MATH 180A (Lecture A00)

mathweb.ucsd.edu/~ynemish/teaching/180a

Today: Correlation. Markov's inequality. Central Limit Theorem

Week 10:

- Homework 7 due Sunday, March 19
- Office hours next week

  YN: Monday, 1-4 PM, AP&M 6321

  SQ: Tuesday, 2-4 PM, HSS 5056

  TG: Tuesday, 5:30-6:30 PM, HSS 4086A

Correlation

Covariance is not particularly good for evaluating the strength of the dependence:

- suppose that $\mathrm{Cov}(X,Y) = 1$, then $\mathrm{Cov}(10X, 10Y) = 100$, but the dependence between X and Y is the same as dependence between $10X$ and $10Y$.

Solution: normalize covariance $\longrightarrow$ correlation

Def. Let X, Y be random variables, $\mathrm{Var}(X) < \infty$, $\mathrm{Var}(Y) < \infty$

The correlation (coefficient) of X and Y is given by

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

Properties of correlation

Thm Let X, Y be random variables, $Var(X) < \infty$, $Var(Y) < \infty$

- $Corr(aX+b, Y) = \frac{a}{|a|} Corr(X,Y)$
- $-1 \le Corr(X,Y) \le 1$
- $Corr(X,Y) = 1$ if and only if $Y = aX + b$, $a > 0$
- $Corr(X,Y) = -1$ if and only if $Y = aX + b$, $a < 0$

Example Let X, Y be random variables satisfying

$$E(X) = 2,\ E(Y) = 1,\ E(X^2) = 5,\ E(Y^2) = 10,\ E(XY) = 1$$

(a) Compute $Corr(X,Y)$

$Var(X) = 5 - 2^2 = 1$, $Var(Y) = 9$, $Cov(X,Y) = 1 - 2 \cdot 1 = -1$

$$Corr(X,Y) = \frac{-1}{\sqrt{1} \cdot \sqrt{9}} = -\frac{1}{3}$$

(b) Find $c \in \mathbb{R}$ such that X and $X + cY$ are uncorrelated.

Moment generating function of a sum of indep. RVs

Def. (Convolution of distributions - Section 7)

Let X and Y be random variables. Then the distribution of $X+Y$ is called the convolution of the distributions of X and Y.

If X and Y are continuous and f_X and f_Y are their PDFs then the PDF of $X+Y$ is given by

$$f_{X+Y}(s) = f_X * f_Y(s) = \int_{-\infty}^{+\infty} f_X(x) f_Y(s-x)\,dx = \int_{-\infty}^{+\infty} f_X(s-y) f_Y(y)\,dy$$

(similar formula for discrete random variables)

If X and Y are independent, it may be easier to compute the MGF

Moment generating function of a sum of indep. RVs

Let X, Y be independent random variables.

Then the MGF of $X+Y$ is

$$E\left(e^{t(X+Y)}\right) =$$

$$M_{X+Y}(t) =$$

1) $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, independent.
   Distribution of $X+Y$?

2) $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, independent.
   Distribution of $X+Y$?

Estimating tail probabilities

Suppose that $X \geq 0$, $E(X) < \infty$. What can we say about $P(X \geq c)$ for $c > 0$?

Thm (Monotonicity of expectation)

- If $P(Z \geq 0) = 1$, then $E(Z) \geq 0$
- If $P(X \geq Y) = 1$, then $E(X) \geq E(Y)$

Markov's inequality:

If X is a nonnegative random variable a.s. (i.e. $P(X \geq 0) = 1$), then for any $c > 0$ $\quad P(X \geq c) \leq \dfrac{E(X)}{c}$

Proof. $X = X \cdot 1 \geq X \cdot \mathbb{1}_{\{X \geq c\}} \geq c\, \mathbb{1}_{\{X \geq c\}} \Rightarrow X \geq c\, \mathbb{1}_{\{X \geq c\}}$

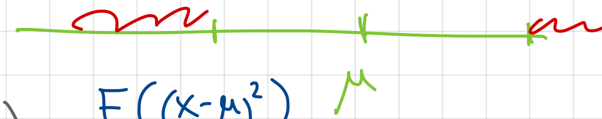
$$\overset{\text{monot.}}{\Rightarrow} E(X) \geq E(c\, \mathbb{1}_{\{X \geq c\}}) = c\, E(\mathbb{1}_{\{X \geq c\}}) = c\, P(X \geq c)$$

Estimating tail probabilities: Chebyshev's inequality

Chebyshev's inequality:

If $E(X)=\mu$, $Var(X)=\sigma^2$, then for any $c>0$

$$P(|X-\mu|\geq c)\leq \frac{\sigma^2}{c^2}$$



Proof. $P(|X-\mu|\geq c) = P((X-\mu)^2\geq c^2) \leq \dfrac{E((X-\mu)^2)}{c}$

In particular, $P(X-\mu\geq c)\leq \dfrac{\sigma^2}{c^2}$, $P(X-\mu\leq -c)\leq \dfrac{\sigma^2}{c^2}$

Remark. Markov/Chebyshev inequalities are sometimes useful, but not always.

Let $X\sim Ber(p)$. $P(X\geq 1) = P(X\geq 0.01) = P(X=1) = p$

Markov's inequality: $P(X\geq 1)\leq \dfrac{E(X)}{1} = p$, $P(X\geq 0.01) = \dfrac{E(X)}{0.01} = 100\cdot p$

Estimating tail probabilities: Chebyshev's inequality

Example Suppose $X \sim Exp(\frac{1}{2})$. Estimate $P(X \geq 6)$

$E(X) = 2$, $Var(X) = 4$

- Markov: $P(X \geq 6) \leq \frac{2}{6} = \frac{1}{3}$
- Chebyshev: $P(X \geq 6) = P(X - 2 \geq 4) \leq \frac{4}{4^2} = \frac{1}{4}$
- Exact value: $P(X \geq 6) = e^{-\frac{1}{2}\cdot 6} = e^{-3} \approx 0.05$

Example X = amount of money earned by a food truck daily.

From past experience we know $E(X) = 5000$

- Markov: $P(X \geq 7000) \leq \frac{E(X)}{7000} = \frac{5000}{7000} = \frac{5}{7}$

Suppose that we additionally know that $Var(X) = 4500$

- Chebyshev: $P(X \geq 7000) = P\left(X - 5000 \geq 2000\right) \leq \frac{4500}{(2000)^2} \approx 0.001$

Weak Law of Large Numbers

Thm Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $\mu, \sigma^2 \in \mathbb{R}$

Let $S_n = X_1 + \cdots + X_n$. Then for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left| \frac{S_n}{n} - \mu \right| < \varepsilon \right) = 1 \quad \left(\frac{S_n}{n} \text{ converges to } \mu \text{ in probability} \right)$$

Proof. Enough to show that $\lim_{n\to\infty} P\left(\left| \frac{S_n}{n} - \mu \right| > \varepsilon \right) = 0$ (complement)

$$E\left(\frac{S_n}{n}\right) = \mu \quad (\text{by linearity}), \quad Var\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

By Chebyshev's inequality

$$P\left(\left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{Var\left(\frac{S_n}{n}\right)}{\varepsilon} = \frac{\sigma^2}{n\varepsilon} \to 0, \quad n \to \infty \quad \blacksquare$$

CENTRAL LIMIT THEOREM (CLT)

<u>Thm</u>. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Let $S_n = X_1 + \cdots + X_n$

Then for any $a, b \in \mathbb{R}$, $a < b$

$$\lim_{n \to \infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

- exactly the same statement as for $X_i \sim Ber(p)$ (Section 4)
- only need the first two moments
- CLT describes the fluctuations of S_n around $n\mu$, which are of order $\sqrt{n}$
- CLT is a family of theorems (there may be different or more general assumptions about (X_i))

Applications of the CLT

Every morning you wake up and start tossing a fair coin until the first Tails comes up. Estimate the probability that in the first 100 days of this experiment you toss the coin at least 220 times (in total).

Denote $X_i = \#$ tosses on day i, $S_{100} = \sum_{i=1}^{100} X_i = \#$ tosses after 100 days

$X_i \sim \text{Geom}(\frac{1}{2})$, $E(X_i) = 2$, $Var(X_i) = 2$

$$P(S_{100} \geq 220) = P\left(\frac{S_{100} - 200}{10 \cdot \sqrt{2}} \geq \frac{220-200}{10 \cdot \sqrt{2}}\right) = P\left(\frac{S_{100} - 200}{\sqrt{2 \cdot 100}} \geq \sqrt{2}\right)$$

$$\approx 1 - \Phi(\sqrt{2})$$

The only relevant information is expectation, variance, independence

If $Y_1, Y_2, \ldots$ are i.i.d with $E(Y_i) = 2$, $Var(Y_i) = 2$, then $P\left(\sum_{i=1}^{100} Y_i \geq 220\right) \approx 1 - \Phi(\sqrt{2})$

Proof of the CLT

Thm (Continuity theorem for the MGF)

Let X be a random variable with continuous CDF.

Suppose that the MGF of X $M_X(t)$ is finite on $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$.

Suppose that $Y_1, Y_2, \ldots$ be a sequence of random variables such that

$$\lim_{n\to\infty} M_{Y_n}(t) = M_X(t) \text{ for all } t \in (-\varepsilon, \varepsilon).$$

Then for any $a \in \mathbb{R}$ $\lim_{n\to\infty} P(Y_n \le a) = P(X \le a)$.

In particular, if $X \sim N(0,1)$, $M_X(t) = e^{\frac{t^2}{2}}$ and $M_{Y_n}(t) \to e^{\frac{t^2}{2}}$, $n \to \infty$

then for any $a \in \mathbb{R}$

$$P(Y_n \le a) \to \Phi(a), \quad n \to \infty$$

Proof of the CLT

$$\left(1+\frac{\alpha}{n}\right)^n \to e^{\alpha}$$

Recall that X_i are i.i.d., $E(X_i)=\mu$, $Var(X_i)=\sigma^2$, $S_n = X_1 + \cdots + X_n$. We want to apply the continuity theorem for MGF to $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Compute the MGF of Y_n

$$M_{Y_n}(t) = E\left(e^{tY_n}\right) = E\left(e^{t\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sqrt{n}\sigma}}\right) E\left(\prod_{i=1}^{n} e^{\frac{t(X_i-\mu)}{\sqrt{n}\sigma}}\right)$$

$$= \prod_{i=1}^{n} E\left(e^{\frac{t(X_i-\mu)}{\sigma\sqrt{n}}}\right) = \left[E\left(e^{\frac{t(X_i-\mu)}{\sqrt{n}\sigma}}\right)\right]^n$$

For n large enough, $\frac{t(X_i-\mu)}{\sqrt{n}\sigma}$ is small, use $e^x \approx 1 + x + \frac{x^2}{2}$

$$e^{\frac{t(X_1-\mu)}{\sigma\sqrt{n}}} \approx 1 + \frac{t(X_1-\mu)}{\sigma\sqrt{n}} + \frac{t^2(X_1-\mu)^2}{\sigma^2 n}, \text{ so } E\left(e^{\frac{t(X_1-\mu)}{\sigma\sqrt{n}}}\right) \approx 1 + 0 + \frac{t^2}{\cancel{\sigma^2} n}\cdot\frac{\cancel{\sigma^2}}{2}$$

$$M_{Y_n}(t) \approx \left[1 + \frac{t^2}{2}\cdot\frac{1}{n}\right]^n \to e^{\frac{t^2}{2}}, \quad n \to \infty$$