Def (Correlation).

$$\mathrm{Corr}(X, Y) = \rho_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}.$$

$$= \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Cov}(X, X)}\sqrt{\mathrm{Cov}(Y, Y)}}.$$

$\cos\theta$.

- $\mathrm{Corr}(aX + b, cY + d) = \mathrm{Corr}(X, Y)$.
- $-1 \le \rho_{X,Y} \le 1$
- $\rho_{X,Y} = \pm 1$ if and only if $Y = aX + b$ (a.e.)

  if $a > 0$. $\rho_{X,Y} = 1$

  if $a < 0$. $\rho_{X,Y} = -1$.
- $X \perp\!\!\!\perp Y \Rightarrow \rho_{X,Y} = 0$.

  $\not\Leftarrow$

- $Y = 100X + 2$, $\rho_{X,Y} = 1$.

  $Y = -0.0000\cdots01 \cdot X + 2$. $\rho_{X,Y} = -1$.

$-1 < \rho_{X,Y} < 1$

$Y = 5X + 0.1\sqrt{X} + 1$

$\rho_{X,Y} = 0 \Rightarrow$ X and Y doesn't have any linear dependence.

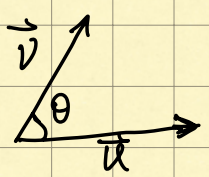
$\{X_i, Y_i\}_{i=1}^{n}$ find a, and b such that

$Y = aX + b + \varepsilon$ → noise term.

fit

$a = \rho \cdot \frac{\sigma_Y}{\sigma_X}$





$$\cos\theta = \frac{\vec{v} \cdot \vec{u}}{|\vec{v}| \cdot |\vec{u}|}.$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Cov}(X,X)} \cdot \sqrt{\text{Cov}(Y,Y)}}.$$

X, Y as vectors

$\text{Cov}(X, Y)$ as inner product.

$\sqrt{\text{Cov}(X, X)}$ as the length of X.

$\text{Corr}(X, Y)$ as the $\cos\theta$, θ is the angle formed by X and Y.

Probability

Statistics.

? . ? ? ?

✓ data ~ / unknown dist (a, B, r ...)

$\mathbb{P}(\text{Event}) = ?$

random variable.

distributions

Geo(p)

Gamma(α, β).

Beta(α, β).

- distributions unknown.
- Geo(p?) Gamma(α?, β?)

use data to know about the distribution, to estimate $\alpha, \beta, \gamma, \ldots$.

Def (Sample. Sample data).

A sample is a collection of random variables $X_1, X_2, \ldots, X_n$. The sample size $= n$.

A sample data. is a collection of observed values of that sample. $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Def (Statistic).

A statistic is any random variable as a function of sample, i.e. $X_1, \ldots, X_n$.

$X_1, \ldots, X_n$.

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n}$$ Sample mean.

$$S^2 := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$$ Sample variance.

$\mathbb{E}(S^2) = \sigma_X^2$ $\quad \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2$

unknown unknown unknown.

X :	x_1	x_2	...	x_n
$P\{X=x_i\}$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

pmf of X.

$$E(X) = \frac{x_1 + \dots + x_n}{n} =: \mu$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 =: \sigma_X^2$$

$x_1, x_2, \dots, x_N'$ $\quad N$

$$\mu = E(X) \xleftarrow{\text{estimate}} \frac{x_1' + \dots + x_N'}{N}$$

$$\sigma_X^2 = Var(X) \xleftarrow{\text{estimate}} \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \frac{x_1' + \dots + x_N'}{N}\right)^2$$

the larger N, the more accurate the estimation.

Confidence interval.

given a threshold 95%, δ

as long as $\frac{x_1' + \dots + x_N'}{N} \in [a, b]$

then we have 95% of confidence that

$$\left|\mu - \frac{x_1' + \dots + x_N'}{N}\right| < \delta$$

A statistic is a $\underline{\text{fun}(X_1, \dots, X_n)}$ random variable so it has a distribution. What does the distribution depend on?

① The underlying distribution of the sample.

$X_1, \cdots, X_n$. $\quad X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

$X_i \sim$ uniform $[0, 1]$.

$\bar{X} \not\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

② The method of sampling.

$X_1, X_2, \cdots, X_n$. $\quad \mathcal{N}(\mu, \sigma^2)$.

method 1: $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ $\quad X_i$'s are independent.

$\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

method 2: $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. $\quad X_2 = X_3 = \cdots = X_n = X_1$

$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{n X_1}{n} = X_1 \sim \mathcal{N}(\mu, \sigma^2)$.

Def (Random sample).

$X_1, \cdots, X_n$ is called a random sample if they are i.i.d (identically independent distributed).

- $X_1, \cdots, X_n$ independent.
- X_i has the same distribution.