Lecture 10, Feb 1, 2023

Variance

Definition

Let X be a random variable with distribution $f(x)$, then the *variance* of X is

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2]$$

which is in the discrete case:

$$\sum_x (x-\mu)^2 f(x)$$

in the continuous case:

$$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

$\sigma = \sqrt{\sigma^2} = \sqrt{\text{var}(X)}$ is known as the *standard deviation* of X

- Variance is a measure of variability, or spread – how wide a range we're going to see values from a distribution
- Example: uniform distribution $f(x) = \begin{cases} \frac{1}{a} & 0 \le x \le a \\ 0 & \text{elsewhere} \end{cases}$
  - $\sigma^2 = \int_0^a \left(x - \frac{a}{2}\right)^2 \frac{1}{a}\,\mathrm{d}x = \frac{1}{a}\left[\frac{x^3}{a} - a\frac{x^2}{2} + a^2\frac{x^2}{4}\right]_0^a = \frac{a^2}{12}$
  - Note $\sigma^2 \to 0$ as $a \to 0$, which would give us a delta function
- $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$
  $= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)\,\mathrm{d}x$
  $= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - 2\mu\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + \mu^2\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$
  $= E[X^2] - 2\mu^2 + \mu^2$
  $= E[X^2] - \mu^2$
  $= E[X^2] - (E[X])^2$
  - This also applies to the discrete case since sums can be split exactly in the same way

Covariance and Correlation

Definition

Let X and Y be random variables with joint distribution $f(x,y)$ and means μ_x, μ_y, then the *covariance* of X and Y is

$$\sigma_{xy} = \text{cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

which is in the discrete case

$$\sum_x \sum_y (x-\mu_x)(y-\mu_y) f(x,y)$$

in the continuous case

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

- Note that $\sigma_{xx} = E[(X - \mu_x)(X - \mu_x)] = E[(X - \mu_x)^2] = \sigma^2$, i.e. the variance is the covariance of a random variable with itself
- The covariance is a measure of correlation
  - If the covariance is positive, then both variables tend to be above their means or below their means at the same time; the two variables would be *positively correlated*
  - If the covariance is negative, then when one is above its mean the other would tend to be below its mean; the two variables would be *negatively correlated*
- $\sigma_{xy} = E[XY] - \mu_x \mu_y$

Definition

Let X and Y be random varaibles with covariance σ_{xy} and standard deviations σ_x and σ_y, then the *correlation coefficient* of X and Y is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The correlation coefficient is a normalized, dimensionless version of the covariance; we always have $\rho_{xy} \in [-1, 1]$
  - Two variables are *uncorrelated* if $\rho_{xy} = 0$
  - Note independence implies correlation 0, but correlation 0 does not imply independence