}0 & \text{male}\\ 1 & \text{female}\end{cases}$$

Let's just put in random numbers for the estimate. Let Y be height.

$$\mathbb{E}[Y|X=1]=64$$
$$\mathbb{E}[Y|X=0]=70$$
$$\mathbb{E}[Y|X]=64X+(1-X)70$$

Again, let's estimate standard deviations

$$\begin{aligned}\operatorname{Var}(Y|X)&=\begin{cases}3^2 & \text{male}\\ 2.5^2 & \text{female}\end{cases}\\&=2.5^2+(1-X)3^2\end{aligned}$$

We also have

$$\mathbb{E}[\mathrm{Var}(Y|X)] = (2.5)^2(0.48) + 3^2(0.52)$$

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

Now, say if $X \sim F$ and $Y \sim G$, then what is $\mathbb{P}(X + Y \leq t)$?

$$\begin{aligned}\mathbb{P}(X+Y \leq t) &= \mathbb{E}[\mathbb{P}(X+Y \leq t|X)] \\ &= \mathbb{E}[G(t-X)] \\ &= \int_{-\infty}^{\infty} g(t-x)\,\mathrm{d}F(x)\end{aligned}$$

This gives us

$$\int_{-\infty}^{\infty} g(t-x)f(x)\,\mathrm{d}x$$

for the density of $X + Y$ if F, G are absolutely continuous.

3 facts:

(a) If X, Y are such that

$$\mathrm{Cov}(g(X), h(Y)) = 0$$

for all g, h, then $X \perp\!\!\!\perp Y$.

(b) If

$$\mathrm{Cov}(g(X), Y) = 0$$

for all g, then they are not independent

Example 8. *If* $\mathrm{Cov}(g(X), Y - k(X)) = 0$*, for all* g*, then* $k(X) = \mathbb{E}[Y|X]$

(c) $\mathrm{Cov}(X, Y) = 0$

Theorem 10 (ECCE Theorem)**.**

$$\mathrm{Cov}(X, Y) = \mathbb{E}[\mathrm{Cov}(X, Y|Z)] + \mathrm{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z])$$

Suppose $Z = X$*. Then*

$$\mathrm{Cov}(X, Y) = 0 + \mathrm{Cov}(X, \mathbb{E}[Y|X]) = \mathrm{Cov}(X, \mathbb{E}[Y|X])$$

Let us have X, Y with pdf $f(x, y)$.

$$f_{Y|X}(y|x) = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y')\,\mathrm{d}y'}$$

so

$$\mathbb{E}[Y|X = x] = \frac{\int_{\mathbb{R}} y f(x,y)\,\mathrm{d}y}{\int_{\mathbb{R}} f(x,y)\,\mathrm{d}y}$$

Let $A = (a, b)$. Find $\mathbb{E}[Y|X \in A]$.

$$f(y|X \in A) = \frac{\int_a^b f(x,y)\,\mathrm{d}x}{\int_{\mathbb{R}} \int_a^b f(x,y')\,\mathrm{d}x\mathrm{d}y'}$$

so

$$\mathbb{E}[Y|X \in A] = \frac{\int_{\mathbb{R}} \int_a^b y f(x,y)\,\mathrm{d}x\mathrm{d}y}{\int_{\mathbb{R}} \int_a^b f(x',y')\,\mathrm{d}y'\,\mathrm{d}x'}$$

This above is probably wrong....

10.1. **Random Sums.**

$$T = \sum_{i=1}^{N} Y_i$$

where the Y_i are random. $N \perp\!\!\!\perp \{Y_i\}$ for $i = 1, 2 \ldots$.

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T|N]] = \mathbb{E}\left[\sum_{i=1}^{N} \mu_i\right]$$

We have $Y_i \sim [mu, \sigma^2]$

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}[N\mu] \\ &= \mu\mathbb{E}[N]\end{aligned}$$

To get the variance, we do

$$\begin{aligned}\operatorname{Var}(T) &= \mathbb{E}[\operatorname{Var}(T|N)] + \operatorname{Var}\mathbb{E}[T|N] \\ &= \mathbb{E}\big[N\sigma^2\big] + \operatorname{Var}(N\mu) \\ &= \sigma^2\mathbb{E}[N] + \mu^2 \operatorname{Var}(N)\end{aligned}$$

If $N \sim \text{Pois}(\lambda)$, $\operatorname{Var}(T) = \lambda(\sigma^2 + \mu^2)$.

Suppose we wish to predict Y from X i.e. minimize

$$\mathbb{E}\big[(Y - g(X))^2|X\big]$$

This is minimized by $\hat{g}(X) = \mathbb{E}[Y|X]$. How do we know this is the best? Suppose we want to minimize

$$\min_c \mathbb{E}\big[(Y-c)^2\big] = \mathbb{E}\big[(Y-\mu)^2\big] = \operatorname{Var}(Y)$$

Then clearly, choosing $c = \mu$ yields the variance.

10.2. **Borel's Paradox.** Given $X_1, X_2 \overset{iid}{\sim}$ Expo, find

$$\mathbb{E}[X_1|X_1 = X_2]$$

(a) Let $U = \frac{X_1}{X_1+X_2} \perp\!\!\!\perp X_1 + X_2 = T$ then

$$\begin{aligned}\mathbb{E}[X_1|X_1 = X_2] &= \mathbb{E}\left[UT \,\middle|\, V = \frac{1}{2}\right] \\ &= \frac{1}{2} \times 2 \\ &= 1\end{aligned}$$

(b) Let

$$\begin{aligned}D &= X_1 - X_2 \\ &\sim SX\end{aligned}$$

$\min(X_1, X_2) \perp\!\!\!\perp D$, so

$$\mathbb{E}[X_1|D = 0] = \mathbb{E}[\min(X_1, X_2)] = \mathbb{E}\left[\frac{1}{2}X\right] = \frac{1}{2}$$

$$\mathbb{E}[X_1|\,|X_1 - X_2| < \varepsilon] \neq \mathbb{E}\left[\left|\,\left|\frac{X_1}{X_1+X_2} - \frac{1}{2}\right| < \varepsilon\right.\right]$$

11. October 8th, 2013

Theorem 11 (Bayes' Theorem)**.**

$$f(y_1|\theta)g(\theta) = p(y_1, \theta) = f_0(y_1)g_1(\theta|y_1)$$

so

$$\begin{aligned}g_1(\theta|y_1) &= \frac{p(y_1,\theta)}{f_0(y_1)} \\ &= \frac{p(y_1,\theta)}{\int p(y_1,\theta)\,\mathrm{d}\theta} \\ &= \frac{f(y_1|\theta)g(\theta)}{\int f(y_1|\theta)g(\theta)\,\mathrm{d}\theta}\end{aligned}$$

For question 5.4, we have

$$\begin{aligned} Y_i &\overset{iid}{\sim} f(y_i|\theta) \\ \theta &\sim g(\theta|\alpha) \\ \alpha &\sim h(\alpha) \end{aligned}$$

Bayes' theorem for this yields

$$g_1(\theta|y_1) = \frac{f(y_1|\theta)\int g(\theta|\alpha)h(\alpha)\,\mathrm{d}\alpha}{\int\int f(y_1|\theta)g(\theta|\alpha)h(\alpha)\,\mathrm{d}\alpha\mathrm{d}\theta}$$

If $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

We divide by $n-1$ to make it unbiased. We have

$$\bar{y} \perp\!\!\!\perp s^2 \sim \sigma^2 \frac{\chi^2_{n-1}}{n-1}$$

Definition 14 (Moment Generating Function)**.** *Let Y be a 1-dimensional real valued random variable and t a real value.*

$$\mathbb{E}\left[e^{tY}\right] = M(t)$$

The MGF exists if there is a T such that $M(t) < \infty$ for $t \in T = (-a, b)$ where $a, b > 0$, that is the interval contains 0.

Example 9. $X \sim$ Expo. *Then*

$$\begin{aligned} \mathbb{E}\left[e^{tX}\right] &= \int_0^\infty e^{tx-x}\,\mathrm{d}x \\ &= \int_0^\infty e^{(t-1)x}\,\mathrm{d}x \\ &= \frac{1}{1-t} \quad \textit{if } t < 1 \end{aligned}$$

$X \sim \mathcal{N}(0,1)$.

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \frac{e^{tx-\frac{1}{2}x^2}}{\sqrt{2\pi}}\,\mathrm{d}x \\ &= e^{\frac{1}{2}t^2}\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(x-t)^2}}{\sqrt{2\pi}}\,\mathrm{d}x \\ &= e^{\frac{1}{2}t^2} \quad t \in \mathbb{R} \end{aligned}$$

$Y \sim \mathcal{N}(\mu, \sigma^2)$. *This means* $Y = \mu + \sigma X$.

$$\begin{aligned} M(t) &= \mathbb{E}\left[e^{t\mu + t\sigma X}\right] \\ &= e^{t\mu}\mathbb{E}\left[e^{t\sigma X}\right] \\ &= e^{t\mu + \frac{1}{2}t^2\sigma^2} \end{aligned}$$

Note that

$$\mathbb{E}\left[e^{tY}\right] = e^{\mathbb{E}[tY] + \frac{1}{2}\operatorname{Var}(tY)}$$

so for $t = 1$,

$$\mathbb{E}\left[e^{Y}\right] = e^{\mathbb{E}[Y] + \frac{1}{2}\operatorname{Var}(Y)}$$

Now let $Z \sim \mathcal{N}(0,1)$. *We look at* $\mathbb{E}[Z^n]$. *If* n *is odd, then this is zero. If it's even, consider* $\mathbb{E}\left[Z^{2m}\right] = \mathbb{E}\left[(Z^2)^m\right]$. *Let* G *be a Gamma*$(1/2)$. *Then*

$$\begin{aligned} \mathbb{E}\left[(Z^2)^m\right] &= 2^m\mathbb{E}[G^m] \\ \mathbb{E}[G^m] &= \int_0^\infty \frac{x^{\frac{1}{2}+m}e^{-x}}{\Gamma(1/2)}\frac{\mathrm{d}x}{x} \\ &= \frac{\Gamma(m+1/2)}{\Gamma(1/2)} \end{aligned}$$

for $m = 2$, *we get*

$$\frac{3}{2}\frac{\Gamma(1/2)^{\frac{1}{2}}}{\Gamma(1/2)}2^m = \frac{3\times 1}{2\times 2}2^2 = 3$$

If $m = 3$, then we have $5 \times 3 \times 1$. If $m = 4$, then we have $7 \times 5 \times 3 \times 1$. Compare

$$\begin{aligned} M(t) &= e^{-\frac{1}{2}t^2} \\ &= 1 + \frac{1}{2}t^2 + \frac{(\frac{1}{2}t^2)^2}{2!}|\frac{(\frac{1}{2}t^2)^3}{3!} + \cdots \\ M(t) &= \mathbb{E}\left[e^{tZ}\right] \\ &= \mathbb{E}\left[1 + tZ + \frac{1}{2}t^2Z^2 + \frac{1}{3!}t^3Z^3 + \cdots\right] \\ &= 1 + \frac{1}{2}t^2\mathbb{E}\left[Z^2\right] + \frac{1}{4!}t^4\mathbb{E}\left[Z^4\right] + \cdots \end{aligned}$$

Some Facts:
If X_1, X_2 are independent (with $M_j(t) = \mathbb{E}\left[e^{tX_j}\right]$), then

$$\mathbb{E}\left[e^{t(a_0+a_1X_1+a_2X_2)}\right] = e^{ta_0}M_1(a_1t)M_2(a_2t)$$

If all the moments exist and are the same, then the distributions are not necessarily the same. Let $Y \sim LN(\mu, \sigma^2)$, the log-normal distribution. Then

$$Y = \exp(\mu + \sigma Z)$$

This gives us

$$\mathbb{E}[Y] = e^{\mu+\frac{1}{2}\sigma^2}$$

to get the variance, we have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 \\ &= \mathbb{E}[\exp(2\mu + 2\sigma Z)] - \exp(2\mu + \sigma^2) \\ &= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \\ &= (\mathbb{E}[Y])^2\{e^{\sigma^2} - 1\} \end{aligned}$$

12. October 10th, 2013

From last time, if $Y \sim LN(\mu, \sigma^2)$, then

$$\begin{aligned} Y &= e^{\mu+Z\sigma} \\ \mathbb{E}[Y] &= e^{\mu+\frac{1}{2}\sigma^2} \\ \mathrm{Var}(Y) &= (\mathbb{E}[Y])^2(e^{\sigma^2} - 1) \end{aligned}$$

To get $\mathbb{E}[Y^n]$, we have

$$\mathbb{E}[Y^n] = e^{n\mu+\frac{1}{2}n^2\sigma^2}$$

However, there is no moment generating function since

$$\begin{aligned} \mathbb{E}\left[e^{tY}\right] &= \sum_{n=0}^{\infty} \frac{\mathbb{E}[tY]^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} e^{\frac{1}{2}n^2\sigma^2} \end{aligned}$$

if $\mu = 0$. This thing diverges since

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

and $n^n = e^{n \log n} < e^{n^2}$, so the $e^{\frac{1}{2}n^2\sigma^2}$ term dominates.

$$\begin{aligned} M(t) &= \mathbb{E}\left[e^{tY}\right] \\ &= \sum_{n=0}^{\infty} \frac{t^n\mathbb{E}[Y^n]}{n!} \end{aligned}$$

the nth raw moment gives

$$M^{(r)}(t) = \frac{r!}{r!}\mathbb{E}[Y^r] + \cdots$$

where the other terms vanish for $t = 0$.

$$\mathbb{E}\Big[e^{t(Y-\mu)}\Big] = \sum_{n=0}^{\infty} \frac{t^n(\mathbb{E}[(Y-\mu)^n]}{n!}$$

$\mu_n = \mathbb{E}[(Y-\mu)^n]$ is the nth central moment.

Definition 15 (Cumulant Function)**.**

$$\begin{aligned} K(t) &\equiv \log \mathbb{E}\big[e^{tY}\big] = \log M(t) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\kappa_n \end{aligned}$$

where κ_n is the nth cumulant of $Y = K_n(Y) = K^{(n)}(t)\Big|_{t=0}$.

$$\begin{aligned} K_1(Y) &= \mathbb{E}[Y] = \frac{M'(t)}{M(t)}\Big|_{t=0} \\ K_2(Y) &= \mathrm{Var}(Y) \end{aligned}$$

In generality,

$$\begin{aligned} M(t) &= e^{K(t)} \\ M'(t) &= K'(t)e^{K(t)} = K'M \\ M''(t) &= K''M + K'M' \\ M''(0) &= K''(0) + 0 \quad \text{ if } M'(0) = 0 \Leftrightarrow \mathbb{E}[Y] = 0 \\ K_2(Y) &= \mathrm{Var}(Y) \end{aligned}$$

Now to try the third derivative,

$$\begin{aligned} M''' &= K'''M + 2K''M' + K'M'' \\ \mathbb{E}\big[(y-\mu)^3\big] &= K'''(0) \\ &= K_3(Y) \end{aligned}$$

The fourth cumulant gives

$$K_4(Y) = \mathbb{E}\big[(y-\mu)^4\big] = -3(\mathbb{E}\big[(y-\mu)^2\big])^2$$

- Normal: Cumulants are 0 after the 2nd one
- Poisson(λ): $K_n(\lambda) = \lambda$.

$$\begin{aligned} K(t) &= \sum_{n=1}^{\infty} \frac{t^n K_n}{n!} \\ &= \sum_{n=1}^{\infty} \frac{t^n \lambda}{n!} \\ &= \lambda(e^t - 1) \end{aligned}$$

Lemma. *If $X \perp\!\!\!\perp Y$, prove that*

$$K_n(X+Y) = K_n(X) + K_n(Y)$$

Proof.

$$\begin{aligned} \log(M_{X+Y}(t) &= \log M_X(t) + \log M_Y(t) \\ K_{X+Y}(t) &= K_X(t) + K_Y(t) \\ \frac{\mathrm{d}^n}{\mathrm{d}t^n} K_{X+Y}(t)\Big|_{t=0} &= \frac{\mathrm{d}^n}{\mathrm{d}t^n} K_X(t)\Big|_{t=0} + \frac{\mathrm{d}^n}{\mathrm{d}t^n} K_Y(t)\Big|_{t=0} \\ K_n(X+Y) &= K_n(X) + K_n(Y) \end{aligned}$$

□

Note: $K_n(aX) = a^n K_n(X)$. If $X_1, \ldots, X_n$ are independent, then

$$K_n(a_0 + a_1X_1 + a_2X_2 + \cdots) = a_1^r K_2(X) 1_+ a_2^r K_2(X_2) + \cdots$$

for $r \geq 2$.

Now, given $Y \sim F_0(y)$ and the MGF exists.

$$K_0(t) = \log(\mathbb{E}\left[e^{tY}\right])$$
$$e^{K_0(t)} = \int_y e^{ty} \, \mathrm{d}F_0(y)$$
$$1 = \int_y e^{\theta y - K_0(\theta)} \, \mathrm{d}F_0(y)$$

NEF = Natural Exponential Family.

$$\mathrm{d}F_0(y) = e^{\theta y - K_0(\theta)} \, \mathrm{d}F_0(y)$$

12.1. **Vectors and Matrices.** Let

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \in \mathbb{R}^k$$

Then

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] = \mu_1 \\ \vdots \\ \mathbb{E}[X_k] = \mu_k \end{pmatrix}$$

If $M = (X - \mu)(X - \mu)'$ is the matrix with entries $((X_i - \mu_i)(X_j - \mu_2))$. then

$$\mathbb{E}[M] = \mathrm{Cov}(X_i, X_j)$$

If $i = j$, then the entry is $\mathrm{Var}(X_i)$.

$$V = \Sigma = \mathrm{Cov}(X, X)$$

Suppose

$$X \sim \left[\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, V \right]$$

Then if

$$Y = \underbrace{A}_{m \times k} \underbrace{X}_{k \times 1} + \underbrace{b}_{m \times 1}$$

we have

$$Y \sim \left[A\mu + b, \underbrace{A}_{m \times k} \underbrace{V}_{k \times k} \underbrace{A'}_{k \times m} \right]$$

13. October 15th, 2013

Skewness of X.

$$K_3\left(\frac{X - \mu}{\sigma}\right) = \frac{\mathbb{E}\left[(x - \mu)^3\right]}{\sigma}$$

Kurtosis of X

$$K_4\left(\frac{X - \mu}{\sigma}\right) = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] - 3\left(\mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^2\right]\right)^2$$

The moment generating function is given by

$$M(t) = \exp(K(t)) = \exp\left(\mathbb{E}[X]\, t + \frac{1}{2} t^2 \, \mathrm{Var}(X) + \frac{1}{3!} y^3 K_3(t) + \cdots\right)$$

If $Y = e^{\mu + \sigma Z}$, then

$$\mathbb{E}[Y] = e^{\mu + \frac{1}{2}\sigma^2 + \frac{1}{6}}$$

Calculate $\mathbb{E}\left[\left\{\sum a_j (X_j - \mu_j)\right\}^4\right]$?

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \sim \mathrm{Dist}\left[\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, V = \begin{pmatrix} v_{11} & \cdots & v_{1k} \\ \vdots & \ddots & \vdots \\ v_{k1} & \cdots & v_{kk} \end{pmatrix} \right]$$

where $v_{ij} = \mathrm{Cov}(X_i, X_j)$.

Definition 16 (Multinomial Distribution)**.** $X \sim \text{Mult}(n, p)$ *has pmf*

$$\begin{aligned}\mathbb{P}(X = x) &= \binom{n}{x_1 \quad \cdots \quad x_k} \prod_{j=1}^{k} p_j^{x_j} \\ &= \frac{n!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k} p_j^{x_j}\end{aligned}$$

Then X has mean np and

$$V = n(D_p - pp')$$

13.1. **Multivariate Normal Distribution.** Start with

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}, \qquad Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$$

We write this as

$$Z \sim \mathcal{N}_k \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, I_k \right)$$

which has density

$$\prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_j^2} = \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{1}{2} \sum_{j=1}^{k} z_j^2}$$

where

$$\sum_{j=1}^{k} z_j^2 = \|\boldsymbol{z}\|^2 = \boldsymbol{z}'\boldsymbol{z}$$

Definition 17 (Multivariate Normal Distribution (by Representation))**.** $Y \sim \text{MVN}$ *if we have*

$$Y = AZ + \mu$$

where Z and μ are both of dimension $k \times 1$. Then

$$Y \sim \mathcal{N}_k(\mu, AA')$$

since

$$\begin{aligned}V &= \mathbb{E}[(Y - \mu)(Y - \mu)')] \\ &= \mathbb{E}[(AZ)(Z'A")] \\ &= A\mathbb{E}[ZZ']\, A' \\ &= AA'\end{aligned}$$

Is this unique though? Is it possible to pick a B such that $Y = BZ + \mu$ and $V = BB' = AA$ but $B \neq A$? We can write

$$A = B\Gamma$$

where $\Gamma_{k \times k}$ is orthogonal ($\Gamma\Gamma' = I_k$).

In fact, we can write

$$V = \Gamma D_\lambda \Gamma'$$

where Γ is orthogonal and D_λ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_k$.

If $V = \mathbb{E}[(Y - \mu)(Y - \mu)']$, then V is non-negative definite, which means that

$$a'Va \geq 0$$

for all a. If instead, we have that $a'Va > 0$ for $a \neq 0$, then we say V is positive definite.

Theorem 12. *A Covariance matrix V is non-negative definite, that is $V \geq 0$.*

Proof. For any a

$$\begin{aligned}a'Va &= a'\mathbb{E}[(Y - \mu)(Y - \mu)']\, a \\ &= \mathbb{E}[(a'(Y - \mu))(a'(Y - \mu))'1] \\ &= \mathbb{E}\big[(a'(Y - \mu))^2\big]\end{aligned}$$

□

We can also write

$$\begin{aligned} V &= \Gamma D_\lambda \Gamma' \\ &= (\Gamma D_{\sqrt{\lambda}})(\Gamma D_{\sqrt{\lambda}})' \\ &= AA' \end{aligned}$$

13.1.1. *Density of the Multivariate Normal.* We have the density of $Z = (Z_1, \ldots, Z_k)'$ is

$$\frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}z'z}$$

If $V > 0$, $Y = AZ + \mu$, then the Jacobian is

$$J = \frac{1}{|\frac{\mathrm{d}y}{\mathrm{d}z}|} = \frac{1}{|AA'|^{1/2}} = \frac{1}{|V|^{1/2}}$$

Since $Z = A^{-1}(Y - \mu)$, then

$$z'z = (y-\mu)' \underbrace{A'^{-1}A'}_{V^{-1}} (y - \mu)$$

So the density of Y is given by

$$f_Y(y) = \frac{1}{(2\pi)^{k/2}} \frac{1}{|V|^{1/2}} e^{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)}$$

To get the MGF,

$$\begin{aligned} \mathbb{E}\left[e^{t'Y}\right] &= e^{\mathbb{E}[t'Y] + \frac{1}{2}\operatorname{Var}(t'Y)} \\ &= e^{t'\mu + \frac{1}{2}t'Vt} \end{aligned}$$

14. October 17th, 2013

If V is non-negative definite, then this is equivalent to

$$\underbrace{t'}_{1\times k} \underbrace{V}_{k\times k} \underbrace{t}_{k\times 1} \geq 0$$

for all $t \in \mathbb{R}^k$. Then we decompose $V = AA'$ using Choleski (for example). Let

$$\underbrace{\vec{y}}_{k\times 1} = \underbrace{A}_{k\times m} \underbrace{\vec{x}}_{m\times 1} + \underbrace{\vec{\mu}}_{k\times 1} \sim \mathcal{N}_k(\mu, V)$$

with

$$\vec{x} \overset{iid}{\sim} \mathcal{N}(0,1)$$

Theorem 13. *$\vec{y}$ is MVN iff every linear combination of the components of $\vec{y}$ is (univariate) Normal*

Proof. ($\Rightarrow$)

$$\vec{t}'\vec{y} = \vec{t}'A\vec{x} + \vec{t}\vec{\mu}$$

is multivariate normal of dimension 1.
($\Leftarrow$) Find the MGF of $\vec{y}$.

$$\begin{aligned} M(\vec{t}) &= \mathbb{E}\left[e^{\vec{t}'\vec{y}}\right] \\ &= \mathbb{E}\left[\vec{t}'\vec{\mu} + \frac{1}{2}\vec{t}'V\vec{t}\right] \end{aligned}$$

which implies that $\vec{y}$ is MVN □

Related: Cramer-Wold device. Any multivariate distribution is determined by the univariate distribution of all the linear combinations.

14.0.2. *Properties of the MVN.*

(1) subvectors of a MVN are MVN

$$\vec{y} = \begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \end{pmatrix}$$

where $\vec{y}_1$ is $k_1 \times 1$ and $\vec{y}_2$ is $k_2 \times 1$, with $k = k_1 + k_2$.

$$\vec{y}_1 = \begin{pmatrix} I & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \end{pmatrix}$$

which is precisely the form of a MVN

(2) Within MVN, uncorrelated implies independence. (A counterexample for a non-normal RV is looking at Z and SZ where S is a random sign. Then

$$\mathbb{E}\left[SZ^2\right] = \mathbb{E}[S]\,\mathbb{E}\left[Z^2\right] = 0$$

so they are uncorrelated but obviously not independent)

Proof. WELOG, $\vec{\mu} = 0$. Uncorrelated implies

$$V = \begin{pmatrix} V_{11} & 0 \\ 0 & V_{22} \end{pmatrix}$$

Then

$$V_{11} = A_{11}A_{11}' \qquad V_{22} = A_{22}A_{22}'$$

which gives that

$$V = \underbrace{\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}}_{A} \begin{pmatrix} A_{11}' & 0 \\ 0 & A_{22}' \end{pmatrix}$$

So our representation is

$$\vec{y} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} A_{11}Z_2 \\ A_{22}Z_2 \end{pmatrix}$$

where

$$Z = \begin{pmatrix} \vec{Z}_1 \\ \vec{Z}_2 \end{pmatrix}$$

with $\vec{Z}_1$,$\vec{Z}_2$ being i.i.d. $\mathcal{N}(0,1)$, so

$$\vec{y}_1 \perp\!\!\!\perp \vec{y}_2$$

The independence here relies on the fact that the our building blocks, Z_i's are independent. □

14.0.3. *Distribution of $\vec{y}_2|\vec{y}_1$.* Idea: Decompose

$$\vec{y}_2 = \vec{y}_{2\cdot 1} + B\vec{y}_1$$

where $\vec{y}_{2\cdot 1} \perp\!\!\!\perp \vec{y}_1$. Clearly, $B\vec{y}_1$ is a function of $\vec{y}_1$. Working backwards, we really have that

$$\vec{y}_{2\cdot 1} = \vec{y}_2 - B\vec{y}_1$$

$$\begin{aligned} 0 &= \mathrm{Cov}(\vec{y}_{2\cdot 1}, \vec{y}_1) \\ &= V_{21} - BV_{11} \end{aligned}$$

which implies

$$B = V_{21}V_{11}^{-1}$$

(assuming V_{11}^{-1} exists)

$$\vec{y}_2|\vec{y}_1 \sim \mathcal{N}_{k_2}(V_{21}V_{11}^{-1}\vec{y}_1, V_{22\cdot 1})$$

where

$$\begin{aligned} V_{22\cdot 1} &= \mathrm{Cov}(\vec{y}_{2\cdot 1}) \\ &= \mathrm{Cov}(\vec{y}_2 - V_{21}V_{11}^{-1}\vec{y}_1, \vec{y}_2 - V_{21}V_{11}^{-1}\vec{y}_1) \\ &= \mathrm{Cov}(\vec{y}_2 - V_{21}V_{11}^{-1}\vec{y}_1, \vec{y}_2) \\ &= \underbrace{V_{22}}_{k_2\times k_2} - \underbrace{V_{21}}_{k_2\times k_1}\underbrace{V_{11}^{-1}}_{k_1\times k_1}\underbrace{V_{12}}_{k_1\times k_2} \end{aligned}$$

Let $Z_1, \ldots, Z_n \overset{iid}{\sim} \mathcal{N}(0,1)$. Show

$$\bar{Z} \perp\!\!\!\perp S^2 = \frac{1}{n-1}\sum_j (Z_j - \bar{Z})^2$$

(1) Basu's theorem

(2) Orthogonal transformations
(3) Conditioning and MVN properties.
Show $\bar{Z} \perp\!\!\!\perp (Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$ To do this, we note that $(\bar{Z}, Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})$ is MVN. It suffices to show that

$$\operatorname{Cov}(\bar{Z}, Z_j - \bar{Z}) = 0$$

Then, by Adam's law,

$$\begin{aligned}\mathbb{E}\big[\bar{Z}(Z_j - \bar{Z})\big] &= \mathbb{E}\big[\mathbb{E}\big[\bar{Z}(Z_j - \bar{Z})|\bar{Z}\big]\big] \\ &= \mathbb{E}\big[\bar{Z}\mathbb{E}\big[Z_j - \bar{Z}|\bar{Z}\big]\big]\end{aligned}$$

but we have

$$\mathbb{E}\big[Z_1|\bar{Z}\big] = \mathbb{E}\big[Z_2|\bar{Z}\big] = \cdots = \mathbb{E}\big[Z_n|\bar{Z}\big]$$

If we add these and use linearity, then

$$n\mathbb{E}\big[Z_1|\bar{Z}\big] = Z_1 + \cdots + Z_n$$

which implies

$$\mathbb{E}\big[Z_j|\bar{Z}\big] = \bar{Z}$$

15. October 22nd, 2013

15.1. **Exponential Families.**

Definition 18 (NEF: Natural Exponential Family in the Univariate Case)**.**

$$\mathrm{d}F_\eta(y) = e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y)$$

where η is called the natural parameter, y is called the natural observation, and F_0 is a CDF not dependent on η. Here $\psi(0) = 0$.

$$P_\eta(Y \in B) = \int_B e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y)$$

Definition 19 (EF: Exponential Family)**.** *If we take a nonlinear transformation of $y = T(x)$, e.g.*

$$e^{\eta(\theta)x^3}A(\theta)h(x)\,\mathrm{d}x$$

Example 10.

- *Take $\mathcal{N}(\mu, 1)$, which has density*

$$\begin{aligned}&\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}(y-\mu)^2} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}}_{\mathcal{N}(0,1)}\, e^{-y^2/2}e^{\mu y - \mu^2/2}\end{aligned}$$

  where $\eta = \mu$ and $\psi(\eta) = \mu^2/2$.
- *Take* $\mathrm{Bin}(n, p)$*, which has PMF*

$$\begin{aligned}&\binom{n}{y}p^y q^{n-y} \\ &= \binom{n}{y}e^{y\log p + (n-y)\log q} \\ &= \binom{n}{y}e^{y logit p + n\log q}\end{aligned}$$

  and here $\eta = \operatorname{logit}(p)$.

Find the MGF of the $\mathrm{d}F_\eta$ distribution.

$$\begin{aligned}\mathbb{E}_\eta(e^{tY}) &= \int e^{ty + ny - \psi(\eta)}\,\mathrm{d}F_0(y) \\ &= \int e^{(t+\eta)y - \psi(\eta)}\,\mathrm{d}F_0(y) \\ &= e^{-\psi(\eta)}e^{\psi(\eta+t)}\int e^{(t+\eta)y - \psi(t+\eta)}\,\mathrm{d}F_\eta(y) \\ &= e^{\psi(t+\eta)-\psi(\eta)}\end{aligned}$$

Since the density integrates to 1,

$$\begin{aligned}
e^{\psi(\eta)} &= \int e^{\eta y}\,\mathrm{d}F_0(y) \\
\psi'(\eta)e^{\psi(\eta)} &= \int y e^{\eta y}\,\mathrm{d}F_0(y) \\
\psi'(\eta) &= \int y e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y) \\
&= \mu \\
&= \mathbb{E}_\eta[Y] \\
\psi''(\eta) &= \int y(y - \underbrace{\psi'(\eta)}_{\mu}) e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y) \\
&= \mathbb{E}_\eta(y(y-\mu)) \\
&= \mathrm{Var}_\eta(Y) \equiv V(\mu)
\end{aligned}$$

This doesn't say that $V(\mu)$ is the variance of our parameter μ, but rather the variance of our random variable as a function of μ. Is is valid to write the variance as a function of the mean i.e. is $\psi'(\eta)$ invertible? This is true since the variance (or second derivative) is positive so that $\psi'(\eta)$ is increasing and one-to-one.

15.2. **NEF-QVF (Quadratic Variance Function).**

- $\mathcal{N}(\mu, 1)$: $V(\mu) = 1$ is constant.
- $\mathrm{Pois}(\mu)$: $V(\mu) = \mu$
- $\mathrm{Bern}(p)$: $V(\mu) = \mu(1-\mu)$
- $\mathrm{Geom}(p)$: Mean: $\frac{q}{p} = \mu$, Variance: $\frac{q}{p^2} = \frac{q}{p}\frac{1}{p} = \mu(1+\mu)$.
- $\mu \cdot \mathrm{Expo}$: Mean: μ, Variance: μ^2.
- NEF-CHS (NEF-GHS) (Carl proved that there are only 6 of them)

An example of a bad family is $\mathrm{Unif}(0, \theta)$, where the support depends on θ.

If we know the variance function V (including the domain), then we know the NEF.

Recall that

$$V(\mu) = \psi''(\eta) = \frac{\mathrm{d}\psi'(\eta)}{\mathrm{d}\eta} = \frac{\mathrm{d}\mu}{\mathrm{d}\eta}$$

Then

$$\begin{aligned}
\int \frac{\mu}{V(\mu)}\,\mathrm{d}\mu &= \int \mu\,\mathrm{d}\eta \\
&= \int \psi(\eta)\,\mathrm{d}\eta \\
&= \psi(\eta) + C
\end{aligned}$$

$$e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y)$$

we know F_0 from the fact that its MGF is $e^{\psi(t)}$.

15.3. **MLE in NEF.** We find η to maximize

$$\mathcal{L}(\eta) = e^{\eta y - \psi(\eta)}$$

Then taking the log-likelihood

$$l(\eta) = ny - \psi(\eta)$$

Taking the derivative gives

$$l'(\eta) = y - \psi'(\eta) = 0$$

and so

$$\psi'(\hat{\eta}) = y = \hat{\mu}$$

To check that we indeed got a maximum,

$$l''(\eta) = -\psi''(\eta) < 0$$

since it is the variance.

15.3.1. *Operations.*

(1) Location-Scale
(2) Generation (exponential-tilting)
(3) Convolution/Division

16. October 24th, 2013

The variance function is quadratic if $V(\mu) = v_2\mu^2 + v_1\mu + v_0$ Assume that $v_2 \neq 0$. The discriminant is written as

$$d = v_1^2 - 4v_1v_0$$

The derivative is

$$V'(\mu) = 2v_2\mu + v_1$$

which is linear. We can rewrite $V(\mu)$ as

$$V(\mu) = \frac{\{V'(\mu)\}^2 - d}{4v_2}$$

It is impossible to have both d and v_2 negative since this implies $V(\mu)$ is negative.

		d		
v_2		Pos	Neg	0
	Pos	Nbin	NEF-CHS	Gamma
	Neg	Bin	Impossible	

- Bin

  $$V(p) = p - p^2$$

  where v_2 is negative and d is positive
- Geom $\mu = q/p$, $V(\mu) = \mu + \mu^2$
- NEF-CHS

  $$V(\mu) = \mu^2 + 1$$

  $d = -1 < 0$ and $v_2 > 0$.

Number of Parameters:

- 1: δ_μ
- 2: $\mathcal{N}(\mu, \sigma^2)$
- 3: Pois(λ), Gamma
- 4: Bin, NBin, CHS

where the parameter types are Location, Scale, Convolution.

If $y_j \overset{iid}{\sim} f(y_j|\theta)$, for $j = 1, 2, \ldots, n$, then the likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(y_j|\theta)$$

If $\theta \sim g(\theta)\,\mathrm{d}\theta$,

$$g(\theta|y) \propto \mathcal{L}(\theta)g(\theta)\,\mathrm{d}\theta$$

If we have an NEF(n, η), then

$$\underbrace{\prod_{i=1}^{r} e^{\eta y_j - \psi(\eta)}}_{\mathcal{L}(\eta)} \underbrace{e^{\eta\{r_0\mu_0\} - r_0\psi(\eta)}}_{g(\eta)} = e^{\eta(r\bar{y} + r\mu_0 - (r+r_0)\psi(\eta)}$$

Then

$$\bar{y}|\mu \sim NEF\left[\mu, \frac{V(\mu)}{r}\right]$$

If the conjugate prior is

$$g(\eta) = ce^{r_0\eta\mu_0 - r_0\psi(\eta)}$$

where $\mu = \psi'(\eta)$, $\frac{\mathrm{d}\mu}{\mathrm{d}\eta} = \psi''(\eta) = V(\mu)$, and

$$\eta = \int \frac{\mathrm{d}\mu}{V(\mu)} = \frac{\mathrm{d}\eta}{\mathrm{d}\mu} = \frac{1}{V(\mu)}$$

$$\begin{aligned}\psi('eta) &= \int \psi(\eta)\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\,\mathrm{d}\mu \\ &= \int \frac{\mu}{V(\mu)}\,\mathrm{d}\mu\end{aligned}$$

and

$$\tilde{g}(\mu) = ce^{r_0\left(\mu_0\int\frac{\mathrm{d}\mu}{V(\mu)} - \int\frac{\mu}{V(\mu)}\,\mathrm{d}\mu\right)}\frac{1}{V(\mu)} - \int\frac{\mu = \mu_0}{V(\mu)}\,\mathrm{d}\mu$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\log f(x)\right) = \frac{Linear(x)}{Quadratic(x)}$$

all continuous on $\mathbb{R}$,

$$f(s) = e^{\int \frac{L(x)}{Q(x)}\,\mathrm{d}x}$$

Take the t-density.

$$f(x) = \frac{c_0}{(1+x^2)^{c_1}}$$

Then

$$\begin{aligned}(\log f(x))' &= (-c_1 \log(1+x^2))' \\ &= -c_1 \frac{2x}{1+x^2}\end{aligned}$$

If $\mu \sim$ Conjugate with $V(\mu)$ quadratic.

$$\begin{aligned}\mathbb{E}[\mu] &= \int \psi'(\eta) g(\eta)\,\mathrm{d}\eta \\ &= c\int \psi'(\eta) e^{nr_0\mu_0 - r_0\psi(\eta)}\,\mathrm{d}\eta \\ &= c\int e^{\eta r_0\mu_0}\psi'(\eta)e^{-r_0\psi(\eta)}\,\mathrm{d}\eta \\ \mu &\sim CD\left[\mu_0, \frac{V(\mu_0)}{r_0 - v_2}\right]\end{aligned}$$

$$\bar{y} = \frac{\mathrm{Bin}(n,p)}{n} \sim \mathrm{Bin}\left[p, \frac{p(1-p)}{n}\right]$$

$p = \mathrm{Beta}$, $\mu_0 = \mathbb{E}[p]$, then

$$\bar{y} \sim \mathrm{Beta}\left(\mu_0, \frac{\mu_0(1-\mu_0}{r_0 - v_2}\right)$$

$$\begin{aligned}\mathbb{E}[\bar{y}] &= \mathbb{E}[\mathbb{E}[\bar{y}|\mu]] \\ &= \mathbb{E}[\mu] \\ &= \mu_0 \\ \mathrm{Var}(\bar{y}) &= \mathbb{E}\left[\frac{V(\mu)}{n}\right] + \mathrm{Var}(\mu) \\ &= \frac{1}{n}\left[V(\mu_0) + v_2\frac{V(\mu_0)}{n} + \frac{V(\mu_0)}{r_0}\right]\end{aligned}$$

17. October 29th, 2013

Short lecture since the mid-term is next week

Theorem 14. *If X has a bounded distribution and is not constant, then X is not infinitely divisible.*

Proof. WELOG assume $0 \le X \le 1$. Write $X = X_1 + X_2 + \cdots + X_n$ where the X_i's are i.i.d. This gives us that $0 \le X_j \le \frac{1}{n}$ almost surely. This implies

$$\mathrm{Var}(X_j) \le \mathbb{E}\big[X_j^2\big] \le \frac{1}{n^2}$$

which implies

$$\mathrm{Var}(X) \le \frac{1}{n}$$

for all n. □

18. November 5th, 2013

The r-norm is

$$\|X\|_r = (\mathbb{E}[|X|^r])^{1/r}$$

If $X \to X - \mu$ and $r = 2$, then we have

$$\|X\|_2 = \sqrt{\mathbb{E}[(X-\mu)^2]} = \mathrm{SD}(X)$$

Theorem 15 (Cauchy-Schwarz)**.**

$$\mathbb{E}[|XY|] \le \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

This also says

$$|\,\mathrm{Cov}(X,Y)| \le SD(X)SD(Y)$$

so $\rho = \mathrm{Corr}(X,Y) \le 1$.

Proof.

$$\mathbb{E}\big[(Y-\beta X)^2\big] = \mathbb{E}\big[Y^2\big] - 2\beta\mathbb{E}[XY] + \beta^2\mathbb{E}\big[X^2\big]$$

Differentiating with respect to β, we get

$$-2\mathbb{E}[XY] + 2\beta\mathbb{E}\big[X^2\big] = 0$$

so

$$\hat{\beta} = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}$$

Plugging this in, we get

$$0 \le \mathbb{E}\big[Y^2\big] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[X^2]}$$

and rearranging this result gives us Cauchy-Schwarz. □

Theorem 16 (Markov's inequality)**.** *If $Y \ge 0$ almost surely, then*

$$\mathbb{P}(Y \ge 1) \le \mathbb{E}[Y]$$

The more standard result says that

$$\mathbb{P}\left(\frac{|Y|}{a} \le 1\right) \le \mathbb{E}\left[\frac{|Y|}{a}\right] = \frac{1}{a}\mathbb{E}[|Y|]$$

But we can take this to the power of r which gives us

$$\mathbb{P}\left(\frac{|Y|^r}{a^r} \le 1\right) \le \mathbb{E}\left[\frac{|Y|^r}{a^r}\right] = \frac{1}{a^r}\mathbb{E}[|Y|^r]$$

Theorem 17 (Chebyshev)**.** *Now, replace Y with $Y-\mu$. $\mathbb{E}[Y] = \mu$ and $r = 2$. This implies*

$$\mathbb{P}(|Y-\mu| \ge a) \le \frac{\mathbb{E}\big[(Y-\mu)^2\big]}{a} = \frac{\mathrm{Var}(Y)}{a^2}$$

18.1. **Weak Law of Large Numbers.** In the easy case, $X_1, \dots, X_n$ are iid as X. Prove

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \to \mu \equiv \mathbb{E}[X]$$

with probability 1. We prove this using Chebyshev's inequality. For all $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\big(|\bar{X}_N - \mu| > \varepsilon\big) &\le \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} \\ &= \frac{1}{n}\frac{\mathrm{Var}(X)}{\varepsilon^2} \\ &\to 0 \text{ as } n \to \infty\end{aligned}$$

18.1.1. *A more general WLLN.* Suppose now that the X_j are independent, but not identically distributed. The variance is $\mathrm{Var}(X_j) = \sigma_j^2$, $\mathbb{E}[X_j] = \mu$. Chebyshev's inequality gives

$$\begin{aligned}\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) &\le \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} \\ &= \frac{1}{n^2\varepsilon^2}\sum_{j=1}^{n}\mathrm{Var}(X_j)\end{aligned}$$

Now, if $\sigma_j^2 \le M < \infty$ for all j, we have

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \le \frac{nM}{n^2\varepsilon^2} \to 0$$

as $n \to \infty$. What if we have $\sigma_j^2 \sim Mj$? This wouldn't work since it won't converge to 0. Say if we had $\sigma_j^2 \le M\sqrt{j}$?

$$\begin{aligned}\sum_{j=1}^{n} j^{1/2} &\approx \int_1^n x^{1/2}\,\mathrm{d}x \\ &= \frac{2}{3}x^{3/2}\Big|_1^n \\ &\approx \frac{2}{3}n^{3/2}\end{aligned}$$

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j)$$

Covariance Inequality:

$$\mathrm{Cov}(g(Y), h(Y)) \geq 0$$

where g and h are increasing.

19. November 7th, 2013

Example 11 (Inequalities Examples)**.** *Let $Z \sim \mathcal{N}(0,1)$, and we want $\mathbb{P}(|Z| > 3) \approx 0.0027$. We can get this easily in R, but what if we don't know this? What are some quick and easy bounds?*

(1) Markov's inequality says

$$\mathbb{P}(|Z| > 3) \leq \frac{\mathbb{E}[|Z|]}{3} = \frac{\sqrt{2/\pi}}{3} \approx 0.27$$

(2) Alternatively, we can use Chebyshev's inequality.

$$\mathbb{P}(|Z| > 3) = \mathbb{P}\big(Z^2 > 9\big) \leq \frac{\mathbb{E}\big[Z^2\big]}{9} = \frac{1}{9} \approx 0.11$$

(3) Chernoff Bound: Let $t > 0$.

$$\mathbb{P}(|Z| > 3) = 2\mathbb{P}(Z > 3) = 2\mathbb{P}\big(e^{tZ} > e^{3t}\big) \leq \frac{2\mathbb{E}\big[e^{tZ}\big]}{e^{3t}} = 2\frac{e^{\frac{1}{2}t^2}}{e^{3t}} = 2e^{\frac{1}{2}t^2 - 3t}$$

We can optimize this over t, and this give us that $t = 3$ will give us the best bound, 0.022.

(4) Mill's Ratio: for $t > 0$,

$$\mathbb{P}(|Z| > t) \leq 2\frac{\phi(t)}{t}$$

where $\phi(t)$ is the density of Z. The outline of proof comes from the fact that

$$\mathbb{P}(Z > t) = \int_t^\infty \phi(z)\,\mathrm{d}z$$

and then put z/t into the integral, which bounds the probability from above (since $z > t$). This implies that

$$\frac{2\phi(3)}{3} \approx 0.00295$$

Example 12 (Cauchy Schwarz Example)**.** *Let X be a non-negative, integer-valued, discrete random variable with upper bound $\mathbb{P}(X = 0)$ in terms of moments. Write*

$$X = X\mathbb{1}(X > 0)$$

Then

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}(X > 0)] \\ &\leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[\mathbb{1}(X > 0]}\end{aligned}$$

which implies

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$$

or

$$\mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X^2]}$$

Example 13 (Near-birthday Problem)**.** *How many people do you need to have at least a 50% chance of having the same or one-day apart birthday? The answer is 14, compared to 23 for the standard birthday problem. Solving this analytically is difficult, so let us do some approximations by bounding it.*

Let X be the number of near-matches and approximate it by Poisson.

$$X \sim \mathrm{Pois}\left(\binom{n}{2}\frac{3}{365}\right)$$

Let us take

$$\mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X^2]}$$

We can rewrite X as

$$X = I_1 + I_2 + \cdots + I_{\binom{n}{2}}$$

where $I_j \sim \mathrm{Bern}\left(\frac{3}{365}\right)$ which are pairwise independent, but not fully independent. Then the variance of X is

$$\mathrm{Var}(X) = \binom{n}{2}\frac{3}{365}\left(1 - \frac{3}{365}\right)$$

so

$$\mathbb{P}(X=0) \le \frac{\mathrm{Var}(X)}{\mathbb{E}[X^2]} = \frac{\binom{n}{2}\frac{3}{365}\left(1-\frac{3}{365}\right)}{\mathrm{Var}(X)+\left(\frac{n}{2}\frac{3}{365}\right)^2}$$

For $n = 14$ people,

$$\mathbb{P}(X=0) \le 0.573$$

19.1. **Convergence.**

- Convergence almost surely: $X_n \overset{a.s.}{\to} X$ means $\mathbb{P}(X_n \to X) = 1$.
- Convergence in probability: $X_n \overset{\mathbb{P}}{\to} X$ means $\forall \varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$.
- Convergence in distribution: $F_n(x) \to F(x)$ at continuity points of F where $F_n(x)$ is the cdf of X_n and $F(x)$ is the cdf of X. ($X_n \overset{D}{\to} X$ or $X_n \overset{\mathcal{L}}{\to} X$)

Convergence a.s. $\Rightarrow$ Convergence in probability $\Rightarrow$ Convergence in distribution

A trivial counterexample as to why the converse isn't true is if we take $X_n = X_1 \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{N}(0,1) \perp\!\!\!\perp X_1$. Then clearly, the distributions are the same always, but since $X_1 \perp\!\!\!\perp X$, the random variables don't get "close".

Example 14. *Let $X_n = \frac{1}{n}$. We would want $X_n \overset{D}{\to} 0$.*

Example 15.

$$X_n \overset{P}{\to} X \not\Rightarrow X_n \overset{a.s.}{\to} X$$

Let $X_n \overset{iid}{\sim} \mathrm{Bern}(1/n)$, $X = 0$. Let $0 < \varepsilon < 1$. Then

$$\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(X_n = 1) = \frac{1}{n} \to 0$$

However, X_n does not converge to 0 almost surely.

$$\mathbb{P}(\textit{infinitely many 1's}) = 1$$

which is done by using Borel-Cantelli.

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

Example 16. *Suppose that $X_n \overset{a.s.}{\to} X$ and $Y_n \overset{a.s.}{\to} Y$. Then is it true that*

$$X_n + Y_n \overset{a.s.}{\to} X + Y?$$

This happens to be true. What if $X_n \overset{\mathbb{P}}{\to} X$ and $Y_n \overset{\mathbb{P}}{\to} Y$, then does

$$X_n + Y + n \overset{\mathbb{P}}{\to} X + Y?$$

*This requires a bit more work, but it also true. If $X_n \overset{D}{\to} X$ and $Y_n \overset{D}{\to} Y$, then we do **NOT** have that*

$$X_n \to Y_n \overset{D}{\to} X + Y$$

However, there is a special case where this is true. This is given by Slutsky's theorem.

Theorem 18 (Slutsky's Theorem)**.** *If $X_n \overset{D}{\to} X$ and $Y_n \overset{D}{\to} c$ where c is a constant, then*

$$\begin{cases} X_n + Y_n \overset{D}{\to} X + c \\ X_n Y_n \overset{D}{\to} cX \end{cases}$$

19.1.1. *Equivalence of convergence in distribution.*

$$X_n \overset{D}{\to} X \Leftrightarrow \mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$$

for all bounded, continuous h.

Theorem 19 (Skorohod's Theorem)**.** *Let $X_n \overset{D}{\to} X$. Then there exist $X_n^* \sim X_n$ and $X^* \sim X$ on some space, with $X_n^* \overset{a.s.}{\to} X$.*

Idea of proof. Coupling argument: Let $U \sim$ Uniform and

$$X_n^* = F_n^{-1}(U)$$

here we are bringing the X_n^* together using the same U. Then

$$F_n^{-1}(U) \overset{a.s.}{\to} F^{-1}(U)$$

□

An easy corollary is

Theorem 20 (Continuous Mapping Theorem)**.** *If $X_n \overset{D}{\to} X$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \overset{D}{\to} g(X)$.*

Proof. Create X_n^* and X^* as in Skorohod. Let h be bounded and continuous. Then

$$\mathbb{E}[h(X_n)] = \mathbb{E}[h(X_n^*)] \to \mathbb{E}[h(X^*)] = \mathbb{E}[h(X)]$$

where the limit is true by bounded convergence theorem. □

20. November 12th, 2013

Let $X_j \sim [\mu_j, \sigma_j^2]$ be independent (not necessarily identically distributed).

$$S_n = \sum_{j=1}^{n} X_j$$

We standardize this, looking at

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{s_n}$$

where

$$s_n^2 = \sum_{j=1}^{n} \sigma_j^2 = \operatorname{Var}(S_n)$$

In fact,

$$Z_n \sim [0, 1]$$

We question either $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Central Limit Theorem holds if:

(5) iid case
(4) Linear combinations of the iid case
(3) Fourth cumulant condition
(2) Lyapunov condition: $2 + \delta$ moments, $\delta > 0$.
(1) Lindeberg-Feller

Let

$$Y_j = \frac{X_j - \mu_j}{\sigma_j} \sim [0, 1]$$

which gives us

$$Z_n = \frac{\sum_{j=1}^{n} \sigma_j Y_j}{s_n}$$

and we wish to prove that this goes to $\mathcal{N}(0,1)$.

Continuity Theorem (From Chapter 11): This says that all we have to show is that the characteristic function of Z_n goes to the characteristic function of a $\mathcal{N}(0,1)$ random variable.

$$\phi_n(t) = \mathbb{E}\left[e^{itZ_n}\right] \to \mathbb{E}\left[e^{itZ}\right] = e^{-\frac{1}{2}t^2} \leq 1$$

Theorem 21 ("Continuity Theorem")**.**

$$\phi_n(t) \to \phi(t) \Leftrightarrow \textit{Convergence in Law}$$

for all $t \in \mathbb{R}$.

If we take the log of the characteristic function

$$\begin{aligned}\log \mathbb{E}\left[e^{itZ_n}\right] &= \log \mathbb{E}\left[e^{it\frac{1}{s_n}(\sigma_1 Y_1 + \sigma_2 Y_2 + \cdots \sigma_n Y_n)}\right] \\ &= \sum_{j=1}^{n} \log \mathbb{E}\left[e^{it\sigma_j \frac{Y_j}{s_n}}\right]\end{aligned}$$

Let

$$u_n = \max_{1 \leq j \leq n} \frac{\sigma_j^2}{s_n^2}$$

to go to zero as $n \to \infty$. This is called the uniformly asymptotically negligible condition.

$$\mathbb{E}\left[e^{itX}\right] = \phi(t)$$

Expanding this out, we get

$$\mathbb{E}\left[1 + itx + \frac{1}{2!}(itx)^2 + \cdots\right]$$

If $\mathbb{E}[X] = 0$, then

$$\phi(t) \approx 1 - \frac{1}{2}t^2 \operatorname{Var}(X)$$

Call $\psi(t) = \log \phi(t)$, which is the complex cumulant generating function. If t is small, $\psi(t) \approx -\frac{1}{2}t^2 \operatorname{Var}(X)$.

$$\begin{aligned}\psi_n(t) &= \log \mathbb{E}\left[e^{itZ_n}\right] \\ &= \sum_{j=1}^{n} \log\left(1 - \frac{t^2}{2}\frac{\sigma_j^2}{s_n^2} + \text{small } t^3\right) \\ &= \sum_{j=1}^{n}\left(-\frac{t^2}{2}\right)\frac{\sigma_j^2}{s_n^2} + \text{small } s_n \\ &= -\frac{t^2}{2} + \text{something small as } n \to \infty\end{aligned}$$

Chf result:

$$e^{ix} = \cos(x) + i\sin(x)$$

So

$$\phi(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]$$

This immediately shows that the characteristic function is bounded, since

$$\begin{aligned}|\mathbb{E}\left[e^{itX}\right]| &\leq \mathbb{E}\left[|e^{itX}\right] \\ &= \mathbb{E}\left[\sqrt{\cos^2(tX) + \sin^2(tX)}\right] \\ &= 1\end{aligned}$$

so $|\phi(t)| \leq 1$ and the characteristic function exists for all random variables and $\phi(0) = 1$.

If $\mathbb{E}[|X|^r] < \infty$, then

$$\phi(t) = 1 + it\mathbb{E}[X] - \frac{1}{2!}t^2(\mathbb{E}[X])^2 + \cdots + o(|t|^r)$$

as $t \to \infty$. where the little o-notation means

$$\frac{o(t)}{t} \to 0$$

as $t \to \infty$.

Theorem 22. *This is found in page 171 in the book. $Y \sim [0,1]$. ccgf:*

$$\psi(y) = \log \mathbb{E}\left[e^{itY}\right] = -\frac{t^2}{2}\left(1 + z_2(t)\mathbb{E}\left[Y^2 \min(1, |tY|)\right]\right)$$

if $|t| \leq \frac{1}{2}$ and $|z_2(t)| \leq C < \infty$.

Let

$$m(t) = \mathbb{E}\left[Y^2 \min(1, |tY|)\right]$$

Then from before,

$$\begin{aligned}\psi_n(t) &= \log \mathbb{E}\left[e^{itZ_n}\right] \\ &= \sum_{j=1}^{n} -\frac{t^2}{2}\frac{\sigma_j^2}{s_n^2}\left(1 + z_n\mathbb{E}\left[Y_j^2 \min\left(1, |tY|\frac{\sigma_j}{s_n}\right)\right]\right)\end{aligned}$$

if $|t|\frac{\sigma_j}{s_n} < \frac{1}{2}$ for $j = 1, \ldots, n$. However, by UAN if we take n large enough, we can force this to be approximation to hold.

$$\psi_n(t) = -\frac{t^2}{2} - \frac{t^2}{2}C\sum_{j=1}^{n}\frac{\sigma_j^2}{s_n}\mathbb{E}\left[Y_j^2 \min\left(1, |t||Y_j|\sqrt{u_n}\right)\right]$$

where C is the bound for z_i. and the entire sum is called the fundamental bound when $t = 1$. We wish to show that the sum goes to zero as $n \to \infty$.

Pg.171 in the notes:

(b) Lindeberg-Feller: For all $\varepsilon > 0$

$$\sum_{j=1}^{n}\mathbb{E}\left[\frac{\sigma_j^2}{s_n^2}Y_j^2\mathbb{1}\{\frac{\sigma_j}{s_n}|Y_j| > \varepsilon\}\right] \to 0$$

as $n \to \infty$.

(c) Lyapunov:

$$\sum_{j=1}^{n}\mathbb{E}\left[\left(\frac{\sigma_j|Y_j|}{s_n}\right)^r\right] \to 0$$

for $r > 2$.

(d) UAN (Uniformly Asymptotically Neglibible) holds and $K_4(Z_n) \to K_4(Z) = 0$.
(e) Linear combinations of iid Y_j.

$$Z_n = \frac{\sum_{j=1}^n \sigma_j Y_j}{\sqrt{\sum_{j=1}^n a_j^2}} \to \mathcal{N}(0,1)$$

if

$$\frac{\max(a_j^2)}{\sum_{j=1}^n a_j^2} \to 0$$

as $n \to \infty$.

What if Y_j are iid and $\sigma_j = 1$? Then

$$Z_n = \frac{\sum_{j=1}^n Y_j}{\sqrt{n}} = \sqrt{n}\bar{Y} \to Z$$

UAN

21. November 14th, 2013

Carl's lecture: Central Limit Theorem:

$$Z_n = \frac{\sum_{j=1}^n X_j}{s_n} = \frac{\sum_{j=1}^n \sigma_j Y_j}{s_n}$$

where $X_j \sim [0, \sigma^2]$ and $Y_j = \frac{1}{\sigma_j} X_j$.

If $X_j \sim \text{Bern}(p_j)$, do we have that

$$\sum_{j=1}^n \frac{X_j - p_j}{\sqrt{p_j q_j}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)?$$

We must find the fourth cumulant, we first find the third cumulant by

$$K_3(X_j) = \left[\frac{\mathrm{d}}{\mathrm{d}p_j}(p_j - p_j^2)\right] = (1 - 2p_j)(p_j - p_j^2)$$

We differentiate again to get the fourth cumulant

$$\begin{aligned} K_4(X_j) &= (p_j q_j)\left\{1 - 6p_j + 6p_j^2\right\} \\ &= (p_j q_j)\left\{1 - 6p_j q_j\right\} \end{aligned}$$

Our UAN condition is

$$\frac{\max(p_j q_j)}{\sum_{j=1}^n p_j q_j} \to 0$$

If $s_n^2 = \sum_{j=1}^n p_j q_j \to \infty$.

$$\begin{aligned} K_4(Z_n) &= K_4\left(\sum_{j=1}^n \frac{X_j - p_j}{s_n}\right) \\ &= \frac{1}{s_n^4}\sum_{j=1}^n (p_j q_j - 6(p_j q_j)^2) \\ &= \underbrace{\frac{1}{\sum_{j=1}^n p_j q_j}}_{\text{UAN}} - 6\frac{\sum_{j=1}^n p_j q_j}{s_n^4} \\ &\to 0 \end{aligned}$$

Problem 13.5.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$\hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

To make things easier, let $\bar{x} = 0$.

$$\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2} = \beta_1 + \frac{\sum y_i \varepsilon_i}{\sum x_i^2}$$

and then

$$Z_n = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum x_i^2}}$$

Theorem 23 (Berry Esseen Theorem)**.** *$X_1, \dots, X_n \overset{iid}{\sim} [\mu, \sigma^2]$ and*

$$Z_n = \frac{\sum(X_i - \mu)}{\sigma\sqrt{n}}$$

then

$$\sup_x |F_n(x) - \Phi(x)| \leq .4784 \frac{\mathbb{E}\left[\left|\frac{X-\mu}{\sigma}\right|^3\right]}{\sqrt{n}}$$

22. November 19th, 2013

$M_0, M_1, M_2, \dots$ is martingale with respect to $X_0, X_1, \dots$ if

(1) $\mathbb{E}[|M_n|] < \infty$
(2) $M_n \in \sigma(X_0, \dots, X_n)$
(3) $\mathbb{E}[M_{n+1}|X_0, \dots, X_n] = M_n$. Equivalently, we can write $\mathbb{E}[M_{n+1} - M_n|X_1, \dots, X_n] = 0$.

Martingales are supposed to capture the idea of a "fair game".

- Supermartingale: We replace $=$ with $\leq$
- Submartingale: We replace $=$ with $\geq$

Definition 20 (Filtration)**.** *We have a filtration if we have*

$$\begin{aligned} F_0 &= \sigma(X_0) \\ &\subseteq F_1 = \sigma(X_0, X_1) \\ &\subseteq F_2 = \sigma(X_0, X_1, X_2) \\ &\vdots \end{aligned}$$

Fact: If (M_n) is a martingale with respect to (X_n), then (M_n) is a martingale with respect to (M_n), that is it's a martingale with respect to itself.

Proof.

$$\begin{aligned} \mathbb{E}[M_{n+1}|M_0, M_1, \dots, M_n] &= \mathbb{E}[\mathbb{E}[M_{n+1}|M_0, \dots, M_n, X_0, \dots, X_n]\,|M_0, \dots, M_n] \\ &= \mathbb{E}[M_{n+1}|X_0, \dots, X_n] \\ &= M_n \end{aligned}$$

□

Example 17.

(1) Simple symmetric random walk on integers.

$$S_0 = 0, \qquad S_n = X_1 + \cdots + X_n$$

where X_j are iid random signs.

$$\mathbb{E}[S_{n+1}|X_1, \dots, X_n] = S_n$$

(2) Pólya urn (sampling with double replacement).
Suppose we start with one red ball and one blue ball. Let M_n be the proportion of balls that is red. ($M_0 = \frac{1}{2}$). This is, in fact, a martingale. Our question of interest is what happens to M_n as $n \to \infty$? We actually have

$$M_n \overset{n\to\infty}{\longrightarrow} U \sim \mathcal{U}(0, 1)$$

(3) Likelihood Ratios
If $X_1, X_2, \dots$ are iid, do these data come from pdf f or pdf g?

$$M_n = \frac{f(X_1)}{g(X_1)} \frac{f(X_2)}{g(X_2)} \cdots \frac{f(X_n)}{g(X_n)}$$

This is a martingale if the true distribution of X_n is g. If $X_n \sim f$, then we have a submartingale.

(4) Branching Process

$$X_{n+1} = \sum_{n=1}^{X_n} C_{nj}$$

where C_{nj} is equal to the number of children the jth amoeba in the nth generation has. What happens to X_n as $n \to \infty$? Let $M = \mathbb{E}[C_{nj}]$. Note that

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n]\,\mu$$

Then $M_n = \frac{X_n}{\mu^n}$ is a martingale.

(5) S_n random walk, look at S_n^2. This thing is a bit more complicated. What's interesting it that $S_n^2 - n$ is a martingale.

Definition 21 (Brownian Motion)**.** *We have $(B_t)_{t\geq 0}$*

(1) Continuous sample paths
(2) $B_t \sim \mathcal{N}(0,t)$
(3) Martingale
(4) Markov process

Theorem 24 (Martingale Convergence Theorem)**.** *If (M_n) is a submartingale with*

$$\sup_n \mathbb{E}\left[M_n^+\right] < \infty$$

where $M_n^+ = \max(M_n, 0)$, then $M_n \xrightarrow{a.s.} M_\infty$ for some M_∞ with $\mathbb{E}[|M_\infty|] < \infty$.

Corollary. *If you have a non-negative supermartingale, then it converges.*

For the branching process example, note that $M_n \geq 0$, then we know by martingale convergence theorem that

$$M_n = \frac{X_n}{\mu^n} \xrightarrow{a.s.} M_\infty$$

If $\mu < 1$, then $X_n \xrightarrow{a.s.} 0$ or also

$$\mathbb{P}(X_n = 0 \text{ for all } n, \text{ eventually}) = 1$$

If $\mu = 1$, then $M_n = X_n \xrightarrow{a.s.} M_\infty$.

We can show by conditional expectation that $\mathbb{E}[M_n] = \mathbb{E}[M_0]$. Does this imply that $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ for T random (non integer valued random variable). There are two main counterexamples to think about.

(1) "psychic powers"
e.g. Random walk starting at $\$1 (M_0 = 1)$.
(2) S_n random,

$$T = \inf\{n : S_n = 10^{12}\}$$

Then

$$\mathbb{E}[S_T] = 10^{12} \neq \mathbb{E}[S_0] = 0$$

23. November 21st, 2013

23.1. **Stopping Times.** $T : \Omega \to \{0, 1, 2, \ldots, \infty\}$ is a **stopping time** relative to (X_n) if the event

$$\{T \leq n\} \in \sigma(X_0, \ldots, X_n) = \{\{(X_0, \ldots, X_n) \in B\}, B \text{ Borel in } \mathbb{R}^{n+1}\}$$

The events $\{T \leq n\}$ and $\{T = n\}$ are equivalent here.

Intuitively,

$$I(T \leq n) = g(X_0, X_1, \ldots, X_n)$$

Theorem 25. *A stopped martingale is a martingale, i.e. if (M_n) is a martingale with respect to (X_n) and T is a stopping time with respect to (X_n), then $M_{\min(T,n)}$ is a martingale. ($M_{\min(T,n)}(\omega) = M_{\min(T(\omega),n)}(\omega)$)*

Example 18 (Simple Symmetric Random Walk S_n)**.** *Fix a positive integer b. $T = \inf\{n : S_n = b\}$ is a stopping time. It follows that $M_n = S_{\min(T,n)}$ is a martingale. Note that $M_n \leq b$. By the Martingale Convergence Theorem,*

$$M_n \xrightarrow{a.s.} M_\infty$$

for some M_∞. This implies that $\mathbb{P}(T < \infty) = 1$.

Also, we have

$$\mathbb{E}[S_T] = b \neq \mathbb{E}[S_0] = 0$$

Theorem 26 (Optional Stopping Theorem)**.** *Let M_n be a martingale and T a stopping time. Then*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0]$$

if any of the following hold:

(1) $T \overset{a.s.}{\leq} n$, where n is some constant. (Bounded time)
(2) $|M_n| \overset{a.s.}{\leq} c$ and $\mathbb{P}(T < \infty) = 1$ (Bounded space)
(3) $|M_n - M_{n-1}| \overset{a.s.}{\leq} c$, $\mathbb{E}[T] < \infty$.

Example 19 (Back to our original example)**.** *We have that $\mathbb{E}[T] = \infty$ by the third condition of the Optional Stopping Theorem.*

Proof of Optional Stopping Theorem Part 1.

$$\begin{aligned} M_T &= M_0 + \sum_{j=1}^{T} (M_j - M_{j-1}) \\ &= M_0 + \sum_{j=1}^{n} (M_j - M_{j-1}) I(T \geq j) \end{aligned}$$

Now taking expectation of each term in the sum, and using the fact that

$$\begin{aligned} I(T \geq j) &= 1 - I(T < j) \\ &= 1 - I(T \leq j-1) \end{aligned}$$

we have

$$\begin{aligned} \mathbb{E}[(M_j - M_{j-1}) I(T \geq j)] &= \mathbb{E}[\mathbb{E}[(M_j - M_{j-1}) I(T \geq j) | X_0, \ldots, X_{j-1}]] \\ &= \mathbb{E}[I(T \geq j) \mathbb{E}[M_j - M_{j-1} | X_0, \ldots, X_{j-1}]] \\ &= \mathbb{E}[I(T \geq j) \times 0] \\ &= 0 \end{aligned}$$

Therefore,

$$\mathbb{E}[M_T] = \mathbb{E}[M_0]$$

□

Proof of Optional Stopping Theorem Part 2. "Truncation Argument". Assume $|M_n| \overset{a.s.}{\leq} c$ and $\mathbb{P}(T < \infty) = 1$. Let

$$T_n = \min(T, n)$$

Note that this is a bounded stopping time. If we take limits,

$$\begin{aligned} \lim_{n \to \infty} T_n &\overset{a.s.}{=} T \\ \lim_{n \to \infty} M_{T_n} &\overset{a.s.}{=} M_T \end{aligned}$$

We know from the first condition of the Optional Stopping Theorem that

$$\mathbb{E}[M_{T_n}] = \mathbb{E}[M_0]$$

let $n \to \infty$ and then use Bounded Convergence Theorem. □

23.2. **Gambler's Ruin.** Suppose Gambler A starts with $\$a$ and Gambler B starts with $\$b$. Let $p = \frac{1}{2}$.

$$T = \inf\{n : S_n = -a \text{ or } b\}$$

$\mathbb{P}(T < \infty) = 1$. That is when S_n reaches $-a$ or b, then the corresponding gambler has won.

The probability that Gambler A wins is

$$\mathbb{P}(A \text{ wins}) = \frac{a}{a+b}$$

This was proved in STAT 110 with difference equations. Let us apply the Optional Stopping Theorem.

$$\begin{aligned} 0 &= \mathbb{E}[S_T] \\ &= (-a)\mathbb{P}(A \text{ loses}) + b\mathbb{P}(A \text{ wins}) \end{aligned}$$

which implies that

$$\mathbb{P}(A \text{ wins}) = \frac{a}{a+b}$$

This is easy to get, even in 110, but what is a natural question we would want to ask that is more difficult to solve in STAT 110 methods?

We have that

$$\mathbb{E}[T] = ab$$

Clearly, we would be using a martingale, but which one? Note that $S_n^2 - n$ is a martingale. We can get this result by looking at $\mathbb{E}\big[S_T^2 - T\big]$

Now let $p \neq \frac{1}{2}$ and $q = 1 - p$. Let

$$M_n = \left(\frac{q}{p}\right)^{S_n}$$

Note that

$$M_1 = \begin{cases} \frac{q}{p} & \text{with prob. } p \\ \frac{p}{q} & \text{with prob. } q \end{cases}$$

which has expected value 1.

23.2.1. *Maximal Inequality for Margtinales.* Let $(M_n)_{n=0}^{\infty}$ be a positive martingale and $a > 0$.

$$\mathbb{P}\left(\max_{1\le j\le n} M_j \ge a\right) \le \frac{\mathbb{E}[M_0]}{a}$$

Note that this is very similar to Markov's inequality. Furthermore, the right hand side does not even depend on n.

Proof. Let $T = \min(\inf\{j : M_j \ge a\}, n)$. If

$$\max_{0\le j\le n} M_j \ge a$$

then $M_T \ge a$. Therefore,

$$\begin{aligned} \mathbb{P}\left(\max_{1\le j\le n} M_j \ge a\right) &\le \mathbb{P}(M_T \ge a) \\ &\le \frac{\mathbb{E}[M_T]}{a} \quad \text{by Markov's inequality} \\ &= \frac{\mathbb{E}[M_0]}{a} \quad \text{by the Optional Stopping Theorem} \end{aligned}$$

□

23.3. **Say "Red" Problem.** The Say "Red" Problem is a game where you are dealt cards until you say stop. When you say stop, if the next card is red, you win. If it's black, you lose. What's a good strategy for this?

Let M_n be the proportion of remaining cards that are red after n cards have been dealt. $(0 \le n \le 51)$. Again, $(M_n)_0^{51}$ is a martingale. Let T be a stopping rule and $\mathbb{E}[M_T] = \mathbb{E}[M_0] = \frac{1}{2}$. So if you are looking at the game from the beginning, there is no best strategy. (However, during the game, the probabilities may change).

There is also an exchangeability argument. Say you have T cards revealed and $52 - T$ cards remaining. If you switch any of the unrevealed cards, it does not make any difference, so say you just take the last one. This one has a probability of $1/2$ of being red or black, so it does not matter.

23.4. **ABRACADABRA.** Let T be the time until this appears.

$$\mathbb{E}[T] = 26^{11} + 26^4 + 26$$

The idea for this is to imagine an infinite sequence of gamblers. Think of martingales as a "fair game", that is the gamblers' expected gains/losses are equal to 0. Then apply the Optional Stopping Theorem to this and the result will pop out.

24. November 26th, 2013

NOT IN THE BOOK
Review Section 1 Dec 5. 4pm - 6pm. Office Hours same day 1-2 pm.

24.1. **Coupling Inequality.** One widely used metric to compare probability distributions is the total variation distance

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|$$

Theorem 27 (Coupling Inequality)**.** $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) \le \mathbb{P}(X \ne Y)$

Proof.

$$\begin{aligned} d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) &= \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\ &= \sup_A |\cancel{\mathbb{P}(X \in A, X = Y)} + \mathbb{P}(X \in A, X \ne Y) - \cancel{\mathbb{P}(Y \in A, X = Y)} - \mathbb{P}(Y \in A, X \ne Y)| \\ &= \sup_A |\mathbb{E}[I_{X\ne Y}(I_{X\in A} - I_{Y\in A})]| \\ &\le \sup_A \mathbb{E}[I_{X\ne Y}] \\ &= \mathbb{P}(X \ne Y) \end{aligned}$$

□

Suppose we have two Markov Chains, $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$. If there is a random time T such that $X_t = Y_t$ for all $t \ge T$,

$$d_{TV}(\mathcal{L}(X_t), \mathcal{L}(Y_t)) \le \mathbb{P}(X_t \ne Y_t) \le \mathbb{P}(T > t)$$

Example 20 (Deck of cards (My own notes))**.** *Each shuffling of a deck (52! of them) is a node in the Markov Chain. The X's start from an ordered deck. The Y's are an already shuffled deck.*

Let's bound $\mathbb{P}(T > t)$. This is analogous to the toy problem, where there are n types of toys to collect and how long it takes to collect all of them.

$$\begin{aligned}\mathbb{P}(T > t) &\leq \mathbb{P}(\text{after } t \text{ steps, at least one card has not been named}) \\ &\leq n\left(1 - \frac{1}{n}\right)^t \\ &\leq ne^{-t/n}\end{aligned}$$

so let $t = n\log(n + c) \leq e^{-c}$

Example 21 (Stochastic Domination)**.** *We have $X \preceq Y$ if $F_X(a) \geq F_Y(a)$. The claim is that*

$$X \preceq Y \Leftrightarrow \exists X^*, Y^* \text{ with } X^* \sim X, Y^* \sim Y, X^* \preceq Y^*$$

Use the Probability Integral Transform. $X^* = F_X^{-1}(U), Y^* = F_Y^{-1}(U)$. $U \sim$ Uniform.

24.1.1. *Socks in a Drawer Problem.* There are n pairs of socks, which are all different, shuffled, in a drawer in the dark. On average, how many socks do you need to draw to get a pair?

Let N be the number of socks. We wish to find $\mathbb{E}[N]$. Clearly, this is a discrete problem. This is difficult to do, so the idea is to convert this to a continuous time problem. Imagine that we have $2n$ iid Uniform random variables.

X_i will be the time between the ith sock and the $(i-1)$th sock. The Nth sock is picked at time T.

$$T = \sum_{i=1}^{N} X_i$$

N is independent of the X_i's since the ordering of the socks do depend on the X_i's.

$$\mathbb{E}[T] = \mathbb{E}[N]\,\mathbb{E}[X_1]$$

What's the distribution of the X_i's in terms of representation?
We have $U_1, \ldots, U_{2n}$ which are $2n$ Uniform order statistics.

$$U_{(j)} = \frac{Y_1 + \cdots + Y_j}{Y_1 + \cdots + Y_{2n+1}}$$

where the Y_i's are iid Expo.

$$X_1 \sim \text{Beta}(1, 2n - 1 + 1)$$

Therefore,

$$\mathbb{E}[T] = \mathbb{E}[N]\,\frac{1}{2n+1}$$

So how do we get $\mathbb{E}[T]$ now? Let $T = \min(T_1, \ldots, T_n)$, where T_j is the time at which sock pair j is complete. The key thing to note here is that $T_1, \ldots, T_n$ are independent. Consider one of them.

$$T_1 = \max(T_{11}, T_{12}) \sim \text{Beta}(2, 1) \sim \sqrt{U}$$

where U is Uniform.

$$T \sim \min(\sqrt{U_1}, \ldots, \sqrt{U_n}) = \sqrt{\min(U_1, \ldots, U_n)} \sim \sqrt{\text{Beta}(1, n)}$$

Then

$$\mathbb{E}[T] = n\int_0^1 x^{3/2}(1-x)^n \frac{\mathrm{d}x}{x(1-x)}$$

We know that $\Gamma(1/2) = \sqrt{\pi}$, so $\Gamma(3/2) = \frac{1}{2}\sqrt{\pi}$.

$$\mathbb{E}[T] = \frac{n\Gamma(3/2)\Gamma(n)}{\Gamma(3/2+n)}$$

which implies

$$\mathbb{E}[N] = \frac{(2^n n!)^2}{(2n)!} \approx \sqrt{\pi n}$$

Say $X \sim \text{Mult}_k(n,p)$, $p = (p_1, \ldots, p_k)$ and

$$\sum_{j=1}^{k} p_j = 1$$

We also have $\sum_{j=1}^{k} X_j = n$. If $n = 1$,

$$\mathbb{P}(X_j = x_j) = \begin{cases} x_j = 1 & p_j \\ x_j = 0 & q_j = 1 - p_j \end{cases}$$

and $X_j \sim \text{Bern}(p_j)$. We can write this as

$$p_j^{x_j}(1-p_j)^{1-x_j}$$

In the "Bernomial/Categorial" case ($n = 1$), we have

$$\mathbb{P}\left(X = x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}\right) = \prod_{j=1}^{k} p_j^{x_j}$$

If $n \geq 2$, then we have

$$\binom{n}{x_1, \ldots, x_k} \prod_{j=1}^{k} p_j^{x_j}$$

Is this a NEF? Yes, since

$$\prod_{j=1}^{k} p_j^{x_j} = \exp\left(\sum_{j=1}^{k} x_j \log p_j\right)$$

The expected value is

$$\mathbb{E}[X_j] = \mathbb{E}[\text{Bin}(n, p_j)] = np_j$$

The covariance of two of these (if $n = 1$) is

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] \\ &= 0 - p_1 p_2 \end{aligned}$$

The first term is 0 because only one of these can happen.

$$\text{Cov}(X) = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ & p_2(1-p_2) & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & p_k(1-p_k) \end{bmatrix}$$

If $X \sim \text{Mult}_k(n,p)$, then

$$X \sim [np, n(D_k - pp)]$$

so

$$Y = a'X \sim \left[na'p, na'D_k a - (a'p)^2\right]$$

25.1. **Fisher's Trick.** Let $Y_1, \ldots, Y_k$ be independent with $Y_j \sim \text{Pois}(\lambda_j)$. Then

$$(Y_1, \ldots, Y_k) | Y_1 + \cdots + Y_k = n \sim \text{Mult}_k(n, \vec{p})$$

with

$$p_j = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_k}$$