# 1 Probability and random variables

- **Probability**: $S$ sample space (all possible states of the system), $F \subset \mathcal{P}(S)$ a $\sigma$-algebra, $P : F \to \mathbb{R}$ a measure, such that $P(S) = 1$.
- **Random variable**: $X : S \to \mathbb{R}$, such that preimages of open sets are in $F$ (i.e. has a well defined probability).
- **Cumulative distribution function** of random variable: $F_X(t) = P(X \leq t)$.
- **Probability distribution** of random variable: $g$ such that $F_X(t) = \sum_{x \leq t, x \in C} g(x)$.
- **Probability density function**: $f$ such that $F_X(t) = \int_{-\infty}^{t} f(s)ds$.
- Two random variables have the **same distribution** if they have the same cdf.

Example: **uniform distribution**:

- $S$ a finite interval $[a, b]$
- $F$: Set of Borel sets on $S$ (sets with a well defined "length")
- $P$: Borel measure ("length") divided by $b - a$
- $X = id$.

## 1.1 Expectation of random variables and their functions

- $X$ is a random variable, the **expectation** of $X$ is $E[X] = \int_S X dP$.
- The **variance** of $X$ is $E[(X - E[X])^2]$.
- The $k$-th **moment** of $X$ is $E[X^k]$.
- The **moment generating function** of $X$ is $E[e^{Xt}]$ (two sided Laplace transform)
- The **characteristic function** of $X$ is $E[e^{itX}]$ (Fourier transform)

Since expectation is defined via integration, one can use the properties of integration to prove statements regarding expectation.

Example: **Chebyshev's theorem**: $E[X] = 0$, $E[X^2] = 1$, then $P(|X| < k) \geq 1 - \frac{1}{k^2}$.
Proof:

$$1 = E[X^2] = \int_S X^2 dP \geq k^2 \int_{|X| \geq k} 1 dP = k^2(1 - P(|X| < k))$$

Example: If $X$ has p.d.f. $f_X$, then $E[g(X)] = \int_{-\infty}^{\infty} g f_x dt$. We prove it when $g(X)$ is bounded via Fubini's theorem:

$$E[g(X)] = \int_S g(X) dP$$

$$= \int_{g(X)\geq 0} \int_0^{g(X)} 1 dy dP - \int_{g(X)<0} \int_{g(X)}^{0} 1 dy dP$$

$$= \int_0^{\infty} \int_{g^{-1}([y,\infty])} f_X(t) dt dy - \int_{-\infty}^{0} \int_{g^{-1}([-\infty,y])} f_X(t) dt dy$$

$$= \int_{-\infty}^{\infty} g f_x dt$$

There is a multivariate version of this formula, and one can also write down $E[g(X)]$ when only the c.d.f. of $X$ is known (via Fubini's theorem or integration by parts).

Can you write down a random variable with neither probability distribution nor p.d.f.?

Can you write down a random variable with no expectation?

## 1.2 Independence and conditional probability for random events

- $A, B \in F$ are **independent** iff $P(A \cap B) = P(A)P(B)$.
- If $P(B) \neq 0$, $P(A \cap B) = P(B)P(A|B)$. Here $P(A|B)$ is the **conditional probability** of $A$ when $B$ is known to happen.

## 1.3 Joint distribution, marginal distribution, conditional distribution

### 1.3.1 Joint distribution

- $X$ and $Y$ are two random variables. The **joint cumulative distribution function** is $F(s,t) = P(X \leq s, Y \leq t)$.
- If $F(s,t) = \sum_{(x,y)\in C, x\leq s, y\leq t} g(s,t)$, we call $g$ the **joint probability distribution**.
- If $F(s,t) = \int_{(-\infty,s]\times(-\infty,t]} f(x,y) dx dy$ we call $f$ the **joint probability density function**.
- $X$ and $Y$ are called independent iff the joint c.d.f. is $F(x,y) = F_X(x)F_Y(y)$.
- The **covariance** between $X$ and $Y$ is $E[(X - E[X])(Y - E[Y])]$

Example: $X$ and $Y$ are two independent random variable with uniform distribution on $[0, 1]$. What is the joint distribution function of $X$ and $Y$? How about $max(X, Y)$ and $min(X, Y)$? What are their covariances?

### 1.3.2 Marginal distribution

Knowing the joint c.d.f. of $X$ and $Y$, the c.d.f. of $X$ or $Y$ are called the **marginal cumulative distribution function**, their p.d. or p.d.f. the **marginal p.d.** or **marginal p.d.f**.

### 1.3.3 Conditional distribution

- If $A$ is a set such that $P(Y \in A) > 0$, then the **conditional cumulative distribution function** of $X$ is $F_{X|Y\in A}(t) = P(X \leq t|Y \in A) = P(X \leq t \cap Y \in A)/P(Y \in A)$. The **conditional p.d.f.**, **conditional p.d.** and **conditional expectation** are defined similarly.

- If $P(Y \in A) = 0$ there isn't a definition of conditional distribution that works in all cases. For example, if $X, Y$ has joint p.d.f. $f_{X,Y}$, and the marginal p.d.f. of $Y$, denoted as $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)dx$, exists and is non zero at $y_0$, then the conditional p.d.f. at $Y = y_0$ is defined as $f_{X|Y=y_0} = f_{X,Y}(x, y_0)/f_Y(y_0)$. The conditional c.d.f. is its integral.

Remark: The definition of conditional distribution for the case $P(Y \in A) = 0$ depends on $Y$ and not just $Y^{-1}(A)$. For example, if $Z = Ye^X$, $f_{X|Y=0} \neq f_{X|Z=0}$.

Example: $X$ is a random variable with uniform distribution on $[0, 1]$, $P(Y = 1|X = p) = p$ (i.e. $P(Y = 1|X \in A) = \int_A p dF_x(p)$), $P(Y = 0|X = p) = 1 - p$. Find the conditional distribution of $X$ when $Y = 1$.

When there are $N$ random variables, $N \geq 3$, the joint/marginal/conditional distributions can be defined analogously.

# 2 Special probability distributions, central limit theorem

## 2.1 Special discrete distributions

- **Bernoulli distribution**: $f(1) = \theta$, $f(0) = 1 - \theta$.

- **Binomial distribution** (sum of iid Bernoulli): $f(x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1 - \theta)^{n-x}$, $x = 0, 1, \ldots, n$.

- **Negative Binomial distribution** (waiting time for the $k$-th success of iid trials): $f(x) = \left(\begin{array}{c} x-1 \\ k-1 \end{array}\right) \theta^k (1-\theta)^{x-k}$, $x = k, k+1, \ldots$. When $k = 1$ it is the **geometric distribution**.

- **Hypergeometric distribution** (randomly pick $n$ elements at random from $N$ elements, the number of elements picked from a fixed subset of $M$ elements) $f(x) = \left(\begin{array}{c} M \\ x \end{array}\right) \left(\begin{array}{c} N-M \\ n-x \end{array}\right) \left(\begin{array}{c} N \\ n \end{array}\right)^{-1}$.

- **Poisson distribution** (limit of binomial as $n \to \infty$, $n\theta \to \lambda$) $f(x) = \lambda^x e^{-\lambda}/x!$.

- **Multinomial distribution** $f(x_1, \ldots x_k) = \left(\begin{array}{c} n \\ x_1, \ldots, x_k \end{array}\right) \theta_1^{x_1} \ldots \theta_k^{x_k}$, $\sum_i x_i = n$, $\theta_i \theta_i = 1$.

- **Multivariate Hypergeometric distribution** $f(x_1, \ldots, x_k) = \prod_i \left(\begin{array}{c} M_i \\ x_i \end{array}\right) \cdot \left(\begin{array}{c} N \\ n \end{array}\right)^{-1}$. $\sum_i x_i = n$, $\sum_i M_i = N$.

## 2.2 Special continuous distributions

- **Uniform distribution**: $f(x) = \begin{cases} 1/(b-a) & x \in (a,b) \\ 0 & x \notin (a,b) \end{cases}$.

- **Normal distribution**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

- **Multivariate Normal distribution**: $x \in \mathbb{R}^d$, $\Sigma$ positive definite $d \times d$ symmetric matrix, $f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$.

- $\chi^2$ **distribution** $d$: degrees of freedom. Squared sum of $d$ normal distributions: $f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \le 0 \end{cases}$.

- **Exponential distribution** $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \le 0 \end{cases}$.

- **Gamma-distribution**: $f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \le 0 \end{cases}$

- **Beta distribution**: (conjugate prior of Bernoulli distribution) $f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in (0,1) \\ 0 & x \notin (0,1) \end{cases}$.

## 2.3 Law of Large Numbers and Central Limit Theorem

### 2.3.1 Convergence

- **Convergence in distribution**: cdf pointwise convergence.
- **Convergence almost surely**: $P(\lim_i X_i \neq X) = 0$.

Example: $X$ uniform on $[0,1]$, $Y_i = \begin{cases} 1 & \exists n \in \mathbb{Z}(X + n \in [\sum_{j=1}^{i} \frac{1}{j}, \sum_{j=1}^{i+1} \frac{1}{j}]) \\ 0 & \text{otherwise} \end{cases}$.
Then $Y_i$ converges to 0 in distribution but not almost surely.

### 2.3.2 CLT and weak LLN

**Levy's continuity theorem**: If $\phi_{X_j} \to \phi_X$ pointwise, then $X_j$ converges to $X$ in distribution.

**Weak Law of Large Numbers** $X_i$ i.i.d. with expectation $\mu$. $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then $S_n$ converges to $\mu$ in distribution.

**(Levy's) Central Limit Theorem** $X_i$ i.i.d. with expectation $\mu$ and variance $\sigma^2 > 0$. $Y_n = \sqrt{\frac{1}{n\sigma^2}} \sum_i (X_i - \mu)$, then $Y_n$ converges in distribution to standard normal distribution (normal distribution with $\mu = 0$ and $\sigma^2 = 1$).

Proof of both theorems (assume $X_i$ bounded): Taylor expansion of the characteristic function.

One can also use the continuity of moment generating function, which is the argument in the textbook.

### 2.3.3 Strong Law of Large Numbers

**Borel-Cantelli Lemma** $A_i$ events, $i = 1, 2, \ldots$, $\sum_i (A_i) < \infty$, then $P(\cap_i (\cup_{j>i} A_j)) = 0$. (the probability of infinitely many $A_i$ happening is 0)

Proof: $P(\cap_i (\cup_{j>i} A_j)) \leq P(\cup_{j>i} A_j) \leq \sum_{j>i} P(A_j)$ which converges to 0 as $i \to \infty$.

**Strong Law of Large Numbers** $X_i$, $i = 1, 2, \ldots$ i.i.d. (independent with identical distribution) and $E(X_i) = \mu$, then $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ converges a.s. to constant $\mu$.

Proof (assume $X_i$ bounded by $M$): Suppose $Var(X_i) = m$. $\sqrt{\frac{n}{m}}(Y_n - \mu)$ has expectation 0 and variance 1, so $P(|Y_n - \mu| > C\sqrt{\frac{m}{n}}) < 1/C^2$ by Chebyshev's theorem. Now let $n_k = k^4$, $C_k = k$, then $Y_{n_k} = Y_{k^4}$ converges a.s. to $\mu$ by Borel-Cantelli.

$Y_n = (\lfloor n^{1/4}\rfloor^4 Y_{\lfloor n^{1/4}\rfloor^4} + X_{\lfloor n^{1/4}\rfloor^4+1} + \cdots + X_n)/n = Y_{\lfloor n^{1/4}\rfloor^4} + (M+|\mu|)\frac{n-\lfloor n^{1/4}\rfloor^4}{n}$.
The first term converges to $\mu$ as $n \to \infty$, and the second converges to 0.

# 3 Sample statistics

## 3.1 Some important distributions

- Standard Normal Distribution: $\mathcal{N}(0,1)$
- $\chi^2(k)$: squared sum of $k$ independent standard normal distribution.
- $t$ distribution: $Z$ standard normal, $Y \sim \chi^2(k)$, $Z$ and $Y$ independent, then $T = \frac{Z}{\sqrt{Y/k}}$ is said to have $t$-distribution with $k$ degrees of freedom.
- $F$ distribution: $U$ and $V$ independent, $U \sim \chi^2(m)$, $V \sim \chi^2(n)$, then $F = \frac{U/m}{V/n}$ is said to have $F$ distribution with degrees of freedom $m$ and $n$,

## 3.2 Sample statistics

$X_1, \dots X_n$ i.i.d. (independent with identical distributions). Sample statistics: a random variable computed from $n$ other random variables.

- **Sample mean**: $\overline{X} = \frac{\sum_i X_i}{n}$
  - $E[\overline{X}] = E[X_1]$, $Var(\overline{X}) = \frac{1}{n} Var(X_1)$.
    Proof:
    $$E[\overline{X}] = E[\frac{1}{n}\sum_i X_i] = \frac{1}{n}\sum_i E[X_i] = E[X_1]$$
    $$Var(\overline{X}) = E[(\overline{X} - E[X_1])^2] = \frac{1}{n^2}E[\sum_i (X_i - E[X_i)^2] = \frac{1}{n}Var(X_1)$$
  - If $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
    Proof: By calculation using MGF.
  - If $n \to \infty$, $\sqrt{\frac{n}{Var(X_1)}}(\overline{X} - E[X_1])$ converges to standard normal by distribution.
    Proof: This is just central limit theorem.
- **Sample variance**: $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_i X_i^2 - n\overline{X}^2)$.
  - $E[S^2] = Var(X_1)$.
    Proof:
    $$E[S^2] = \frac{1}{n-1}\sum_i E[(X_i - \overline{X})^2] = \frac{1}{n-1}\sum_i E[(\frac{n-1}{n}X_i - \sum_{j\neq i}\frac{1}{n}X_j)^2]$$

$$= \frac{1}{n-1} \sum_i \left( \frac{(n-1)^2}{n^2} E[X_i^2] + \sum_{j \neq i} \frac{1}{n^2} E[X_j^2] - \sum_{j \neq i} \frac{2n-2}{n^2} E[X_i] E[X_j] \right.$$

$$\left. + \sum_{j \neq i, k \neq i, j \neq k} \frac{2}{n^2} E[X_j] E[x_k] \right)$$

$$= E[X_1^2] - E[X_1]^2 = Var(X_1)$$

- If $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, then
  - $\overline{X}$ and $S^2$ are independent
    Proof: Calculate joint cdf, do a change of variables.
  - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
    Proof:

    $$\frac{(n-1)S^2}{\sigma^2} + n \frac{(\overline{X} - E[X_1])^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_i (X_i - E[X_1])^2 \sim \chi^2(n)$$

    Now use moment generating function and the independence between $S^2$ and $\overline{X}$.
  - $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.
    Proof: By definition of $t$-distribution.
- If $S_1^2$ is the sample variance of $n_1$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables $Y_i$, $S_2^2$ the sample variance of $n_2$ i.i.d. $\mathcal{N}(\mu', \sigma'^2)$ random variables $Z_j$ independent from $Y_i$, then $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$
  Proof: By definition of $F$-distribution.

- **Order statistics** The $k$-th order statistics is the $k$-th smallest element in $\{X_i\}$, denoted as $Y_k$. Then, if $X_1$ has pdf $f$, then

$$f_{Y_k}(t) = \frac{d}{dt} F_{Y_k}(t) = \lim_{\delta \to 0} \frac{F_{Y_k}(t+\delta) - F_{Y_k}(t)}{\delta}$$

$$= \lim_{\delta \to 0} \frac{1}{\delta} \binom{n}{k-1, 1, n-k} (\int_0^t f ds)^{k-1} \int_t^{t+\delta} f ds (\int_{t+\delta}^{\infty} f ds)^{n-k}$$

$$= \frac{n!}{(k-1)!(n-k)!} (\int_0^t f ds)^{k-1} f(t) (\int_t^{\infty} f ds)^{n-k}$$

## 3.3 PDF of $\chi^2$-, t- and F- distributions

### 3.3.1 $\chi^2$

Let $X_i$ be iid standard normal, their joint distribution is

$$f(x_1, \dots x_n) = (2\pi)^{-n/2} e^{-\sum_i x_i^2/2}$$

Hence the pdf of $\chi^2$ is:

$$f_{\chi^2(n)}(r) = \frac{d}{dr} \int_{\sum_i x_i^2 \le r} (2\pi)^{-n/2} e^{-\sum_i x_i^2/2} dx_1 \dots dx_n$$

which is easy to see must be proportional to $r^{\frac{n-2}{2}} e^{-r/2}$.

## 3.4 $t$

Let $X$ and $Y$ be independent with pdf: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $f_Y(y) = \frac{1}{2^{d/2}\Gamma(d/2)} y^{\frac{d-2}{2}} e^{-y/2}$. Then

$$f_{t(d)}(s) = \frac{d}{ds} P(X \le s\sqrt{Y/d}) = \frac{d}{ds} \int_0^\infty dy \int_{-\infty}^{s\sqrt{y/d}} dx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{2^{d/2}\Gamma(d/2)} y^{\frac{d-2}{2}} e^{-y/2}$$

$$= \int_0^\infty dy \sqrt{y/d} \frac{1}{\sqrt{2\pi}} e^{-s^2 y/2d} \frac{1}{2^{d/2}\Gamma(d/2)} y^{\frac{d-2}{2}} e^{-y/2}$$

Do change of variables $z = (s^2/d + 1)y$ we get that it is proportional to $(s^2/d+1)^{-\frac{d+1}{2}}$.

The calculation for the pdf of $F$ is similar.

# 4 Point estimators and their properties

Basic setting:

- $\mathcal{F}$: a family of possible distributions (represented by a family of cdf, pdf, or pd)
- $\theta : \mathcal{F} \to \mathbb{R}$ population parameter
- $X_1, \dots X_n$ i.i.d. with distribution $F \in \mathcal{F}$
- $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ a function of $X_i$, which is an estimate of $\theta(F)$, is called a point estimate.

Example: $\mathcal{F}$: all distributions with an expectation, then $\overline{X}$ is a point estimate of the expectation.

$\hat{\theta}$ is a point estimate of $\theta$.

- The **bias** is $E[\hat{\theta}] - \theta$. $\hat{\theta}$ is called unbiased if $E[\hat{\theta}] = \theta$.
- The **variance** is $Var(\hat{\theta})$.
- $\hat{\theta}$ is called **minimum variance unbiased estimate** if it has the smallest variance among all unbiased estimates.
- $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimates, the relative efficiency is the ratio of their variance. When they are biased, one can use the mean squared error $E[(\hat{\theta} - \theta)^2]$ instead.
- $\hat{\theta}$ is called **asymptotically unbiased** if bias converges to 0 as $n \to \infty$.
- $\hat{\theta}$ is called **consistent** if $\hat{\theta}$ converges to $\theta$ in distribution.

Example: Estimate of the expectation and variance of binomial distribution

- Expectation can be estimated by sample mean, which is unbiased and consistent.
- Variance can be estimated by sample variance which is unbiased and consistent, or $\overline{X}(1 - \overline{X})$, which is consistent but biased.

Example: Estimate $t$ for uniform distribution on $[0, t]$.
The following estimates are all unbiased and consistent:

- $2\overline{X}$
- $\frac{n+1}{n} Max(X_i)$
- $Max(X_i) + Min(X_i)$

Can you calculate their variance? Which is the best among the three?

Answer:

$$Var(2\overline{X}) = \frac{4}{n} \cdot Var(X_1) = \frac{t^2}{3n}$$

$$Var(\frac{n+1}{n} Max(X_i)) = \frac{(n+1)^2}{n^2} \cdot n! \cdot \int_0^t dx_n \int_0^{x_n} dx_{n-1} \cdots \int_0^{x_2} dx_1 \cdot \frac{(x_n - t)^2}{t^n}$$

$$= \frac{(n+1)^2}{n} \int_0^t \frac{(x_n - \frac{nt}{n+1})^2 x_n^{n-1}}{t^n} dx_n = \frac{t^2}{n(n+2)}$$

$$Var(Max(X_i) + Min(X_i)) = \frac{n!}{t^n} \cdot \int_0^t dx_n \int_0^{x_n} dx_1 \int_{x_1}^{x_n} dx_{n-1} \ldots d_{x_2} \cdot (x_n + x_1 - t)^2$$

$$= \frac{n(n-1)}{t^n} \int_0^t dx_n \int_0^{x_n} dx_1 (x_n + x_1 - t)^2 (x_n - x_1)^{n-2} = \frac{2t^2}{(n+1)(n+2)}$$

If an asymptotically unbiased estimate has variance $\to 0$ when $n \to \infty$, it must be consistent.

Cramer-Rao inequality:

$$Var(\hat{\theta}) \geq \frac{1}{nE[(\frac{d}{d\theta}\log f)^2]}$$

When equality is reached we get minimal variance unbiased estimate.

Example: $X_i$ iid normal, then $\overline{X}$ is MVUE.

$$Var(\overline{X}) = \sigma^2/n$$

$$\frac{1}{nE[(\frac{d}{d\theta}\log f)^2]} = \frac{1}{nE[(X-\mu)^2/\sigma^4]} = \sigma^2/n$$

# 5 Method of moments, Maximum likelihood

## 5.1 MLE

Suppose $X_i \sim F \in \mathcal{F}$, i.i.d., where $\mathcal{F}$ is the family of possible distributions of $X_i$, and $F$ is unknown and belongs to $\mathcal{F}$. We want to find a point estimate for some function $\theta : \mathcal{F} \to \mathbb{R}$. The Method of Maximal Likelihood is:

$$\hat{\theta}(X_1, \dots X_k)_{MLE} = \theta(\arg\max_{F \in \mathcal{F}} L(X_1, \dots X_k, F)$$

- When $F$ is a continuous distribution with p.d.f. $f(x)$, let $L(x_1, \dots, x_k, F) = \prod_i f(x_i)$
- When $F$ is a discrete distribution with p.d. $g(x) = P(X = x)$, let $L(x_1, \dots, x_k, F) = \prod_i g(x_i)$

Example: $X_i$ i.i.d. and has binomial distribution with $n = 5$ and unknown $p$, find MLE for $p$.

Answer: If $X_i$ satisfies the binomial distribution with $n = 5$ and let $p$ be some unknown value, the likelihood function is:

$$L(X_1, \dots X_k) = \prod_i \binom{5}{X_i} p^{X_i} (1-p)^{5-X_i}$$

The $p$ that maximizes it is $p = \frac{\sum_i X_i}{5k}$, hence $\hat{p}_{MLE} = \frac{\sum_i X_i}{5k}$

Example: $X_i$ i.i.d. and has uniform distribution on $[a, a+t]$. Find MLE for $a$ and $t$.

Answer: If $[a, a+t]$ fails to contain any of the $X_i$ the likelihood must be 0, so $a \leq \min\{X_i\}$, $a+t \geq \max\{X_i\}$. To maximize the likelihood in this case, one need to minimize $t$, hence $\hat{a}_{MLE} = \min\{X_i\}$ and $\hat{t}_{MLE} = \max\{X_i\} - \min\{X_i\}$.

Example: $X_i$ i.i.d. and has normal distribution with expectation $\mu$ variance $\sigma^2$. Find MLE for $\sigma^2$.

Answer: Write down the likelihood function, take derivative for both $\mu$ and $\sigma^2$ and set both to be 0, we get that $\hat{\sigma^2}_{MLE} = \frac{1}{n}\sum_i (X_i - \overline{X})$.

### 5.2 MOM

MOM is a less popular approach but does have some advantages in some situations.

Empirical distribution: Given $x_1, \ldots x_k \in \mathbb{R}$, the empirical distribution $X'$ is defined as $P(X' = x_i) = \frac{m_i}{k}$ where $m_i$ is the multiplicity of $x_i$.

Method of moments means estimating the parameters in such a way that the first few moments of $X_i$ under these parameters match the first few moments of empirical distribution obtained from $X_1, \ldots X_k$.

Example: $X_i$ i.i.d. uniform on $[a, a+t]$, find MOM estimate for $a$ and $t$.

Example: $X_i$ i.i.d. normal, find MOM estimate for expectation and variance.

## 6 Maximum a posteriori

## 7 Hypothesis testing

## 8 Examples of hypothesis testing

## 9 Confidence interval

## 10 Linear Regression

## 11 ANOVA

## 12 Example of non parametric methods