1 Probability and random variables

- **Probability**: S sample space (all possible states of the system), $F \subset \mathcal{P}(S)$ a σ-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): $f(x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} \theta^k (1-\theta)^{x-k}$, $x = k, k+1, \ldots$. 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) $f(x) = \begin{pmatrix} M \\ x \end{pmatrix} \begin{pmatrix} N-M \\ n-x \end{pmatrix} \begin{pmatrix} N \\ n \end{pmatrix}^{-1}$.
- **Poisson distribution** (limit of binomial as $n \to \infty$, $n\theta \to \lambda$) $f(x) = \lambda^x e^{-\lambda}/x!$.
- **Multinomial distribution** $f(x_1, \ldots x_k) = \begin{pmatrix} n \\ x_1, \ldots, x_k \end{pmatrix} \theta_1^{x_1} \ldots \theta_k^{x_k}$, $\sum_i x_i = n$, $\theta_i \theta_i = 1$.
- **Multivariate Hypergeometric distribution** $f(x_1, \ldots, x_k) = \prod_i \begin{pmatrix} M_i \\ x_i \end{pmatrix} \cdot \begin{pmatrix} N \\ n \end{pmatrix}^{-1}$. $\sum_i x_i = n$, $\sum_i M_i = N$.

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**: $x \in \mathbb{R}^d$, Σ positive definite $d \times d$ symmetric matrix, $f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$.
- χ^2 **distribution** d: degrees of freedom. Squared sum of d normal distributions: $f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$.
- **Exponential distribution** $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$.
- **Gamma-distribution**: $f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \leq 0 \end{cases}$
- **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 μ. $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then S_n converges to μ in distribution.

(Levy's) Central Limit Theorem X_i i.i.d. with expectation μ 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 μ.

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 μ 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 μ 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

$X_1, \ldots X_n$ 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

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

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

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

3.3 PDF of χ^2-, t- and F- distributions

3.3.1 χ^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 χ^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 θ.

- 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 θ 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 μ variance σ^2. Find MLE for σ^2.

Answer: Write down the likelihood function, take derivative for both μ and σ^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 $X_1, \ldots X_k$, i.e. the sample moments $M'_n = \frac{1}{k}\sum_i X_i^n$.

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

Example: X_i i.i.d. exponential, $f(x) = \frac{1}{c}e^{-x/c}$, find MOM estimate for c.

Example: X_i i.i.d. binomial with $p = \frac{1}{2}$. Find MOM and MLE for n. Are they the same?

5.3 Point estimate for non i.i.d. random variables

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

One can define bias, variance and consistency similar to the i.i.d. case. The MLE (and MAP which will be discussed later) works for non i.i.d. case as well!

Example: $X_1, \ldots X_n$ uniform on $[a, a+t]$, $Y_1, \ldots, Y_n$ uniform on $[b, b+t]$, find MLE of t.

Midterm 1 review

Some key topics:

- pdf from cdf, probability from pdf, expectation from pdf.
- LLN and CLT
- Sample mean and sample variance: general case and normal population
- Bias, variance, mean squared error and consistency for point estimate Ways to check consistency:
  - Definition
  - LLN
  - mean squared error
- MLE

Review problem:

1. X_i i.i.d. $P(X_i = 0) = a$, $P(X_i = 1) = b$, $P(X_i = 2) = c$, $a + b + c = 1$. Find $E[X_i]$, the MLE of $E[X_i]$, and the bias and variance of said MLE. Find $Var(X_i)$ and its MLE. Are these MLEs consistent?

2. X_i i.i.d., with pdf $\frac{1}{\sqrt{2\pi}}(ce^{-x^2/2} + (1-c)e^{-(x-1)^2/2}$. Find MLE of c. Is it consistent?

Digression: The idea of substitution

Examples:

- If X_i i.i.d. with p.d.f. f, then likelihood function $L = \prod_i f(X_i)$. If Y_i i.i.d., tY_i has $\chi^2(2)$ distribution (p.d.f. $f(x) = \frac{1}{2}e^{-x/2}$ when $x \geq 0$), what is the likelihood function?
- Let $f_\theta(\cdot)$ be the p.d.f. of θ, $f_{X|\theta}(\cdot, theta)$ the conditional p.d.f. of X, $f_{\theta|X}(\cdot, X)$ the conditional p.d.f. of θ, then

$$f_{\theta|X}(\theta, x) = \frac{f_\theta(\theta) f_{X|\theta}(x, \theta)}{\int_{\mathbb{R}} f_\theta(\theta) f_{X|\theta}(x, \theta) d\theta}$$

6 Bayesian statistics

6.1 The basic idea of Bayesian statistics

- Input:
  - Some (possibly vector valued) random variable Θ with given distribution (**prior**)
  - Some (possibly vector valued) random variable X with known conditional distribution conditioned at a value of Θ, $X \sim F(X|\Theta)$. (**observable**)
- Output: the conditional distribution of Θ conditioned at a value of X (**posterior**) $\Theta \sim F(\Theta|X)$.

Example 1:

- **Prior** $Y \sim Bernoulli(\frac{1}{100})$
- **Observable** X_1, X_2 conditionally i.i.d. when $Y = y$, and their conditional distribution is Bernoulli with $p = \frac{1+8Y}{10}$.

Calculation of the posterior:

$$P(Y = 1|X_1, X_2) = \frac{P(Y = 1, X_1, X_2)}{P(X_1, X_2)}$$
$$= \frac{P(X_1, X_2|Y = 1)P(Y = 1)}{P(X_1, X_2|Y = 0)|P(Y = 0) + P(X_1, X_2|Y = 1)|P(Y = 1)}$$
$$= \frac{(9/10)^{X_1+X_2}(1/10)^{2-X_1-X_2} \times \frac{1}{100}}{(9/10)^{X_1+X_2}(1/10)^{2-X_1-X_2} \times \frac{1}{100} + (1/10)^{X_1+X_2}(9/10)^{2-X_1-X_2} \times \frac{99}{100}}$$
$$= \frac{9^{X_1+X_2}}{9^{X_1+X_2} + 99 \times 9^{2-X_1-X_2}}$$

So, for example, if we know both X_i takes a value of 1, then the probability of $Y = 1$ is 9/20.

We can answer many questions using posterior, for example:

- What is the probability of Θ taking value in A given X?
- What is the "most likely" value of Θ? $\hat{\Theta}_{MAP} = \arg\max_s f_{\Theta|X}(s)$, where f is p.d.f. when $\Theta|X$ is continuous and p.d. when it is discrete. This is called the **maximum a posteriori (MAP)** estimate.
- What is the average value of Θ? $\hat{\Theta} = E[\Theta|X]$. This is called the **Bayesian point estimate with L^2 lost**.
- In general, let $l(\cdot, \cdot)$ be a lost function (a positive function such that $l(a, a) = 0$), then $\hat{\Theta} = \arg\min_\theta E[l(\Theta, \theta)|X]$ is called the **Bayesian point estimate**.

6.2 Comparison between non-Bayesian and Bayesian

MLE:

- Input: Assumption on the distribution of X: $X \sim F(\alpha)$. A likelihood function $L(X, \alpha)$.
- Output: $\hat{\alpha}_{MLE} = \arg\max_\alpha L(X, \alpha)$.

Bayesian statistics:

- Input: Prior: $\alpha \sim F_0$, Conditional distribution: $X|\alpha \sim F(\alpha)$.
- Calculated output: Posterior: $\alpha|X \sim F'(X)$
- MAP Point estimate: $\hat{\alpha} = \arg\max_\alpha f_{alpha|X}(\alpha)$
- L^2-Bayesian Point estimate: $\hat{\alpha} = E[\alpha|X]$.

6.3 Example of point estimate using Bayesian statictics

Example 2:

Input:

- $\mu \sim \mathcal{N}(0, 1)$
- $X_i|\mu$ cond. i.i.d., $\sim \mathcal{N}(\mu, 1)$

Posterior:

$$f_{\mu|X_i}(s) = \frac{f_{\mu,X_i}(s, X_1, \ldots, X_n)}{f_{X_i}(X_1, \ldots X_n)} = \frac{f_{\mu,X_i}(s, X_1, \ldots, X_n)}{\int_{\mathbb{R}} f_{\mu,X_i}(t, X_1, \ldots X_n)dt}$$

$$= \frac{\prod_i f_{X_i|\mu=s}(X_i) f_\mu(s)}{\int_{\mathbb{R}} \prod_i f_{X_i|\mu=t}(X_i) f_\mu(t)dt} = \frac{(2\pi)^{-\frac{n+1}{2}} e^{-\sum_i (X_i - s)^2/2 - s^2/2}}{\int_{\mathbb{R}} (2\pi)^{-\frac{n+1}{2}} e^{-\sum_i (X_i - t)^2/2 - t^2/2} dt}$$

So

$$\mu|X_i \sim \mathcal{N}(\frac{\sum_i X_i}{n+1}, \frac{1}{n+1})$$

The MAP and L^2 Bayesian estimate of μ are both $\hat{\mu} = \frac{\sum_i X_i}{n+1}$.

From the computation above we get:

$$f_{\mu|X}(s) \propto f_{X|\mu=s}(X) f_\mu(s)$$

This works for discrete μ or X as well!

Example 3: P uniform on $[0, 1]$, $X|P \sim$ Binomial$(5, P)$, then $f_{P|X}(s) \propto s^X (1-s)^{5-X} \cdot 1$, hence $P|X \sim Beta(X + 1, 6 - X)$.

6.4 Hierarchical Models

This section is beyond the scope of our exams.

Often in practice we build "hierarchical models" by stacking multiple layers of Bayesian and non Bayesian models together. For example:

$$\sigma_i^2 \sim \Gamma(\alpha, \beta)$$
$$\sigma^2 \sim \Gamma(\alpha', \beta')$$
$$\mu_i \sim \mathcal{N}(0, \sigma^2)$$
$$X_{ij} \text{ ind. } \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

How would you estimate σ_i and μ_i from the values of X_{ij}?

We will talk about models like this if we have more time at the end of the semester.

6.5 More Examples

Example 4: t has p.d.f. $f_t(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x > 0 \end{cases}$. $P(Y = n|t) = (1 - e^{-t})e^{-nt}$. Knowing Y, find $\hat{t}_{MAP}$ and $E[t|Y]$.

Example 5: a, t indep. $\sim$ Uniform($[0, 1]$). $X_i|a, t$ i.i.d. $\sim$ Uniform($[a, a + t]$), find $\hat{t}_{MAP}$.

Answer: $M = \max(X_i)$, $m = \min(X_i)$, then:

$$f_{a,t|X_i} \propto \begin{cases} t^{-n} & 0 \le a \le m \le M \le a + t \le a + 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$f_{t|X_i} \propto \begin{cases} t^{-n} \cdot (\min(1, m) - (M - t)) & M - \min(1, m) \le t \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{t}_{MAP} = \min(1, \frac{n}{n-1}(M - \min(1, m)))$$

7 Hypothesis testing

7.1 Definitions

- Problem: want to know if the distribution of X satisfy certain propositions (**null hypothesis**), for example:
  - Will the coronavirus kill more than a million people in the end?

  - Will the expectation of our midterm 2 grade be better than midterm 1?
  - Is the performance of a machine learning algorithm better than random chance?

- Solution: Find a random variable Z (**test statistics**) depending on X and a set A (**critical region**), and reject the hypothesis when $Z \in A$.
- (Z, A) is called a **statistical test** to null hypothesis H_0.
- If $Z \in A \iff Z' \in A'$ we consider (Z, A) and (Z', A') to be the same test.
- If H_0 completely determines $P(Z \in A)$ (**simple hypothesis**), $p = P(Z \in A|H_0)$ is called the **significance level**.

The key reasoning behind statistical tests: **Suppose H_0 is true. If (Z, A) is a test with a very small significance level, then $Z \in A$ is highly unlikely. If, however, we do actually observe that $Z \in A$, then this can only tell us that H_0 is unlikely to be true.** It is basically a kind of "statistical" proof by contradiction.

Example 1: Suppose your grade for midterm 1 is X_1, your grade for midterm 2 is X_2, $Y = X_2 - X_1$ satisfies normal distribution with variance 25. How do we test the null hypothesis $E[Y] = 0$?

- Answer 1: $Z = Y$, $A = (-\infty, -M) \cup (M, \infty)$.

  $p = P(Y < -M \cup Y > M|H_0) = P(Y < -M|Y \sim \mathcal{N}(0, 25)) + P(Y > M|Y \sim \mathcal{N}(0, 25))$

  $$= 2\int_M^\infty \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$$

- Answer 2: $Z = Y$, $A = (M, \infty)$, $p = \int_M^\infty \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$
- Answer 3: $Z = Y$, $A = (-M, M)$, $p = \int_{-M}^M \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$

Which of the three is more reasonable?

- **Alternative hypothesis**: an alternative to the null hypothesis H_0, called H_1.
- $P(Z \in A|H_0)$ is called **significance level** or **type I error**.
- If H_1 is a simple hypothesis, $P(Z \notin A|H_1)$ is called **type II error**.
- If H_1 is a simple hypothesis, $1 - P(Z \notin A|H_1) = P(Z \in A|H_1)$ is called **(statistical) power**

- If $X \sim F(\theta)$, $\pi(\theta) = P(Z \in A|\theta)$ is called the **power function**. If $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, then p-value is $\pi(\theta_0)$ and power is $\pi(\theta_1)$.

In Example 1, let $Y = \mathcal{N}(\theta, 25)$, what is the power function of the three tests?

Answer: Let $f(y) = \frac{1}{\sqrt{50\pi}} e^{-y^2/50}$. Then for Test 1,

$$\pi_1(\theta) = \int_{(-\infty,-M)\cup(M,\infty)} f(t-\theta)dt = \int_{(-\infty,-M-\theta)\cup(M-\theta,\infty)} f(s)ds$$

So $\frac{d\pi}{d\theta} = f(M-\theta) - f(-M-\theta)$, which is positive when $\theta > 0$ and negative when $\theta < 0$. So, if the alternative hypothesis is $\theta = \theta_1 >> 0$ or $\theta = \theta_1 << 0$, it is possible to find some M which make significance level small and power large.

For Test 2,

$$\pi_2(\theta) = \int_{(M,\infty)} f(t-\theta)dt = \int_{(M-\theta,\infty)} f(s)ds$$

So $\frac{d\pi}{d\theta} = f(M-\theta) > 0$. So if the alternative hypothesis is $\theta = \theta_1 >> 0$ it is possible to find some M which make significance level small and power large. In other words, this test can only capture the case when $E[Y] > 0$ but not $E[Y] < 0$, which is consistent with our expectation.

For Test 3, the power function is

$$\pi_3(\theta) = \int_{(-M,M)} f(t-\theta)dt = \int_{(-M-\theta,M-\theta)} f(s)ds$$

So $\frac{d\pi}{d\theta} = -f(M-\theta) + f(-M-\theta)$ which is negative when $\theta > 0$ and positive when $\theta < 0$, so as a consequence the type I and type II errors always sum up to something larger than 1, which means that it is a very bad test.

Example 2: Y_i i.i.d. $\sim \mathcal{N}(\theta, 25)$, $H_0 : \theta = 0$. What is the power function for the test $(\overline{Y}, (-\infty, -M) \cup (M, \infty))$?

Example 3: Y_i i.i.d. Bernoulli distribution with parameter θ, $H_0 : \theta = 0.5$. What is the power function for the test $(\overline{Y}, (0, 1/2 - \epsilon) \cup (1/2 + \epsilon, 1))$?

Example 4: X_i $i = 1, \ldots 6$ *i.i.d.*, Bernoulli with $P(X_i = 1) = p$. $H_0 : p = 0.5$, $H_1 : p = 0.9$. Test statistics: $Z = \sum_i X_i$. $A = [M, 6]$, M is an integer.

Then power function is:

$$\pi(p) = P(Z \geq M|p) = \sum_{i=M}^{6} \binom{6}{i} p^i (1-p)^{6-i}$$

p-value is $\pi(0.5) = \frac{1}{64} \sum_{i=M}^{6} \binom{6}{i}$. Power is $\pi(0.9) = \sum_{i=M}^{6} \binom{6}{i} (0.9)^i (0.1)^{6-i}$.

- $M = 6$: significance=0.0156, power=0.531
- $M = 5$: significance=0.109, power=0.886
- $M = 4$: significance=0.344, power=0.984

There is trade-off between significance and power. Which M to choose depends on the purpose of the test, in particular whether false positive or false negative would be more costly.

7.2 Likelihood ratio test

Recall that the likelihood function is $L(x, \theta) = f_{X|\theta}(x)$, which is the p.d.f. when X is continuous and p.d. when X is discrete. The Neyman-Pearson test for $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ is:

$$(X, \{x : L(x, \theta_0)/L(x, \theta_1) \leq k\})$$

Example 4, Neyman-Pearson test: $p_0 = 0.5$, $p_1 = 0.9$

$$L(X_1, \ldots, X_6, p_0) = \prod_i p_0^{X_i}(1 - p_0)^{1-X_i} = \frac{1}{2^6}$$

$$L(X_1, \ldots, X_6, p_1) = \prod_i p_1^{X_i}(1 - p_1)^{1-X_i}$$

$$= 0.9^{\sum_i X_i} \cdot 0.1^{6-\sum_i X_i} = 0.1^6 \cdot 9^{\sum_i X_i}$$

So likelihood ratio decreases with $\sum_i X_i$.

Sometimes we need to consider **composite hypothesis**, i.e. cases when H_0 and H_1 does not completely determine the distribution of X. Suppose $H_0 : \theta \in D_0$, $H_1 : \theta \in D_1$, the likelihood ratio test becomes:

$$(X, \{x : \frac{\sup_{\theta \in D_0} L(x, \theta)}{\sup_{\theta \in D_0 \cup D_1} L(x, \theta)} \leq k\})$$

How would you do likelihood ratio test for the following examples:

- X_i i.i.d. Bernoulli(p). $H_0 : p = 0.5$, $H_1 : p \neq 0.5$.
- X_i i.i.d. $\mathcal{N}(\mu, 1)$. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.

Answer:

- Likelihood under H_0 is

$$L_0 = \prod_i 0.5^{X_i}(1 - 0.5)^{1-X_i} = 0.5^n$$

maximum likelihood under H_0 or H_1 is

$$L_1 = \sup_p \prod_i p_i^X (1-p)^{1-X_i}$$

$$= \sup_p p^{\sum_i X_i} (1-p)^{n-\sum_i X_i}$$

$$= (\sum_i X_i/n)^{\sum_i X_i} (1 - \sum_i X_i/n)^{n-\sum_i X_i}$$

It is easy to see that the likelihood ration L_0/L_1, as a function of $\sum_i X_i$, is symmetric with regards to $n/2$, and takes its maximum at $\sum_i X_i = n/2$. So the likelihood ratio test must be of the form: $|\sum_i X_i - n/2| \geq C$ for some C.

- Likelihood under H_0 is

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi}} e^{-X_i^2/2} = (2\pi)^{-n/2} e^{-\frac{\sum_i X_i^2}{2}}$$

maximum likelihood under H_0 or H_1 is

$$L_1 = \sup_\mu \prod_i \frac{1}{\sqrt{2\pi}} e^{-(X_i-\mu)^2/2}$$

$$= \sup_\mu (2\pi)^{-n/2} e^{-\frac{\sum_i (X_i-\mu)^2}{2}}$$

$$= (2\pi)^{-n/2} e^{-\frac{\sum_i X_i^2 - (\sum_i X_i)^2/n}{2}}$$

So

$$L_0/L_1 = e^{-\frac{(\sum_i X_i)^2}{2n}}$$

So the likelihood ratio test must be of the form $|\sum_i X_i| \geq C$.

7.3 Proof of Neyman-Pearson Lemma

Neyman-Pearson test has the highest power for given significance, and lowest significance level for given power.

Proof in continuous case: Let X taking value in $\mathbb{R}^n$, k be the threshold of the Neyman-Pearson test with significance α. In other words,

$$\int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} f_{X|H_0}(x)dx = \alpha$$

Then its power is $\beta_0 = \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} f_{X|H_1}(x)dx$.

Suppose another test (Z, A) has significance α, then by definition of conditional p.d.f.,

$$\int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_0}(x)dx = \alpha$$

While the power is

$$\begin{aligned}
&\int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_1}(x)dx \\
&= \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \in A|X) f_{X|H_1}(x)dx + \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_1}(x)dx \\
&= \beta_0 - \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \notin A|X) f_{X|H_1}(x)dx + \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_1}(x)dx \\
&\leq \beta_0 - \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \notin A|X) f_{X|H_0}(x)dx + \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_0}(x)dx \\
&= \beta_0 - \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} f_{X|H_0}(x)dx + \frac{1}{k} \int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_0}(x)dx \\
&= \beta_0
\end{aligned}$$

8 Examples of hypothesis testing

8.1 Significance and p-value

$X \sim F(\theta)$, $H_0 : \theta \in D_0$.

Suppose a family of statistical tests with parameter k is $X \in A(k)$.

Then:

- The significance level of the test $X \in A(k)$ is $\alpha = \sup_{\theta \in D_0} P(X \in A(k)|\theta)$
- The p-value for x, which is an observed value of X, is

$$p = \inf_{k \in \{k : x \in A(k)\}} \sup_{\theta \in D_0} P(X \in A(k))$$

- Suppose the test $X \in A(k_0)$ has significance level α_0. Then $x \in A(k_0)$ (i.e. $X = x$ results in rejection of H_0 under this test) implies that x has a p-value no larger than α_0, and x has p-value less than α_0 implies that $x \in A(k_0)$.

  Proof: Let $\alpha(k) = \sup_{\theta \in D_0} P(X \in A(k)|\theta)$, then because $P(X \in A(k)|\theta)$ is non-increasing, $k \mapsto \alpha(k)$ is non increasing. Furthermore, by assumption, $\alpha(k_0) = \alpha_0$, and $\alpha(k) > \alpha_0 \implies k > k_0$, and the p-value for x is

$$p = \inf_{k \in \{k : x \in A(k)\}} \alpha(k)$$

Suppose $x \in A(k_0)$, then the p-value of x is $p = \inf_{k \in \{k: x \in A(k)\}} \alpha(k) \leq \alpha(k_0) = \alpha_0$.

Now suppose the p-value of x is less than α_0, then there is some k' such that $x \in A(k')$ and $\alpha(k') < \alpha_0$. Hence, $k' \leq k_0$, $x \in A(k') \subset A(k_0)$.

- Suppose a statistical test with significance level 0.05 is used to test covid-19, null hypothesis being not having covid-19. If your test come out positive, what do you know about your probability of getting covid-19?
- Let p be a function that sends observed value X to a p-value. What can you say about the c.d.f. of random variable $p(X)$?

8.2 Some Examples

Example 1

X_i i.i.d., Bernoulli distribution with parameter p. $H_0 : p = p_0$, $H_1 : p \neq p_0$.

Likelihood ratio test:

$$\frac{\prod_i p_0^{X_i}(1-p_0)^{1-X_i}}{\sup_p \prod_i p^{X_i}(1-p)^{1-X_i}} \leq k$$

$$\frac{p_0^{\sum_i X_i}(1-p_0)^{n-\sum_i X_i}}{(\frac{1}{n}\sum_i X_i)^{\sum_i X_i}(1-\frac{1}{n}\sum_i X_i)^{n-\sum_i X_i}} \leq k$$

$$\log(LHS) = n\overline{X}(\log(p_0) - \log(\overline{X})) + n(1-\overline{X})(\log(1-p_0) - \log(1-\overline{X}))$$

Which is non positive and 0 iff $\overline{X} = p_0$. So for k close to 1 the test should be of the form:

$$|\overline{X} - p_0| > \epsilon$$

From CLT, if $n >> 1$, under H_0, $\sqrt{\frac{n}{p_0(1-p_0)}} \cdot (\overline{X} - p_0)$ has distribution close to $\mathcal{N}(0,1)$, so the test with significance level α is roughly $|\overline{X} - p_0| \geq \Phi^{-1}(1 - \alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}$ where Φ is the cdf of $\mathcal{N}(0,1)$. The reason is that under H_0, if a test of the form $|Z| \geq c$ has significance α, where $Z \sim \mathcal{N}(0,1)$, then

$$\alpha = P(Z \leq c \cup Z \geq c) = 2P(Z \geq c) = 2(1 - \Phi(c))$$

So $c = \Phi^{-1}(1 - \alpha/2)$. Now let $Z = \sqrt{\frac{n}{p_0(1-p_0)}} \cdot (\overline{X} - p_0)$ one gets the answer.

And the p-value for given $\overline{X} = \overline{x}$ is

$$p = \inf\{\alpha : |\overline{x} - p_0| \geq \Phi^{-1}(1 - \alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}\}$$

$$= 2(1 - \Phi(\sqrt{\frac{n}{p_0(1-p_0)}} \cdot |\overline{x} - p_0|))$$

Suppose $n = 100$, $p_0 = 0.5$, 60 of the X_i has a value of 1 and 40 has a value of 0. We want to test if $H_0 : p = p_0$ is true with a significance level 0.05.

- **Method 1**: The test with significance level 0.05 is roughly $|\overline{X} - p_0| \geq \Phi^{-1}(1 - 0.05/2)\sqrt{\frac{p_0(1-p_0)}{n}} = 0.0980$. $\overline{X} - p_0 = 0.1$ which is larger than the threshold, hence we should reject H_0.
- **Method 2**: Calculate the p-value, we get $p = 2(1 - \Phi(\sqrt{\frac{n}{p_0(1-p_0)}} \cdot |\overline{X} - p_0|)) = 0.0455 \leq 0.05$, so we should reject H_0.

As a review:

- Significance level of a test: highest possible probability of false positive under H_0. It is a increasing function of the threshold k.
- p-value of a possible value of X: the significance level of the test with the lowest threshold that rejects H_0.
- How to test H_0 with given significance level α:
  - Method I: Find the threshold k corresponding to α, test the observed value of X using threshold k.
  - Method II: Find the p-value corresponding to the observed value of X, compare it with α.

Example 2

X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, here μ and σ^2 are both unknown. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.

Likelihood ratio test:

$$\frac{\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2}\prod_i e^{-X_i^2/2\sigma^2}}{\sup_{\mu,\sigma^2}(2\pi\sigma^2)^{-n/2}\prod_i e^{-(X_i-\mu)^2/2\sigma^2}} \leq k$$

Do the optimization we get the optimal μ is $\overline{X}$, the optimal σ^2 in denominator is $\frac{1}{n}\sum_i X_i^2$, and the optimal σ^2 in the numerator is $\frac{1}{n}\sum_i(X_i - \overline{X}) = \frac{1}{n}\sum_i X_i^2 - \overline{X}^2$. (Recall examples we did in MLE).

Hence

$$\log(k) \geq \log(LHS) = -\frac{n}{2}(\log(\frac{1}{n}\sum_i X_i^2) - \log(\frac{1}{n}\sum_i X_i^2 - \overline{X}^2)) + \frac{n}{2} - \frac{n}{2}$$

$$= \frac{n}{2}\log(1 - \frac{\overline{X}^2}{\frac{1}{n}\sum_i X_i^2}) = h(|\frac{\overline{X}}{\sqrt{S^2/n}}|)$$

Where $h(t) = \frac{n}{2}\log(1 - \frac{1}{1+\frac{1}{(n-1)t^2}})$.

So the LRT must be of the form $\left|\frac{\overline{X}}{\sqrt{S^2/n}}\right| \geq M$. From the definition of t-distribution, we know that if

$$X_i \sim \mathcal{N}(0, \sigma^2)$$

Then

$$(n-1)S^2/\sigma^2 \sim \chi(n-1)$$
$$\overline{X}/\sqrt{\sigma^2/n} \sim \mathcal{N}(0,1)$$

So

$$\frac{\overline{X}}{\sqrt{S^2/n}} = \frac{\overline{X}/\sqrt{\sigma^2/n}}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} \sim t(n-1)$$

For any observed value x_i, let $\overline{x}$ and s^2 be the sample mean and sample variance, then the largest threshold M which yield positive result (which corresponds to the smallest k) is:

$$M_0 = \left|\frac{\overline{x}}{\sqrt{s^2/n}}\right|$$

The p-value, which is the significance level of the test with threshold M_0, is:

$$p = P(\left|\frac{\overline{X}}{\sqrt{S^2/n}}\right| \geq M_0 | \frac{\overline{X}}{\sqrt{S^2/n}} \sim t(n-1))$$

$$= 2(1 - T(\left|\frac{\overline{x}}{\sqrt{s^2/n}}\right|)$$

Where T is the cdf of $t(n-1)$.

Example 3

X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, here μ and σ^2 are both unknown. $H_0 : \mu \leq 0$, $H_1 : \mu > 0$.

Likelihood ratio test:

$$\frac{\sup_{\mu\leq 0,\sigma^2}(2\pi\sigma^2)^{-n/2}\prod_i e^{-(X_i-\mu)^2/2\sigma^2}}{\sup_{\mu,\sigma^2}(2\pi\sigma^2)^{-n/2}\prod_i e^{-(X_i-\mu)^2/2\sigma^2}} \leq k$$

The likelihood ratio is 1 if $\sum_i X_i \leq 0$, and the same as Example 2 if $\sum_i X_i > 0$. Hence, the LRT is of the form:

$$\left|\frac{\overline{X}}{\sqrt{S^2/n}}\right| \geq M \text{ and } \overline{X} > 0$$

Hence

$$\frac{\overline{X}}{\sqrt{S^2/n}} \geq M$$

Hence, for given significant level α we let

$$M = T^{-1}(1-\alpha)$$

For given value x_i we can calculate the p-value as

$$p = 1 - T(\frac{\overline{x}}{\sqrt{s^2/n}})$$

Where $\overline{x}$ and s^2 are the calculated sample mean and sample variance.

Some conceptual questions

- Suppose a statistical test with significance level 0.05 is used to test covid-19, null hypothesis being not having covid-19. If your test come out positive, what do you know about your probability of getting covid-19?
- Let p be a function that sends observed value X to a p-value. What can you say about the c.d.f. of random variable $p(X)$ when H_0 is true?

Answer:

- Nothing. Because $P(\text{infected}|\text{tested positive})$ can not be calculated from $P(\text{tested positive}|\text{uninfected})$ which is known to be 0.05.
- The c.d.f. of $p(X)$ at any value $q \in [0,1]$ is bounded from above by q.

Midterm 2 will cover up to here.

8.3 Commonly used statistical tests

Some common hypothesis testing problems have well known tests, which are usually either LRT or approximated LRT. We will illustrate via examples how to use some of the tests in Chapter 13 of the textbook.

Usually a statistical test is stated as follows:

Testing H_0 against H_1, test statistics $z = z(X)$, critical region of size (significance level) α is $z \in D_\alpha$.

For example, for the **One sample, One sided t-test**:

$X_1, \dots X_n$, i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Testing $\mu \leq 0$ against $\mu > 0$. Test statistics $t = \frac{\overline{X}}{\sqrt{S^2/n}}$ Critical region $t \geq T^{-1}(1-\alpha)$, where T is the cdf of $t(n-1)$.

To make use of it, say $n = 5$ and X_i are $-1, 0, 1, 2, 1$. The t statistics can be calculated as 1.1767. $T^{-1}(1-0.05) = 2.1318$, so we can not reject H_0 when significance level is chosen to be 0.05. The minimal α such that 1.1767 is in the critical region is $1 - T(1.1767) = 0.1523$, so the p-value is 0.1523.

If X_i are $0, 1, 2, 3, 4$ however, $t = 2.8284 \geq 2.1318$, so reject H_0 under significance level 0.05. The p-value is 0.0237.

Sometimes we make use of a test indirectly by transforming the observed random variables: from some observed random variables X, we build random variables Y, and use a known test on Y. For example: X_i $i = 1, \dots 10$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, Y_i, $i = 1, \dots 10$, i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, X_i and Y_j are all independent. Want to test if $\mu_1 = \mu_2$. One way to do so would be to consider $Z_i = X_i - Y_i$, which are i.i.d.normal, and test if their expectation is 0.

This approach usually won't give us the most powerful test as we are losing information during the transformation. However in many situations this is good enough.

8.3.1 Tests about expectation and variance

X_i i.i.d. $i = 1, \dots, n, \sim \mathcal{N}(\mu, \sigma^2)$.

- Test $\mu = \mu_0$ against $\mu \neq \mu_0$. $t = \frac{\overline{X} - \mu_0}{\sqrt{S^2/n}}$, critical region $|t| \geq F^{-1}_{t(n-1)}(1 - \alpha/2)$, where $F_{t(n-1)}$ is the c.d.f. of $t(n-1)$.
- Test $\mu \leq \mu_0$ against $\mu > \mu_0$, same t as above, critical region $t \geq F^{-1}_{t(n-1)}(1 - \alpha, n-1)$.
- Test $\sigma^2 = \sigma_0^2$: $\chi^2 = (n-1)S^2/\sigma_0^2$, critical region $\chi^2 \in (0, F^{-1}_{\chi(n-1)}(\alpha/2)] \cup [F^{-1}_{\chi(n-1)}(1 - \alpha/2), \infty)$
- Test $\sigma^2 \leq \sigma_0^2$ against $\sigma^2 > \sigma_0^2$: χ same as above, critical region $\chi^2 \geq F^{-1}_{\chi(n-1)}(1 - \alpha)$.
- Test $\sigma^2 \geq \sigma_0^2$ against $\sigma^2 < \sigma_0^2$: χ same as above, critical region $\chi^2 \leq F^{-1}_{\chi(n-1)}(\alpha)$.

X_i, $i = 1, \dots n_1$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, Y_i $i = 1, \dots n_2$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, X_i, Y_j indep.

- Test for $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$, knowing σ_1^2 and σ_2^2. $z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_2^2/n_1 + \sigma_2^2/n_2}}$. Critical region $|z| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)$.
- If σ_i^2 unknown but number of samples is large, can approximate them with S^2.
- $\sigma_1^2 = \sigma_2^2$ but unknown, test $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(1/n_1 + 1/n_2) \cdot \left(\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2} \right)}}$$

  Critical region $|t| \geq F^{-1}_{t(n_1+n_2-2)}(1 - \alpha/2)$.

- One sided tests are similar.

X_i, $i = 1, \dots n_1$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, Y_i $i = 1, \dots n_2$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, X_i, Y_j indep.

- Testing $\sigma_1^2 = \sigma_2^2$ against $\sigma_1^2 \neq \sigma_2^2$. $f = S_X^2/S_Y^2$. Critical region $f \in (0, F^{-1}_{F(n_1-1,n_2-1)}(\alpha/2)] \cup [F^{-1}_{F(n_1-1,n_2-1)}(1-\alpha/2), \infty)$.
- One sided tests are similar.

One can check by calculation that all these tests have the significance level α.

The final exam is open book so one doesn't need to memorize any of these statistical tests, just need to know how to use them would be enough!

8.3.2 Pearson's χ^2-test for Goodness of fit

X_i i.i.d. taking values at $\{1, 2, \dots m\}$. Test for null hypothesis: $P(X = j) = e_j$, where $e_j = f(j, \theta_1, \dots, \theta_k)$. Let n_j be the number of X_i taking value j. Then likelihood ratio test gives:

$$\frac{\sup_{\theta_1,\dots,\theta_k} \prod_j e_j^{n_j}}{\sup_{p_j, \sum_j p_j = 1} \prod_j p_j^{n_j}} \leq k$$

The optimal p_j is n_j/n where $n = \sum_j n_j$. So

$$\log(LHS) = \sum_j n_j(-\log(\frac{n_j/n}{\hat{e_j}})) = \sum_j n_j(-\log(1 + \frac{n_j - n\hat{e_j}}{n\hat{e_j}}))$$

Taylor expansion at $n_j = ne_j$, we get approximated LRT:

$$\sum_j \frac{(n_j - n\hat{e_j})^2}{\hat{e_j}} \geq m$$

When n is large, and with some additional assumptions, the test statistics $\sim \chi^2(m - k - 1)$.

Midterm 2 Review

- MOM
- Bayesian-based point estimates: expectation of posterior, MAP, etc.
- Neyman-Pearson test (the proof that it is optimal will not be tested in the exam)
- Likelihood ratio test
- Significance, power, and p-value

9 Confidence interval

Setting: X has p.d.f. (or p.d.) $f(x, \theta)$, where θ is unknown.

- **Point estimate**: find a random variable $\hat{\theta}$ based on X, which is close to θ.
- **Hypothesis testing**: given θ_0, we can tell how unlikely it is to get the observed value of X if $\theta = \theta_0$.
- **Confidence interval** is related to both of these concepts:
  - Conceptually, confidence interval is an extension of point estimate: this is a random variable taking value in the set of sets, such that θ is in it with probability $1 - \alpha$.
  - Mathematically, confidence intervals are equivalent to certain types of statistical tests.

The $1 - \alpha$-**confidence interval** of θ is a set $I(X)$ depending on X, such that for any possible value of θ, $P(\theta \in I(X)|\theta) = 1 - \alpha$.

Equivalence between confidence intervals and statistical tests:

- If $X \in D(\theta_0)$ is a statistical test of the null hypothesis $H_0 : \theta = \theta_0$, which has significance level α. Then $I(X) = \{\theta_0 : X \notin D(\theta_0)\}$ is a $1 - \alpha$ confidence interval for θ.
- If $I(X)$ is a $1 - \alpha$ confidence interval for X, then $\theta_0 \notin I(X)$ is a statistical test of the null hypothesis $H_0 : \theta = \theta_0$.

Proof:

- Suppose $P(X \in D(\theta)|\theta) = \alpha$. Let

$$I(X) = \{\theta : X \notin D(\theta)\}$$

then

$$\begin{aligned} P(\theta \in I(X)|\theta) &= P(X \notin D(\theta)|\theta) \\ &= 1 - P(X \in D(\theta)|\theta) = 1 - \alpha \end{aligned}$$

- Suppose $P(\theta \in I(X)|\theta) = 1 - \alpha$. Let

$$D(\theta_0) = \{X : \theta_0 \notin I(X)\}$$

Then

$$\begin{aligned} P(X \in D(\theta)|\theta) &= P(\theta \notin I(X)|\theta) \\ &= 1 - P(\theta \in I(X)|\theta) = \alpha \end{aligned}$$

In some textbooks the CI is defined as $P(\theta \in I(X)|\theta) \geq 1-\alpha$, then, they should correspond to statistical tests of significance level $\leq \alpha$. They will not be the focus of this course, but in case we need to mention them in examples, let's call them *CI with confidence level at least* $1-\alpha$.

Example 1: X normal distribution with expectation μ and variance 1. Find the 0.95 confidence interval for μ.

Likelihood ratio test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$:

$$\frac{e^{-(X-\mu_0)^2/2}}{\sup_\mu e^{-(X-\mu)^2/2}} \leq k$$

The optimal μ is X, so the LRT is

$$|X-\mu_0| \geq \sqrt{-2\log(k)}$$

Let Φ be the c.d.f. of standard normal distribution. The significance level is the probability of success under null hypothesis, and under null hypothesis, $X-\mu_0$ is standard normal. So,

$$\alpha = 2(1-\Phi(\sqrt{-2\log(k)}))$$

So the test

$$|X-\mu_0| \geq \Phi^{-1}(0.975)$$

Is a test with significance level α, the confidence interval is

$$I(X) = \{\mu : |X-\mu| \leq \Phi^{-1}(0.975)\} = [X-\Phi^{-1}(0.975), X+\Phi^{-1}(0.975)]$$

One sided confidence interval: Sometimes we want the confidence interval to be one sided, like $I = [a(X), \infty)$. The statistical test associated to it should be $mu < a(X)$, in other words, it should only reject null hypothesis $\mu = \mu_0$ if μ_0 is too small. Hence, let's consider $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$, then the LRT becomes

$$\frac{e^{-(X-\mu_0)^2/2}}{\sup_{\mu \geq \mu_0} e^{-(X-\mu)^2/2}} \leq k$$

So the optimal μ is μ_0 if $X \leq \mu_0$, X if $X > \mu_0$. So the test is

$$X-\mu_0 \geq \sqrt{-2\log(k)}$$

When $k < 1$, and everything when $k = 1$. So

$$\alpha = 0.05 = 1-\Phi(\sqrt{-2\log(k)})$$

$$X-\mu_0 \geq \Phi^{-1}(0.95)$$

$$I(X) = \{\mu : X-\mu \leq \Phi^{-1}(0.95)\} = [X-\Phi^{-1}(0.95), \infty)$$

In general, if $I(X)$ is one-sided, e.g. of the form $[A(X), \infty)$, the corresponding statistical test must be of the form $\theta_0 \leq \alpha(X)$. In other words, the null hypothesis can be rejected is only if θ_0 is too small, i.e. when $\theta_0 < \theta$. Hence the power function of the test must be no more than α on $(-\infty, \theta_0]$, and one can pick the alternative hypothesis as $\theta_0 < \theta$.

- As an exercise, read Chapter 11 and Chapter 13. For every statistical test in 13.2-13.6, find the corresponding confidence interval, if there are any, from 11.2-11.7.
- True or false: suppose based on the statistics up to today, the reproductive number R_0 of covid-19 has a 95% confidence interval $[2.1, 2.5]$. Then the probability of R_0 being between 2.1 and 2.5 is 0.95. (Answer: False)
- True or false: suppose after the covid-19 outbreak we found a very good model for estimating the R_0 of an epidemic, and this model gives a 95% confidence interval to the R_0 of the next pandemic. Then, the probability of R_0 lying in this confidence interval is 0.95. (Answer: True)

The next part in blue will not be in the exam:

One can create some analogy of hypothesis testing and confidence interval under Bayesian statistics as well, which is conceptually much simpler but completely different from the ones we learn in the non-Bayesian setting:

- Recall that the output of Bayesian statistics is the posterior, i.e. conditional distribution of θ conditioned at X.
- For a hypothesis $H : \theta \in D$, we can calculate its probability under this posterior $P(\theta \in D|X)$, and reject it when this probability is small.
- The $1-\alpha$-**credible interval** is $J(X)$ such that $P(\theta \in J(X)|X) = 1 - \alpha$.

Example 2: normal approximation of binomial distribution
$X \sim B(n, p)$, $n >> 1$, p not too close to 0 or 1. Want CI of p.
From what we learned some weeks ago, we have an approximated LRT based on CLT which says that the test for $p = p_0$ against $p \neq p_0$ with significance level α is

$$|X/n - p_0| \geq \sqrt{\frac{p_0(1-p_0)}{n}} \cdot \Phi^{-1}(1 - \alpha/2)$$

Where Φ is the cdf of standard normal.

So the approximated $1 - \alpha$ CI is

$$\{p : |X/n - p| \leq \sqrt{\frac{p(1-p)}{n}} \cdot \Phi^{-1}(1 - \alpha/2)\} = [p_1, p_2]$$

Where

$$X/n - p_1 = \sqrt{\frac{p_1(1-p_1)}{n}} \cdot \Phi^{-1}(1-\alpha/2)$$

$$p_2 - X/n = \sqrt{\frac{p_2(1-p_2)}{n}} \cdot \Phi^{-1}(1-\alpha/2)$$

Because $n >> 0$, $p_1, p_2 \approx X/n$, we have

$$[p_1, p_2] = [X/n - \sqrt{\frac{X(n-X)}{n^3}} \cdot \Phi^{-1}(1-\alpha/2),$$

$$X/n + \sqrt{\frac{X(n-X)}{n^3}} \cdot \Phi^{-1}(1-\alpha/2)]$$

Example 3: Exponential distribution

X has p.d.f. $f(x) = \begin{cases} ce^{-cx} & x \geq 0 \\ 0 & x \leq 0 \end{cases}$. Find the one sided CI of the form $(0, A]$.

LRT with $H_0 : c = c_0$ and $H_1 : c < c_0$.

$$\frac{c_0 e^{-c_0 X}}{\sup_{c \leq c_0} ce^{-cX}} \leq k$$

If $X \leq 1/c_0$ the LHS is 1, if $X > 1/c_0$, the optimal c in denominator is $1/X$, and we get

$$\log(c_0) - Xc_0 \leq \log(k) - \log(X) - 1$$

$$c_0 X - \log(X) \geq \log(c_0) + 1 - \log(k)$$

The LHS is an increasing function, so the test must be of the form $X \geq M$. If we want the significance level to be α,

$$\alpha = P(X \geq M | c = c_0) = \int_M^\infty f(s)ds$$

So $M = -\log(\alpha)/c_0$. The one sided CI is now

$$\{c : X \leq -\log(\alpha)/c\} = (0, -\log(\alpha)/X]$$

Example 4: Making use of the F-test

Suppose there are 2 independent i.i.d. normal samples X_i, $i = 1, \ldots n_1$, Y_j, $j = 1, \ldots n_2$, with variance σ_1^2 and σ_2^2 respectively. Want the one sided CI of σ_1^2/σ_2^2 of the form $(0, A]$.

$H_0 : \sigma_1^2/\sigma_2^2 = r$, $H_1 : \sigma_1^2/\sigma_2^2 < r$. Let $Y_j' = r^{1/2} Y_j$, then the test is for $Var(X_i) = Var(Y_j')$ against $Var(X_i) \leq Var(Y_j')$, use one sided F-test with significance level α is:

$$S_X^2/S_{Y'}^2 = S_X^2/(rS_Y^2) \leq F^{-1}_{F(n_1-1, n_2-1)}(\alpha)$$

So CI with CL $1-\alpha$ is

$$\{r : S_X^2/(rS_Y^2) \geq F^{-1}_{F(n_1-1,n_2-1)}(\alpha)\}$$

$$= (0, \frac{S_X^2}{S_Y^2 F^{-1}_{F(n_1-1,n_2-1)}(\alpha)}]$$

In the textbook they used the relationship $F^{-1}_{F(n_1-1,n_2-1)}(\alpha) = (F^{-1}_{F(n_2-1,n_1-1)}(1-\alpha))^{-1}$.

The CI for t-, χ^2- etc. tests are analogous.

When $n_1 >> 1$, $n_2 >> 1$, CLT allow us to do normal approximation for the χ^2 distribution. This can also be used to derive approximated CI for the ratio of variance:

By definition, $\chi^2(k)$ is the squared sum of k standard normal, so CLT tells us, if $X \sim \chi^2(k)$, when $k \to \infty$, $\frac{X-k}{\sqrt{2k}} \to$ standard normal.

$$\frac{(n_1-1)S_X^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

So

$$\frac{(n_1-1)(S_X^2-\sigma_1^2)}{\sigma_1^2\sqrt{2n_1-2}} \to \mathcal{N}(0,1)$$

Similarly

$$\frac{(n_2-1)(S_Y^2-\sigma_2^2)}{\sigma_2^2\sqrt{2n_2-2}} \to \mathcal{N}(0,1)$$

Hence the distribution of $S_X^2 - rS_Y^2$ is approximately $\mathcal{N}(\sigma_1^2 - r\sigma_2^2, \frac{2\sigma_1^4}{n_1-1} + \frac{2\sigma_2^4}{n_2-1}) \approx \mathcal{N}(\sigma_1^2 - r\sigma_2^2, \frac{2S_X^4}{n_1-1} + \frac{2r^2S_Y^4}{n_2-1})$, so the test is

$$S_X^2 - rS_Y^2 \leq \Phi^{-1}(1-\alpha)\sqrt{\frac{2S_X^4}{n_1-1} + \frac{2r^2S_Y^4}{n_2-1}}$$

You can now use this to get a corresponding approximated CI.

Of course there are many other ways to get CI or, through CLT, approximated CI. Can you write down a few more for this problem?

Extra Example: Normal approximation X_i i.i.d., $i = 1, \ldots n_1$, Y_j i.i.d., $j = 1, \ldots n_2$, both have bounded variance. Find the approximated CI of $E[X_i] - E[Y_j]$ using CLT on both X_i and Y_j, where the variances are estimated via LLN using S^2.

CLT plus LLN implies that

$$\frac{\sum_i (X_i - E[X_i])}{\sqrt{S_X^2 n_1}} \to \mathcal{N}(0,1)$$

$$\frac{\sum_j (Y_j - E[Y_j])}{\sqrt{S_Y^2 n_2}} \to \mathcal{N}(0,1)$$

So we can approximate $\overline{X}$ and $\overline{Y}$ by $\mathcal{N}(E[X_i], S_X^2/n_1)$ and $\mathcal{N}(E[Y_i], S_Y^2/n_2)$ respectively, and $\overline{X}-\overline{Y}$ can be approximated by $(E[X_i]-E[Y_i], S_X^2/n_1+S_Y^2/n_2)$, so the CI is $[\overline{X}-\overline{Y}-F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)\sqrt{S_X^2/n_1+S_Y^2/n_2}, \overline{X}-\overline{Y}+F^{-1}_{\mathcal{N}(0,1)}(1-\alpha/2)\sqrt{S_X^2/n_1+S_Y^2/n_2}]$.

10 Linear Regression

10.1 Single variable LR

Setting: $x_1, \dots x_n$ real numbers, $Y_1, \dots Y_n$ independent, $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$. How do we estimate c and σ^2?

10.1.1 MLE for c and σ^2

Likelihood function:

$$L = \prod_i f_{Y_i}(Y_i) = (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}$$

$$\log(L) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (Y_i - cx_i)^2$$

$$\frac{\partial}{\partial c}\log(L) = -\frac{1}{2\sigma^2}\sum_i (2x_iY_i - 2cx_i^2)$$

$$\hat{c}_{MLE} = \frac{\sum_i x_iY_i}{\sum_i x_i^2}$$

$$\frac{\partial}{\partial \sigma^2}\log(L) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (Y_i - cx_i)^2$$

$$\hat{\sigma^2}_{MLE} = \frac{1}{n}\sum_i (Y_i - \hat{c}_{MLE}x_i)^2 = \frac{1}{n}(\sum_i Y_i^2 - (\sum_i x_iY_i)^2 / \sum_i x_i^2)$$

10.1.2 Prior on c, knowing $\sigma^2 = 1$

Suppose $\sigma^2 = 1$, c has a prior $\mathcal{N}(0, \lambda)$.

Posterior will be proportional to

$$g(c) = \frac{1}{\sqrt{2\pi\lambda}} e^{-c^2/(2\lambda)} (2\pi)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/2}$$

So

$$c|Y_i \sim \mathcal{N}(\frac{\sum_i X_iY_i}{\sum_i x_i^2 + 1/\lambda}, (\sum_i x_i^2 + 1/\lambda)^{-1})$$

10.1.3 Prior on c and σ^2

Suppose σ^2 has a prior $f(s) = \begin{cases} \alpha e^{-\alpha s} & s \geq 0 \\ 0 & s < 0 \end{cases}$, c has a prior $\mathcal{N}(0, \lambda\sigma^2)$.

Posterior will be proportional to

$$g(c, \sigma^2) = \frac{\alpha}{\sqrt{2\pi\lambda}} e^{-c^2/(2\lambda\sigma^2)} e^{-\alpha\sigma^2} (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}$$

MAP estimate:

$$\hat{c}_{MAP} = \frac{\sum_i X_i Y_i}{\sum_i x_i^2 + 1/\lambda}$$

$$\hat{\sigma^2}_{MAP} = \frac{2(\sum_i (Y_i - \hat{c}_{MAP} x_i)^2 + \hat{c}_{MAP}^2/\lambda)}{n + \sqrt{n^2 + 8\alpha(\sum_i (Y_i - \hat{c}_{MAP} x_i)^2 + \hat{c}_{MAP}^2/\lambda)}}$$

Similarly we can calculate the expectation of c and σ^2 under posterior distribution. It is evident that $E[c|Y_i] = \hat{c}_{MAP}$.

$$E[\sigma^2|Y_i] = \frac{\int_0^\infty d\sigma^2 \int_{-\infty}^\infty \sigma^2 g(c, \sigma^2)}{\int_0^\infty d\sigma^2 \int_{-\infty}^\infty g(c, \sigma^2)}$$

10.1.4 Test for hypothesis $H_0 : c = 0$ against $H_1 : c \neq 0$, knowing $\sigma^2 = 1$

LRT:

$$\frac{(2\pi)^{-n/2} e^{-\sum_i Y_i^2/2}}{\sup_c (2\pi)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/2}} \leq k$$

The optimal c is $\frac{\sum_i x_i Y_i}{\sum_i x_i^2}$ from (1), so

$$-\sum_i Y_i^2 + \sum_i (Y_i - \frac{\sum_i x_i Y_i}{\sum_i x_i^2} \cdot x_i)^2 \leq 2\log k$$

$$\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} \geq -2\log k$$

So the test should be

$$|\sum_i x_i Y_i| \geq M$$

Under null hypothesis $\sum_i x_i Y_i \sim \mathcal{N}(0, \sum_i x_i^2)$, so significance level is

$$\alpha = 2(1 - F_{\mathcal{N}(0,1)}(\frac{M}{\sqrt{\sum_i x_i^2}}))$$

The test with significance level α should be

$$|\sum_i x_i Y_i| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2) \sqrt{\sum_i x_i^2}$$

p-value for $Y_i = y_i$ is

$$p = 2(1 - F_{\mathcal{N}(0,1)}(\frac{|\sum_i x_i y_i|}{\sqrt{\sum_i x_i^2}}))$$

10.1.5 CI for c, knowing $\sigma^2 = 1$

We need statistical test for $H_0 : c = c_0$ against $H_0 : c \neq c_0$. Let $Z_i = Y_i - c_0 x_i$, then $Z_i \sim \mathcal{N}(c - c_0)x_i, 1)$. Now make use of the test in (4), we get

$$|\sum_i x_i Z_i| = |\sum_i x_i Y_i - \sum_i c_0 x_i^2| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)\sqrt{\sum_i x_i^2}$$

So the corresponding $1 - \alpha$ CI for c is

$$\{c : |\sum_i x_i Y_i - \sum_i c x_i^2| \leq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)\sqrt{\sum_i x_i^2}\}$$

$$= [\frac{\sum_i x_i Y_i}{\sum_i x_i^2} - \frac{F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)}{\sqrt{\sum_i x_i^2}}, \frac{\sum_i x_i Y_i}{\sum_i x_i^2} + \frac{F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)}{\sqrt{\sum_i x_i^2}}]$$

10.1.6 Test for hypothesis $H_0 : c = 0$ against $H_1 : c \neq 0$, with unknown σ^2

LRT:

$$\frac{\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\sum_i Y_i^2/(2\sigma^2)}}{\sup_{c,\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}} \leq k$$

$$\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\sum_i Y_i^2/(2\sigma^2)} = (2\pi \cdot \frac{\sum_i Y_i^2}{n})^{-n/2} e^{-n/2}$$

$$\sup_{c,\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)} = (2\pi \cdot \frac{\sum_i (Y_i - \hat{c}x_i)^2}{n})^{-n/2} e^{-n/2}$$

Where $\hat{c} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$. So the test becomes

$$\frac{\sum_i (Y_i - \hat{c}x_i)^2}{\sum_i Y_i^2} \leq k^{2/n}$$

$$\sum_i (Y_i - \hat{c}x_i)^2 = \sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$$

So we can rewrite the test as

$$|\frac{\sum_i x_i Y_i / \sqrt{\sum_i x_i^2}}{\sqrt{\frac{1}{n-1} \cdot (\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}}| \leq M$$

By calculation (using multivariable calculus and linear algebra) we can see that, under null hypothesis, $\frac{\sum_i Y_i^2}{\sigma^2} \sim \chi^2(n)$, $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$ is independent from $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$, and $\frac{1}{\sigma^2} \cdot \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} \sim \chi^2(1)$. So,

$$\frac{\sum_i x_i Y_i / \sqrt{\sum_i x_i^2}}{\sqrt{\frac{1}{n-1} \cdot (\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}} \sim t(n-1)$$

The M for significance level α is $F^{-1}_{t(n-1)}(1-\alpha/2)$.

10.1.7 CI for c, unknown σ^2

Use the same technique as in (5), and the test in (6), we get

$$[\frac{\sum_i x_i Y_i}{\sum_i x_i^2} - F^{-1}_{t(n-1)}(1-\alpha/2) \cdot \frac{\sqrt{(\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}}{\sqrt{(n-1)\sum_i x_i^2}}, \frac{\sum_i x_i Y_i}{\sum_i x_i^2} + F^{-1}_{t(n-1)}(1-\alpha/2) \cdot \frac{\sqrt{(\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}}{\sqrt{(n-1)\sum_i x_i^2}}]$$

10.1.8 Digression: Independence of residue and regression coefficient

This part is just for those who remember linear algebra and multivariable calculus.

We will prove the following: if Y_i i.i.d. normal, $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$ is independent from $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$.

Proof: Let $c_1 = [x_i / \sqrt{\sum_i x_i^2}]^T \in \mathbb{R}^n$. $|c_1| = 1$, so we can find an orthonormal basis $\{c_1, \ldots, c_n\}$ of $\mathbb{R}^n$. Let $C = [c_1 \ldots . c_n]^T$, $Y = [Y_1, \ldots Y_n]^T$, $Z = CY$. Because Y_i are i.i.d. normal, the p.d.f. of Y is $f(y) = 2\pi\sigma^{2-n/2} e^{-\frac{1}{2} y^T y}$, so for any set $A \subset \mathbb{R}^n$,

$$P(Z \in A) = P(CY \in A) = \int_{C^{-1}A} (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} y^T y} dy$$

$$= \int_A (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} z^T z} dz$$

So Z_i i.i.d. $\mathcal{N}(0, \sigma^2)$.

By calculation it is easy to verify that $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} = \sum_{i=2}^n Z_i^2$ and $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} = Z_1^2$, hence they must be independent. The same calculation works for $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ as well by change of variable $Y_i = cx_i + Y_i'$.

10.1.9 Test for $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$

Likelihood ratio test:

$$\frac{\sup_c (2\pi\sigma_0^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2 / (2\sigma_0^2)}}{\sup_{c,\sigma^2} (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2 / (2\sigma^2)}} \leq k$$

The optimal c is $\frac{\sum_i Y_i x_i}{\sum_i x_i^2}$. Let $r^2 = \sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2$, log of LHS is

$$-\frac{n}{2}\log(\sigma_0^2) - \frac{r^2}{2\sigma_0^2} + \frac{n}{2}\log(r^2/n) + \frac{n}{2} \leq \log(k)$$

Hence the critical region should be of the form $r^2/n \geq \sigma_0^2 A$ or $r^2/n \leq \sigma_0^2 B$ for some positive numbers $0 < B < 1 < A$. By similar argument as in the previous slides, under H_0, $\frac{1}{\sigma_0^2} r^2 \sim \chi^2(n-1)$, so significance level

$$\alpha = F_{\chi^2(n-1)}(nB) + 1 - F_{\chi^2(n-1)}(nA)$$

$$\log(A) - A = \log(B) - B$$

In practice, we usually just ignore the second equation and let $F_{\chi^2(n-1)}(nB) = 1 - F_{\chi^2(n-1)}(nA) = \alpha/2$, hence the test is

$$\frac{1}{\sigma_0^2}\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2 \notin [F^{-1}_{\chi^2(n-1)}(\alpha/2), F^{-1}_{\chi^2(n-1)}(1-\alpha/2)]$$

10.1.10 CI for σ^2

Using the test on the previous slide, we have the CI:

$$\left[\frac{\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2}{F^{-1}_{\chi^2(n-1)}(1-\alpha/2)}, \frac{\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2}{F^{-1}_{\chi^2(n-1)}(\alpha/2)}\right]$$

Everything below will be beyond the scope of final exam.

10.1.11 Logistic regression

Setting Y_i independent, $Y_i \sim Bernoulli(\frac{e^{cx_i}}{1+e^{cx_i}})$.

Likelihood function

$$L = \prod_i \frac{y_i e^{cx_i} + (1-y_i)}{1+e^{cx_i}}$$

It is easy to see that $\log(L)$ is concave w.r.t. c, hence any local maximum is the MLE, and we can use convex optimization to calculate the optimal c.

This is a first example of Generalized Linear Models (GLM).

10.2 Higher dimensional linear regression

Setting: $x_1, \ldots x_n \in \mathbb{R}^d$, $Y_1, \ldots Y_n$ independent, $\beta \in \mathbb{R}^d$, $Y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$. How do we estimate β and σ^2?

MLE: Log likelihood is

$$\log(L) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (Y_i - \beta^T x_i)^2$$

So

$$\hat{\beta}_{MLE} = \arg\min_{\beta} \sum_i (Y_i - \beta^T x_i)^2$$

Take derivative, we get:

$$2\sum_i (Y_i - \hat{\beta}^T x_i) x_i = 0$$

$$\hat{\beta} = (\sum_i x_i x_i^T)^{-1} (\sum_i Y_i x_i)$$

The MLE for σ^2 is the same as the univariant case.

A special case is linear regression with constant term:

$x_1, \dots x_n \in \mathbb{R}$, $Y_1, \dots Y_n$ independent, $\beta \in \mathbb{R}^d$, $Y_i \sim \mathcal{N}(d + cx_i, \sigma^2)$. Find MLE for c and d.

Let $x_i' = [1, x_i]^T$, $\beta = [d, c]$, then use the formula on the previous slide, we get

$$[\hat{d}, \hat{c}]^T = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

So

$$\hat{d} = \frac{\sum_i x_i^2 \sum_i Y_i - \sum_i x_i \sum_i x_i Y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$\hat{c} = \frac{-\sum_i x_i \sum_i Y_i + n \sum_i x_i Y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

10.2.1 Ridge Regression

Suppose $\sigma = \sigma_0$, and we add a prior to β as $\beta \sim \mathcal{N}(0, \lambda\sigma_0^2 I_d)$, log of posterior will be, up to a constant,

$$-\frac{\beta^T \beta}{2\lambda\sigma^2} - \frac{1}{2\sigma^2} \sum_i (Y_i - \beta^T x_i)^2$$

So the MAP estimate for β is

$$\hat{\beta} = \arg\min \beta \sum_i (Y_i - \beta^T x_i)^2 + \frac{1}{\lambda} \beta^T \beta$$

$$\hat{\beta} = (\sum_i x_i x_i^T + I_d/\lambda)^{-1} (\sum_i Y_i x_i)$$

This works even when $n < d$.

10.2.2 Linear Mixed Model

Ridge regression has an alternative formulation as follows: $x_i, x \in \mathbb{R}^d$, Y_i, Y satisfies joint distribution $\mathcal{N}(0, \sigma^2(K + \delta I))$, with known σ^2 and δ, and where $K = [x_1, \dots x][x_1, \dots x]^T$. Find the conditional expectation of Y with known $Y_1, \dots Y_n$. The log of joint p.d.f. of $[Y_1, \dots, Y_n, Y]^T$ is, up to a constant, proportional to

$$-\frac{1}{2}[y_1, \dots y_n, y](K + \delta I)^{-1}[y_1, \dots y_n, y]^T$$

Let $K_0 = [x_1, \dots x_n][x_1, \dots x_n]^T$, $b = [x_1^T x, \dots x_n^T x]$, then $K + \delta I = \begin{bmatrix} K_0 + \lambda I & b^T \\ b & x^T x + \lambda \end{bmatrix}$, hence

$$(K + \delta I)^{-1} = \begin{bmatrix} * & B^T \\ B & C \end{bmatrix}$$

Where

$$C = (x^T x - b(K_0 + \delta I)^{-1} b^T)^{-1}$$

$$B = -(x^T x - b(K_0 + \delta I)^{-1} b^T)^{-1} b(K_0 + \delta I)^{-1}$$

So the conditional distribution for y is normal, and the expectation is

$$\hat{y} = -\frac{1}{C} B \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = b(K_0 + \delta I)^{-1} \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$$

Let $X = [x_1, \dots x_n]$, $Y = [y_1, \dots y_n]^T$, then this equals

$$x^T X(X^T X + \delta I)^{-1} Y = x^T (XX^T + \delta I)^{-1} XY$$

$$= x^T (\sum_i x_i x_i^T + \delta I_d)^{-1} (\sum_i y_i x_i)$$

So δ takes the role of $\frac{1}{\lambda}$ earlier. This model allows us to get a value for $\frac{1}{\lambda}$, by setting it as $\hat{\delta}_{MLE}$. This is the simplest case of a family of statistical models called mixed models.

10.2.3 Some questions to think about

- Suppose $x_i \in \{1, 2, 3\}$, how do you check $H_0 : Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ against $H_1 : Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$ where f is an arbitrary function?
- How do you check that $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ in general?