Part V: Probability and Statistics

- V.1 Mean, Variance, and Probability
- V.2 Probability Distributions
- V.3 Moments, Cumulants, and Inequalities of Statistics
- V.4 Covariance Matrices and Joint Probabilities
- V.5 Multivariate Gaussian and Weighted Least Squares
- V.6 Markov Chains

V.1 Mean, Variance, Probability

- Mean
  - Sample mean: done an experiment, got some output
  - Expected mean: known probabilities, but have not used them yet
  - Flip coins 0 or 1: sample mean → expected mean (approach with probability 1)
  - Law of large numbers: sample mean does approach ½ with probability 1 as the number of samples gets larger
- Variance
  - Sample variance: distance from the sample mean
  - Variance: expectation, not trial runs

- Mean is the average value or expected value
- Variance measures the average squared distance from the mean
- Probabilities of n different outcomes are positive numbers $p_1, \ldots, p_n$ adding to 1
- Law of large numbers
  - with probability 1, the sample mean will converge to its expected value E[x] as the sample size N increases (e.g., flip coins)

$$\begin{cases} \text{sample mean: } m = \mu = \dfrac{1}{N}\left(x_1 + x_2 + \cdots + x_N\right) \text{ where } N: \text{ actual output} \leftarrow \text{what we got} \\ \text{expected value: } m = E\left[x\right] = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n = \mathbf{p} \cdot \mathbf{x} \approx \int x p\left(x\right) dx \leftarrow \text{what to expect} \\ n: \text{ number of possible different outputs with their probabilties} \end{cases}$$

$$\begin{cases} \text{sample values: (five random freshmen ages) } 18, 17, 18, 19, 17 \rightarrow \text{sample mean: } 17.8 \\ \text{probabilities: (ages in a freshmen class) } 17(20\%),\ 18(50\%),\ 19(30\%) \rightarrow \text{expected age: } 18.1 \end{cases}$$

Variance (around the mean)

- Variance σ^2 measures expected distance (squared) from the expected mean E[x]
- Sample variance S^2 measures actual distance (squared) from the actual sample mean
- Square root is the standard deviation σ or S

sample variance: $S^2 = \frac{1}{N-1}\left[\left(x_1 - m\right)^2 + \cdots + \left(x_N - m\right)^2\right]$

$\left(N-1\right)$? one degree of freedom is already accounted for in the sample mean

$$\sum\left(x_i - m\right)^2 = \sum\left(x_i^2 - 2mx_i + m^2\right) = \sum x_i^2 - 2m\sum x_i + Nm^2 = \sum x_i^2 - Nm^2$$

variance: $\sigma = E\left[\left(x-m\right)^2\right] = p_1\left(x_1 - m\right)^2 + \cdots + p_n\left(x_n - m\right)^2$

$$E\left[\left(x-m\right)^2\right] = E\left[x^2 - 2mx + m^2\right] = E\left[x^2\right] - 2mE\left[x\right] + m^2 = E\left[x^2\right] - \left(E\left[x\right]\right)^2 = \sum p_i x_i^2 - \left(\sum p_i x_i\right)^2$$

sample ages: $m = 17.8,\ S^2 = \frac{1}{5-1}\left[\left(0.2\right)^2 + \left(-0.8\right)^2 + \left(0.2\right)^2 + \left(1.2\right)^2 + \left(-0.8\right)^2\right] = 0.7$

probabilities of ages: $m = 18.1,\ \sigma^2 = \left(0.2\right)\left(-1.1\right)^2 + \left(0.5\right)\left(-0.1\right)^2 + \left(0.3\right)\left(0.9\right)^2 = 0.49$

Continuous Probability Distributions

- Age: year(17, 18, 19: n=3) → days(17 ≤ x ≤ 20: continuous range), probability distribution p(x)
- Uniform distribution
  - All ages between those numbers are “equally likely”
  - Chance F(x) that a random freshmen has age less than x

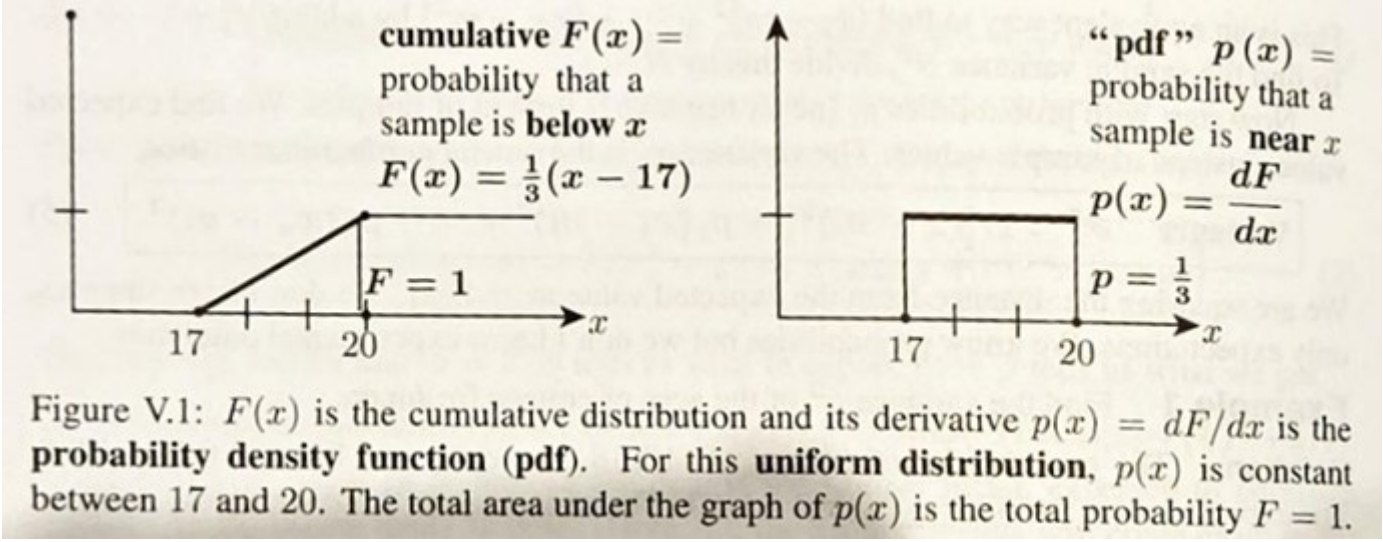

Figure V.1: $F(x)$ is the cumulative distribution and its derivative $p(x) = dF/dx$ is the **probability density function (pdf)**. For this **uniform distribution**, $p(x)$ is constant between 17 and 20. The total area under the graph of $p(x)$ is the total probability $F = 1$.

$p(x)dx$: probability of a sample falling in between x and dx

$$p(x)dx = F(x+dx) - F(x) \rightarrow p(x) = \frac{dF}{dx}, \ (\text{probability of } a \le x \le b) = \int_a^b p(x)dx = F(b) - F(a)$$

$$m = E[x] = \int xp(x)dx \left[= \int_{x=17}^{20} x\left(\frac{1}{3}\right)dx = 18.5 \right]$$

$$\sigma^2 = E\left[(x-m)^2\right] = \int p(x)(x-m)^2 dx \left[= \int_{x=17}^{20} \frac{1}{3}(x-18.5)^2 dx = \frac{3}{4} \right]$$

$$\text{uniform distribution for } 0 \le x \le a \rightarrow \begin{cases} \text{density: } p(x) = \frac{1}{a}, \text{ cumulative: } F(x) = \frac{x}{a} \\ \text{mean: } m = \frac{a}{2} \text{ , variance: } \sigma^2 = \int_0^a \frac{1}{a}\left(x - \frac{a}{2}\right)^2 dx = \frac{a^2}{12} \end{cases}$$

- Normal/Gaussian distribution: Bell-shaped Curve
- Central Limit Theorem
  - The average of N samples of “any” probability distribution approaches a normal distribution as N → ∞

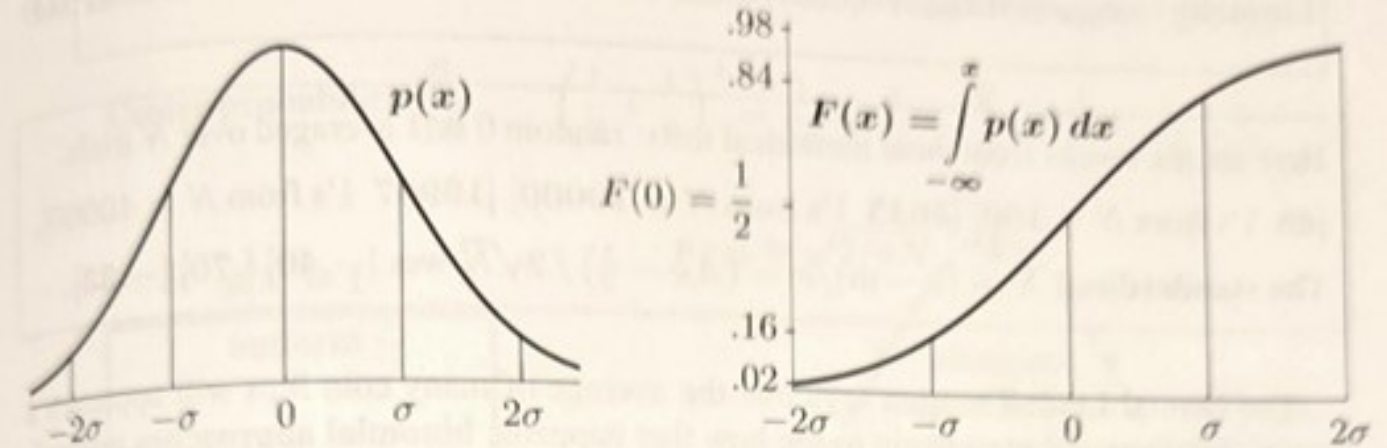

Figure V.2: The standard normal distribution $p(x)$ has mean $m = 0$ and $\sigma = 1$.

standard normal distribution: symmetric $x=0 \rightarrow \underbrace{m=0,\ \sigma^2=1}_{N(0,1)} \rightarrow\ p(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$

$$\begin{cases} \text{total probability: } \int_{-\infty}^{\infty} p(x)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx = 1 \\ \text{mean: } m = E[x] = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}dx = 0 \\ \text{variance: } \sigma^2 = E\left[x^2\right] = \int_{-\infty}^{\infty} p(x)(x-0)^2\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2e^{-x^2/2}dx = 1 \end{cases}$$

$$N(0,\sigma) \rightarrow \begin{cases} F(\sigma)-F(-\sigma) \approx 2/3 \\ F(2\sigma)-F(-2\sigma) \approx 0.95 \end{cases}$$

$$N(0,1) \rightarrow \left[x \xrightarrow{\text{shift}} x-m \xrightarrow{\text{stretch}} \frac{x-m}{\sigma} \right] \rightarrow N(m,\sigma) = p(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-m)^2/2\sigma^2}$$

N coin flips and N $\longrightarrow \infty$

Suppose x is 1 or -1 with equal probabilities $p_1 = p_{-1} = 1/2 \rightarrow \begin{cases} m = 0 \\ \sigma^2 = 1 \end{cases}$

average $A_N = \dfrac{x_1 + \cdots + x_N}{N} \begin{cases} \xrightarrow{\text{law of large numbers}} m = 0 \\ \xrightarrow{\text{linearity}} \sigma_N^2 = \dfrac{\sigma^2}{N^2} + \cdots + \dfrac{\sigma^2}{N^2} = \dfrac{\sigma^2}{N} = \dfrac{1}{N} \end{cases}$

$\rightarrow \begin{cases} \text{Linearity: } x_{new} = a x_{old} + b \\ \rightarrow \begin{cases} m_{new} = a m_{old} + b \\ \sigma^2{}_{new} = a^2 \sigma^2{}_{old} \end{cases} \\ (-1/1) \xrightarrow[b=1/2]{a=1/2} (0/1) \end{cases}$

Change outputs from 1 or -1 to $x = 1$ or $x = 0$. Keep $p_1 = p_0 = 1/2$

$\rightarrow \begin{cases} m = 1/2 \\ \sigma^2 = 1/4 \end{cases} \rightarrow A_N \rightarrow \begin{cases} m = 1/2 \\ \sigma^2 = 1/4N \end{cases}$

$N = 3$: $\underbrace{\text{HHH}}_{\#\text{ of heads}=3}$, $\underbrace{\text{HHT/HTH/THH}}_{\#\text{ of heads}=2}$, $\underbrace{\text{TTH/THT/HTT}}_{\#\text{ of heads}=1}$, $\underbrace{\text{TTT}}_{\#\text{ of heads}=0}$

$\left(\dfrac{1}{2} + \dfrac{1}{2}\right)^3 = \dfrac{1}{8} + \dfrac{3}{8} + \dfrac{3}{8} + \dfrac{1}{8} = 1$, average number of heads in 3 flips: $m = \dfrac{1}{8}(3) + \dfrac{3}{8}(2) + \dfrac{3}{8}(1) + \dfrac{1}{8}(0) = 1.5$

$\sigma^2 = \dfrac{1}{8}(3-1.5)^2 + \dfrac{3}{8}(2-1.5)^2 + \dfrac{3}{8}(1-1.5)^2 + \dfrac{1}{8}(0-1.5)^2 = \dfrac{9+3+3+9}{32} = \dfrac{3}{4}$

$$N=4:\ \left(\frac{1}{2}+\frac{1}{2}\right)^4=\frac{1}{16}+\frac{4}{16}+\frac{6}{16}+\frac{4}{16}+\frac{1}{16}=1$$

For any N, $m=\frac{N}{2}$ and $\sigma^2{}_N=\frac{N}{4}$

subtracting m is "centering" or "detrending"
dividing by σ is "normalizing" or "standardizing"

$$\Big\}\rightarrow X=\frac{x-m}{\sigma}=\frac{x-\frac{N}{2}}{\frac{\sqrt{N}}{2}}\rightarrow\begin{cases}(\text{mean of } X)=0\\(\text{variance of } X)=1\end{cases}$$

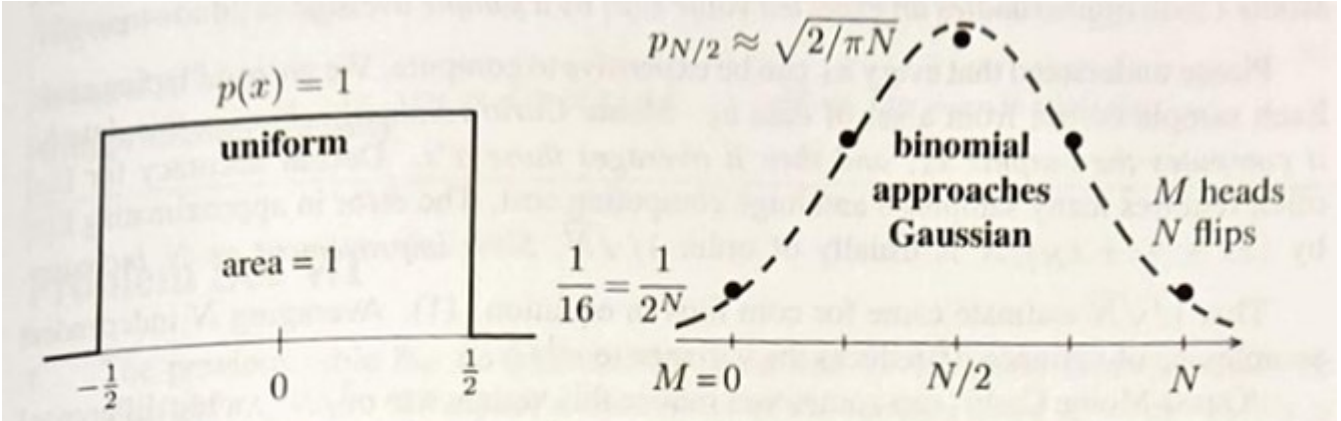

Figure V.3: The probabilities $p=(1,4,6,4,1)/16$ for the number of heads in 4 flips. These p_i approach a Gaussian distribution with variance $\sigma^2=N/4$ centered at $m=N/2$. For X, the Central Limit Theorem gives convergence to the normal distribution $\mathbf{N}(0,1)$.

- Accepting uncertainty in the inputs(b) and estimating the variance in the outputs(x)
  - How to estimate the variance?
  - Often probability distributions p(x) are not known
  - Try different input → compute the outputs → take an average
  - Monte Carlo approximates an expected value E[x] by a sample average: error ~ O(1/√N), slow improvement

samples X_1 to X_N	$m = \frac{X_1 + \cdots X_N}{N}$	$S^2 = \frac{(X_1 - m)^2 + \cdots (X_N - m)^2}{N-1}$
sum of outputs x_i times probabilities p_i	$m = \sum_{i=1}^{n} p_i x_i$	$\sigma^2 = \sum_{i=1}^{n} p_i (x_i - m)^2$
integral of outputs x with probability density	$m = \int x p(x) dx$	$\sigma^2 = \int (x - m)^2 p(x) dx$

V.2 Probability Distributions

- Binomial: tossing a coin n times
- Poisson: rare events
- Exponential: forgetting the past
- Gaussian=Normal: averages of many tries
- Log-normal: logarithm has normal distribution
- Chi-squared: distance squared in n dimensions
- Multivariable Gaussian: probabilities for a vector

V.3 Two Great Inequalities in Statistics

- Markov's inequality (only) applies when all $X_i \geq 0$

$$\text{Prob}\left[X \geq a\right] \leq \frac{\bar{X}}{a} = \frac{mean}{a} = \frac{E\left[X\right]}{a} \leftarrow \bar{X} = E\left[X\right] \begin{cases} = \sum_{\text{all } s} X(s) \text{ times } \left(\text{Probability of } X(s)\right) \\ \geq \sum_{X(s) \geq a} X(s) \text{ times } \left(\text{Probability of } X(s)\right) \\ \geq \sum_{X(s) \geq a} a \text{ times } \left(\text{Probability of } X(s)\right) \end{cases}$$

$$a = 3,\ E\left[X\right] = 1,\ X_i = i,\ \text{show Prob}\left[X \geq 3\right] \leq \frac{1}{3}$$

$$0p_0 + 1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + \cdots = 1 \rightarrow 0p_0 + 1p_1 + 2p_2 + \underbrace{3\left(p_3 + p_4 + p_5 + \cdots\right)}_{\leq 1} + p_4 + 2p_5 + \cdots = 1$$

$$\text{when } 3\left(p_3 + p_4 + p_5 + \cdots\right) = 1?\ p_1 = p_2 = p_4 = p_5 = \cdots = 0 \rightarrow p_3 = 1/3,\ p_0 = 2/3$$

$$\text{if there is equality and Prob}\left[X \geq a\right] = \frac{E\left[X\right]}{a}, \text{ then all probabilities are actually zero except}$$

$$\text{Prob}\left[X = a\right] = \frac{E\left[X\right]}{a} \text{ and Prob}\left[X = 0\right] = 1 - \frac{E\left[X\right]}{a}$$

- Chebyshev inequality (no assumption $X_i \geq 0$)

apply Markov inequality to $Y_i = \left(X_i - m\right)^2$ with same probability p_i

$$\bar{Y} = \sum p_i Y_i = \sum p_i \left(X_i - m\right)^2 = \sigma^2 \rightarrow \left|X_i - m\right| \geq a \rightarrow \left(X_i - m\right)^2 \geq a^2$$

$$\Rightarrow \text{Prob}\left[Y \geq a^2\right] \leq \frac{\bar{Y}}{a^2} = \frac{mean}{a^2} = \frac{\sigma^2}{a^2}$$

V.4 Covariance Matrix

- Linear algebra: M different experiments at once
  - Measure age, height, weight (a, h, w: M=3) of N people
  - M mean values and (separate) variances (→matrix)
  - Connection between the M parallel experiments?
  - p_{ij}: probability that experiment 1 produces x_i and experiment 2 produces y_j

$$\text{covariance: } \sigma_{12} = \sum_{\text{all } i} \sum_{j} p_{ij} \left(x_i - m_1\right)\left(y_i - m_2\right)$$

- Example
  - flip two coins separately vs. glue the coins together

$$\text{for independent experiments, } p_{ij} = \text{probability of } \left(i,j\right) = \left(p_i\right)\left(p_j\right),\ V:\ \text{covariance matrix}$$

$$\begin{cases} \text{coin 1: } \begin{pmatrix} x = 0 \text{ or } 1 \\ \text{with } p = \dfrac{1}{2} \end{pmatrix} \\ \text{coin 2: } \begin{pmatrix} y = 0 \text{ or } 1 \\ \text{with } p = \dfrac{1}{2} \end{pmatrix} \end{cases} \rightarrow \begin{pmatrix} & H & T \\ H & \dfrac{1}{4} & \dfrac{1}{4} \\ T & \dfrac{1}{4} & \dfrac{1}{4} \end{pmatrix} \text{ vs. } \begin{pmatrix} & H & T \\ H & \dfrac{1}{2} & 0 \\ T & 0 & \dfrac{1}{2} \end{pmatrix} \rightarrow \begin{pmatrix} \sigma_{12} = 0 \\ V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \end{pmatrix} \text{ vs. } \begin{pmatrix} \sigma_{12} = 1/4 \\ V = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ \sigma_1^2\sigma_2^2 \geq \left(\sigma_{12}\right)^2 \rightarrow \left|\sigma_{12}\right| \leq \sigma_1\sigma_2 \end{pmatrix}$$

$$\sigma_{12} = \sum_{\text{all } i} \sum_{j} \left(p_i\right)\left(p_j\right)\left(x_i - m_1\right)\left(y_j - m_2\right) = \left[\sum_{i} \left(p_i\right)\left(x_i - m_1\right)\right]\left[\sum_{j} \left(p_j\right)\left(y_i - m_2\right)\right]$$

sample mean: $\bar{\mathbf{X}} = \dfrac{X_1 + \cdots + X_N}{N} \rightarrow$ sample covariance matrix: $\mathbf{S} = \dfrac{\left(X_1 - \bar{X}\right)\left(X_1 - \bar{X}\right)^T + \cdots + \left(X_N - \bar{X}\right)\left(X_N - \bar{X}\right)^T}{N-1}$

probability matrix : $\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$, total probability (all pairs) is 1 : $\sum_{\text{all } i} \sum_{j} p_{ij} = 1$

row sum p_i of $\mathbf{P}$: $\sum_{j=1}^{n} p_{ij} =$ probability p_i of x_i in experiment 1

covariance matrix: $\mathbf{V} = \sum_{\text{all } i} \sum_{j} p_{ij} \underbrace{\begin{bmatrix} x_i - m_1 \\ y_i - m_2 \end{bmatrix} \begin{bmatrix} x_i - m_1 & y_i - m_2 \end{bmatrix}}_{\mathbf{U}\mathbf{U}^T \text{: sum of rank 1 matrices}} = \sum_{\text{all } i} \sum_{j} p_{ij} \begin{bmatrix} \left(x_i - m_1\right)^2 & \left(x_i - m_1\right)\left(y_i - m_2\right) \\ \left(x_i - m_1\right)\left(y_i - m_2\right) & \left(y_i - m_2\right)^2 \end{bmatrix}$

$= \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{bmatrix} = \mathbf{V}^T \rightarrow$ at least positive semidefinite $\begin{cases} p_{ij}\left(x_i - m_1\right)^2 \geq 0 \\ p_{ij}\left(y_i - m_2\right)^2 \geq 0 \rightarrow \begin{pmatrix} \mathbf{V} \text{ is positive definite unless the} \\ \text{experiments are dependent} \end{pmatrix} \\ \det\left(\mathbf{V}_{ij}\right) = 0 \end{cases}$

covariance matrix: $\underset{M \times M}{\mathbf{V}} = E\left[\left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T\right] = \sum p_{ij}\left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T$

$$\begin{cases} \text{mean of } \mathbf{c}^T\mathbf{X}: \; E\left[\mathbf{c}^T\mathbf{X}\right] = \mathbf{c}^T E\left[\mathbf{X}\right] = \mathbf{c}^T\bar{\mathbf{X}} \\ \text{variance of } \mathbf{c}^T\mathbf{X}: \; E\left[\left(\mathbf{c}^T\mathbf{X} - \mathbf{c}^T\bar{\mathbf{X}}\right)\left(\mathbf{c}^T\mathbf{X} - \mathbf{c}^T\bar{\mathbf{X}}\right)^T\right] = \mathbf{c}^T E\left[\left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T\right]\mathbf{c} = \mathbf{c}^T\mathbf{V}\mathbf{c} \geq 0 \end{cases}$$

$\rightarrow$ link between probability and linear algebra : $\mathbf{V} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$

diagonalizing the covariance matrix $\mathbf{V}$ means finding M independent experiments

as combimations of the original M experiments

$\left[\text{mean and variance of } z = x + y\right]$

mean of sum = sum of mean : $\frac{1}{N}\sum_{i=1}^{N}\left(x_i + y_i\right) = \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} y_i$

$$E\left[x+y\right] = \sum_i\sum_j p_{ij}\left(x_i + y_j\right) = \sum_i\sum_j p_{ij}x_i + \sum_i\sum_j p_{ij}y_j = E\left[x\right] + E\left[y\right]$$

$$\sum_i\sum_j p_{ij}x_i = \sum_i\left(p_{i1} + \cdots + p_{iN}\right)x_i = \sum_i p_i x_i = E\left[x\right]$$

$$\sigma_z{}^2 = \sum\sum p_{ij}\left(x_i + y_j - m_x - m_y\right)^2 = \sum\sum p_{ij}\left(x_i - m_x\right)^2 + \sum\sum p_{ij}\left(y_j - m_y\right)^2 + 2\sum\sum p_{ij}\left(x_i - m_x\right)\left(y_j - m_y\right)$$

$$= \sigma_x{}^2 + \sigma_y{}^2 + 2\sigma_{xy}$$

$$\mathbf{X} = \underset{M\times 1}{\begin{bmatrix} x \\ y \end{bmatrix}},\ \mathbf{A} = \underset{K\times M}{\begin{bmatrix} 1 & 1 \end{bmatrix}} \rightarrow \mathbf{Z} = \mathbf{AX}$$

$$\sigma_z{}^2 = \sigma_x{}^2 + \sigma_y{}^2 + 2\sigma_{xy} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_x{}^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y{}^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbf{AVA}^T$$

$$\underset{M\times M}{\mathbf{V_Z}} = \mathbf{A}\ \mathbf{V_X}\ \mathbf{A}^T$$

$\left[\text{correlation } \rho\right]$

rescaling or standardizing the random variables x and y $\left(\text{like } \dfrac{v}{\|v\|}\right)$

$\rightarrow X = \dfrac{x}{\sigma_x}$ and $Y = \dfrac{y}{\sigma_y}$ with $\sigma_X^2 = \sigma_Y^2 = 1$

correlation of x and y $\left(\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y},\ -1 \le \rho_{xy} \le 1\right)$ is the covariance of $X\left(= \dfrac{x}{\sigma_x}\right)$ and $Y\left(= \dfrac{y}{\sigma_y}\right)$