Topics for this week

- Indicator function
- Continuous vs. discrete distribution
- Another example of ‘function of r.v.’
- Revisit Normal distribution (Ch 5.6)
- Bivariate normal distribution (Ch 5.10)

Indicator function

- Indicator function $I(\cdot)$ takes 1 if True; 0 if False
- $I(2+3<6) =?$ 1
- $I(1.2 \in \mathbb{Z}) =?$ 0
- Sometimes written as $I_A(x)$, meaning $I_A(x) = 1$ if $x \in A$, 0 otherwise.
- $I_{\text{nice persons}}(\text{Prof. Shen}) =?$

Indicator function: Example I

- Prove $I_A = 1 - I_{A^c}$
- For any $x \in A$, by definition $I_A(x) = 1$
- and $I_{A^c}(x) = 0$. the result holds.
- Now if $x \in A^c$, $I_A(x) = 0$, and $I_{A^c}(x) = 1$.
- The result holds, too.. Done.

Indicator function: Example II

- Say $f(y) = 3$ if $y < -1$, $f(y) = 6$ if $y \in [-1, 3]$ and $f(y) = 0$ otherwise. How to write f using the indicator function?

$$
\begin{aligned}
f(y) &= 3 \times I(y < -1) + 6 \times I(-1 \leq y \leq 3) + 0 \times I(y > 3) \\
f &= 3 \times I_{(-\infty,-1)} + 6 \times I_{[-1,3]}
\end{aligned}
$$

- Multiply each value by its corresponding indicator function, then sum them up...

Discrete vs. continuous distribution

- X is discrete.. what does that mean?
- That means you can write down X's values and their probabilities...
- Now... what if X is continuous?
- You can't do it.. X's values are within a range...
- Any value has prob 0...

- X follows normal distribution, $P(X=3)=?$
- $P(X=999)=?$
- $P(X=\text{any value})=0...$
- Wait.. but still I feel the chance of $X=3$ is bigger than $X=999$? What's going on?
- That's why we introduce the definition of probability density function (pdf)....

$$f(X=3)=\frac{P(X\in[3,3+\epsilon))}{\epsilon}..\epsilon\to 0.$$
$$f(X=999)=\frac{P(X\in[999,999+\epsilon))}{\epsilon}..$$

- Say both X_1 and X_2 are continuous... how do I understand $X_1 + X_2$?
- Well... the sum is random.. (why?)
- The sum is still continuous.. (why?)
- How about $X_1 \times X_2$?
- What if X_1 is discrete?

Another example of 'function of r.v.'

Suppose $X_1 \sim$Poisson(λ_1), and $X_2 \sim$Poisson(λ_2), independently.

What's the distribution of $Y = X_1 + X_2$?

- MGF. independence, linear combination...

Method I: mgf approach

The mgf of X ~Poisson(λ) is

$$M_X(t) = \exp(\lambda(e^t - 1)).$$

So the mgfs of X_1 and X_2 are

$$M_{X_1}(t) = \exp(\lambda_1(e^t - 1)).$$
$$M_{X_2}(t) = \exp(\lambda_2(e^t - 1)).$$

The mgf of Y is

$$\begin{aligned} M_Y(t) &= M_{X_1}(t) \times M_{X_2}(t) \\ &= \exp(\lambda_1(e^t - 1)) \times \exp(\lambda_2(e^t - 1)) \\ &= \exp((\lambda_1 + \lambda_2)(e^t - 1)). \end{aligned}$$

Therefore, Y is also Poisson r.v. with parameter $\lambda_1 + \lambda_2$.

Method II: use definition only

- Discrete r.v.
- What's the range of Y?
- Well, X_1 takes values in $\{0, 1, 2, 3, \ldots\}$, so does X_2.
- Their sum Y also takes values in $\{0, 1, 2, 3, \ldots\}$

$$P(Y=k) = P(X_1 + X_2 = k) \qquad\qquad **\forall k \in N$$

$$= \sum_{k_1=0}^{k} P(X_1 + X_2 = k | X_1 = k_1) P(X_1 = k_1) \quad * \text{Law of total prob}$$

$$= \sum_{k_1=0}^{k} P(X_2 = k - k_1 | X_1 = k_1) P(X_1 = k_1)$$

$$= \sum_{k_1=0}^{k} P(X_2 = k - k_1) P(X_1 = k_1) \qquad *** \text{independence}$$

$$= \sum_{k_1=0}^{k} \frac{\lambda_2^{k-k_1} e^{-\lambda_2}}{(k-k_1)!} \times \frac{\lambda_1^{k_1} e^{-\lambda_1}}{k_1!}$$

$$= \sum_{k_1=0}^{k} \frac{1}{k_1!(k-k_1)!} \times \lambda_1^{k_1} \lambda_2^{k-k_1} \times e^{-\lambda_1 - \lambda_2}$$

$$
= \sum_{k_1=0}^{k} \frac{1}{k_1!(k-k_1)!} \times \lambda_1^{k_1} \lambda_2^{k-k_1} \times e^{-\lambda_1-\lambda_2}
$$

$$
= \sum_{k_1=0}^{k} \underline{\frac{k!}{k_1!(k-k_1)!} \frac{\lambda_1^{k_1} \lambda_2^{k-k_1}}{(\lambda_1+\lambda_2)^k}} \times \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}
$$

$$
****\text{pmf of Bin}(k, \frac{\lambda_1}{\lambda_1+\lambda_2})
$$

$$
= \left\{ \sum_{k_1=0}^{k} \underline{\frac{k!}{k_1!(k-k_1)!} \frac{\lambda_1^{k_1} \lambda_2^{k-k_1}}{(\lambda_1+\lambda_2)^k}} \right\} \times \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}
$$

$$
= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k}{k!}
$$

$$
***\ \text{pmf of Poisson r.v. with parameter } \lambda_1+\lambda_2
$$

What did we learn from it?

- Try mgf first
- Two random items together, fix one of them, evaluate the other one, then apply the law of total prob.
- The trick of ‘adding extra terms’.

Review: Normal distribution (Ch 5.6)

- Another name: Gaussian distribution.
- Bell-shape, symmetric, on real line, mean = mode.
- pdf of $X \sim N(\mu, \sigma^2)$ as

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Linear combination of independent normal r.v. is still normal.
- $X_1, \ldots, X_n \overset{ind}{\sim}$ normal, then $\sum_{i=1}^n c_i X_i$ is still normal (proof using mgf).

Why do we use normal distribution so often?

- (i) It's simple, and mathematically beautiful.
- (ii) It approximates many distributions well, when the sample size is large (Powerball, insurance, grocery sales, social network data, president election etc.)
- Roughly speaking, asymptotically, sample mean follows a normal distribution regardless of what the true distribution is.
- More on this, in Ch 6.

(iii) not required

- (iii) Any continuous distributions can be approximated by a mixture of normal (independent normal r.v.)

- $$f(x) \approx \frac{1}{2}N(0,1) + \frac{1}{2}N(3,1)$$

- It means half time $N(0,1)$ and half time $N(3,1)$.

- Or more generally, a weighted average,

$$f \approx \sum_{i=1}^{N} w_i N(\mu_i, \sigma_i^2)$$

- The idea of basis expansion, e.g., Taylor expansion

$$f(x) = f(0) + \frac{f'(0)}{1!}\underline{x} + \frac{f''(0)}{2!}\underline{x^2} + \frac{f'''(0)}{3!}\underline{x^3} + \cdots$$

- Or in linear algebra, the basis vectors for a linear space...

$$\left(\begin{array}{c}1\\0\\0\end{array}\right),\left(\begin{array}{c}0\\1\\0\end{array}\right),\left(\begin{array}{c}0\\0\\1\end{array}\right)$$

- Together they form a linear space on R^3.

- (iv) Univariate normal distribution can be easily extended to bivariate (multi-variate) situations, i.e., joint distribution.

Review: bivariate/multivariate distribution

- Model the distribution of X and Y <u>at the same time</u>
- e.g., What's $\text{P}(X = 1, Y = 2)$?
- Why do we care?
- Associations (pairwise/multivariate)
- Weight and height, risk and profit, parent and child's disease.
- Multivariate assoc: Lung cancer $\leftarrow$ heavy use of cigarettes $\leftarrow$ drinking problem $\leftarrow$ anxiety $\leftarrow$ lung cancer..

Review: joint/marginal distribution

- Joint: look at the dist of X,Y,Z <u>at the same time</u>
- Marginal: look at the dist of X,Y,Z <u>individually/separately</u>
- Joint distribution $\Rightarrow$ Marginal dist. (by integrating out)
- Marginal dist $\Rightarrow$ Joint dist? NO!!!
- So when marginal $=$ joint? Independence.

Bivariate normal distribution (Ch 5.10)

Definition: We say X_1 and X_2 follow a bivariate normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation ρ, if their pdf is given by

$$f(x_1, x_2) = \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_1\sigma_2}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right.$$

We write

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

Mean vector, covariance matrix.

Things you should look at for bivariate (multivariate) distributions

- Mean, covariance structure, mode
- Joint pdf, shape
- Marginal dist. (marginal pdf)
- Association (usually complicated, check the sign).

Nice properties

- Mean and covariance. $N\left(\left(\begin{array}{c}\mu_1\\ \mu_2\end{array}\right), \left(\begin{array}{cc}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{array}\right)\right)$
- Symmetric around the mean, mode $=$ mean
- Marginal $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma^2)$.
- Association: $\rho \in (-1, 1)$. Positive (negative) ρ means positive (negative) assoc.
- Positve assoc: X_1 increases, X_2 increases then.
- $\rho = 0$ implies no association, and moreover, independence (this property <u>only</u> holds for joint normal condition).
- Zero correlation doesn't imply independence in general? it DOES for bivariate normal.

Figure: Bivariate normal

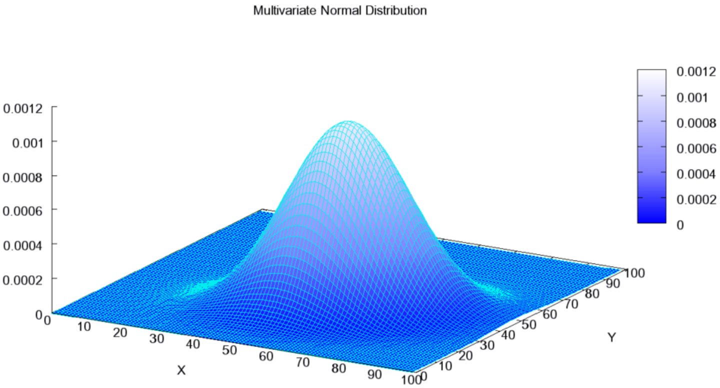
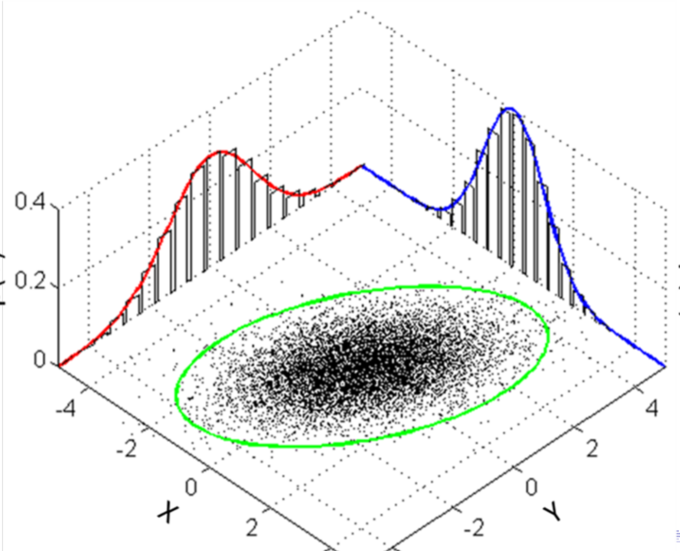


Figure: marginal/joint pdf



- What if $\rho = \pm 1$?

- Recall the definition of correlation

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Recall Cauchy-Schwartz inequality,

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y),$$

where the equality holds if and only if $X = cY$ with prob 1.

- Now corr$= \pm 1$, meaning $X = cY$ for some constant c.

Everything is normal...

- Used to have: sum of independent normal r.v. is normal.
- Now: any linear combo of joint normal is normal.
- Just need to figure out the mean and var.
- Say $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then

$$
\begin{aligned}
& c_1X_1 + c_2X_2 \sim N(\mu, \sigma^2), \\
& \mu = E(c_1X_1 + c_2X_2) = c_1E(X_1) + c_2E(X_2) = c_1\mu_1 + c_2\mu_2, \\
& \sigma^2 = \mathsf{Var}(c_1X_1 + c_2X_2) \\
& \quad = c_1^2\mathsf{Var}(X_1) + 2c_1c_2\mathsf{Cov}(X_1, X_2) + c_2^2\mathsf{Var}(X_2) \\
& \qquad\qquad **\text{Recall } \mathsf{Var}(cX) = c^2\mathsf{Var}(X). \\
& \quad = c_1^2\sigma_1^2 + 2c_1c_2\rho\sigma_1\sigma_2 + c_2^2\sigma_2^2.
\end{aligned}
$$

- Marginal is normal: $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$
- Conditional dist is also normal:

$$X_1 | X_2 = x_2 \sim N(\mu_1 + \rho\sigma_1 \frac{x_2 - \mu_2}{\sigma_2}, (1 - \rho^2)\sigma_1^2)$$

- Compare with the marginal of X_1, we have different mean/variance, what does that mean?
- The marginal of X_1 is like when you first met someone, the first impression.
- Hang out a few times, know that person better
- This extra information/experience is X_2.

- Initially, you have a guess/impression on X_1, then you're updating it by using the extra information from X_2.
- You're thinking like a Bayesian now!
- Update it sequentially: $X_1 \to X_1|X_2 \to X_1|X_2|X_3 \cdots$
- Statistically, it reflects on the variance, roughly speaking, smaller variance means more accuracy.
- Var: $\sigma_1^2 \to (1-\rho^2)\sigma_1^2$.
- Note $\rho \in (-1,1)$, so it's getting smaller, more accurate..
- $\rho = 0$, independence, nothing useful from X_2.

Multivariate normal distribution

- Bivariate normal vs. multivariate normal, there is no fundamental difference.

- Joint normal distribution for $X_1, \ldots, X_d$:

$$N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \ldots \\ \mu_d \end{pmatrix}, \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \ldots & \text{Cov}(X_1, X_d) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \ldots & \text{Cov}(X_2, X_d) \\ \ldots & & & \\ \text{Cov}(X_1, X_d) & \text{Cov}(X_2, X_d) & \ldots & \text{Var}(X_d) \end{pmatrix}\right)$$

- Mean vector, $d \times d$ covariance matrix.

Bivariate normal distribution (Ch 5.10)

Definition: We say X_1 and X_2 follow a bivariate normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation ρ. We write

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

Mean vector, covariance matrix.

Everything is normal

$X_1, \dots, X_d$ follow a joint normal distribution, then

- Any marginal dist is normal, e.g., X_1, X_2, X_3...
- Any lower-dimensional dist is normal, e.g., (X_1, X_3), (X_3, X_4, X_5)
- Linear combination $\sum_{i=1}^{d} c_i X_i$ is normal, e.g., $2X_3 + 5X_10$
- Conditional distribution e.g., $X_1|X_3$, $(X_1, X_3, X_5)|(X_2, X_6)$ is normal,
- X_j and X_k are uncorrelated, then they're independent.

Not required...

- How do we define joint normal? Pdf looks too complicated...
- $(X_1, \ldots, X_d)$ jointly normal if any linear combination of X is normal.
- This result is fascinating!
- My view: a joint normal is like a person/ 3D object in real-life.
- Linear combination is like taking a pic of that person/object, a pic is 2D, no depth.
- What this result is saying: if we take infinitively many pictures like this, we can actually fully recover that in the 3D world.