Statistical Model

$(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$

Parametric Model Dim of Θ is finite. $\Theta \subseteq \mathbb{R}^d$

Identifiable Parameter The parameter θ is called *identifiable* if and only if the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective (Veryfy by solve CDF/PMF and see if uniquely determined by θ).,

$$\theta \neq \theta' \implies \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently,

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies \theta = \theta'.$$

Quantile $F(q_\alpha) = P(X \leq q_\alpha) = 1 - \alpha$

Convergence and Inequality

Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Law of Large Numbers

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}, a.s.} \mu.$$

Central Limit Theorem

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right).$$

Multi CLT Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[X] = \mu$, $\mathrm{Cov}(X) = \Sigma$, then

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

Hoeffding's Inequality Let $X, X_1, \ldots X_n$ be i.i.d. random variables such that $\mathbb{E}[X] = \mu$ and $X \in [a, b]$ almost surely. Then,

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}} \quad \forall \epsilon > 0$$

Markov Inequality If $X \geq 0$ and $a > 0$, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$

Chebyshev Inequality Variable is unlikely to be far from the mean. $\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

Almost Surely (a.s.) Convergence

$$T_n \xrightarrow[n\to\infty]{a.s.} T \iff \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1$$

Convergence in Probability

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \iff \mathbb{P}\left(|T_n - T| \geq \epsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall \epsilon > 0$$

Convergence in Distribution

$$T_n \xrightarrow[n\to\infty]{(d)} T \iff \mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)]$$

for all continuous and bounded function f.

The Delta Method

Let $(Z_n)_{n\geq 1}$ be a sequence of random variables that satisfies

$$\sqrt{n}\left(Z_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right)$$

for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at the point θ. Then

$$\sqrt{n}\left(g(Z_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \left(g'(\theta)\right)^2 \sigma^2\right).$$

Multivariate Delta Method

Let $(T_n)_{n\geq 1}$ sequence of random vectors in $\mathbb{R}^d$ such that

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma),$$

for some $\theta \in \mathbb{R}^d$ and some covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ $(k \geq 1)$ be continuously differentiable at θ. Then,

$$\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \nabla g(\theta)^\intercal \Sigma \nabla g(\theta)\right),$$

where $\nabla g(\theta) = \frac{\partial g(\theta)}{\partial \theta} = \left(\frac{\partial g_j}{\partial \theta_i}\right)_{\substack{1\leq i\leq d \\ 1\leq j\leq k}} \in \mathbb{R}^{d\times k}$

Estimation

Consistent Estimator

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}\ (\text{resp. } a.s.)} \theta \quad (\text{w.r.t. } \mathbb{P}).$$

Asymptotic Normal

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right)$$

Jensen's Inequality If the function $f(x)$ is convex,

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Total Variation

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A\subset E} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|$$

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{x\in E} |p_\theta(x) - p_{\theta'}(x)|$$

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int |f_\theta(x) - f_{\theta'}(x)|\, dx$$

Kullback-Leibler(KL) Divergence : positive and definite (0 means same distribution), but not meet triagular inequality and symmetrical

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \sum_{x\in E} p_\theta(x) \log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) & \text{if } E \text{ is discrete} \\ \int_E f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx & \text{if } E \text{ is continuous} \end{cases}$$

MLE

MLE estimator

$$\hat{\theta}_n^{\mathrm{MLE}} = \operatorname*{argmax}_{\theta\in\Theta} L(X_1, \ldots, X_n, \theta),$$

$$= \operatorname*{argmax}_{\theta\in\Theta} \sum_{i}^{n} log(f_\theta(X_i))$$

Fisher Information On average how curved is the log-likelihood function

$$\ell(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d.$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the Fisher information of the statistical model is defined as

$$I(\theta) = \mathbb{E}\left[\nabla\ell(\theta)\nabla\ell(\theta)^\intercal\right] - \mathbb{E}\left[\nabla\ell(\theta)\right]\mathbb{E}\left[\nabla\ell(\theta)\right]^\intercal = -\mathbb{E}\left[\mathbb{H}\ell(\theta)\right].$$

If $\Theta \subset \mathbb{R}$, we get

$$I(\theta) = \mathrm{Var}\left[\ell'(\theta)\right] = -\mathbb{E}\left[\ell''(\theta)\right].$$

Asymptotical Normality

1. The parameter is identifiable.
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ.
3. θ^* is not on the boundary of Θ.
4. $I(\theta)$ is invertible in a neighborhood of θ^*.
5. A few more technical conditions.

Then, $\hat{\theta}_n^{\mathrm{MLE}}$ satisfies

- $\hat{\theta}_n^{\mathrm{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- $\sqrt{n}\left(\hat{\theta}_n^{\mathrm{MLE}} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, I^{-1}(\theta^*)\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

EM algorithm

Randomly initialize all parameters θ for latent variable Z and observable variable X

1. **E-step:** (Complete data by replacing Z_i with conditional expectation $\mathbb{E}[Z_i|X_i]$ when Z_i is Bernoulli = $\mathbb{P}(Z_i = 1|X_i)$)

$$p(j|i) = \frac{p_j \mathcal{N}\left(\mathbf{x}^{(i)}; \mu^{(j)}, \sigma_j^2 \mathbf{I}\right)}{p(\mathbf{x}|\theta)}$$

where likelihood

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{K} p_j \mathcal{N}\left(\mathbf{x}^{(i)}; \mu^{(j)}, \sigma_j^2 \mathbf{I}\right)$$

2. **M-step:** (Plug Z_i in likelihood and optimize with respect to parameter of X)

$$\widehat{n}_j = \sum_{i=1}^{n} p(j|i), \widehat{p}_j = \frac{\widehat{n}_j}{n}$$

$$\widehat{\mu}^{(j)} = \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} p(j|i)\, \mathbf{x}^{(i)}$$

$$\widehat{\sigma}_j^2 = \frac{1}{\widehat{n}_j d} \sum_{i=1}^{n} p(j|i)(||\mathbf{x}^{(i)} - \mu^{(j)}||)^2$$

M-estimation

M-estimation

1. Find the loss function $\rho : E \times \mathcal{M} \to \mathbb{R}$ where $\mathcal{M}$ is the set of all possible values for the unknown μ, such that

$$Q(\mu) := \mathbb{E}\left[\rho(X_1, \mu)\right]$$

achieves its minimum at $\mu = \mu^*$.

2. Estimator is then $\widehat{\mu} = \operatorname*{argmin}_{\mu} \frac{1}{n} \sum_i^n \rho(X_i, \mu)$

- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2$, for all $x, \mu \in \mathbb{R}$: $\mu^* = \mathbb{E}[X]$.
- If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x, \mu) = \|x - \mu\|_2^2$, for all $x, \mu \in \mathbb{R}^d$: $\mu^* = \mathbb{E}[X] \in \mathbb{R}^d$.
- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = |x - \mu|$, for all $x, \mu \in \mathbb{R}$: μ^* is a **median** of $\mathbb{P}$.
- If $E = \mathcal{M} = \mathbb{R}$, $\alpha \in (0, 1)$ is fixed and $\rho(x, \mu) = C_\alpha(x - \mu)$, for all $x, \mu \in \mathbb{R}$: μ^* is a α-quantile of $\mathbb{P}$.

$$C_\alpha = \begin{cases} -(1-\alpha)x & \text{if } x < 0 \\ \alpha x & \text{if } x \geq 0. \end{cases}$$

Method of Moment Estimator

Moment Generating Function

$$M_X(t) = \mathbb{E}e^{[tX]}$$

$$\mathbb{E}[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0}$$

Population Moments Let $m_k(\theta) = \mathbb{E}_\theta\left[X_1^k\right]$, $1 \leq k \leq d$.

Empirical Moments Let $\hat{m}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$,

$1 \leq k \leq d$.

$$(\hat{m}_1, \ldots, \hat{m}_d) \xrightarrow[n\to\infty]{\mathbb{P}/a.s.} (m_1(\theta), \ldots, m_d(\theta))$$

Moments Estimator Let

$$M : \Theta \to \mathbb{R}^d$$

$$\theta \mapsto M(\theta) = (m_1(\theta), \ldots, m_d(\theta))$$

Assume M is one-to-one:

$$\theta = M^{-1}\left(m_1(\theta), \ldots, m_d(\theta)\right)$$

Moments estimator of θ:

$$\widehat{\theta}_n^{\text{MM}} = M^{-1}\left(\widehat{m}_1, \ldots, \widehat{m}_d\right)$$

Generalized Method of Moment

$$\sqrt{n}\left(\widehat{\theta}_n^{\text{MM}} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \Gamma(\theta)\right),$$

where $\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta} M(\theta)\right]^\top \Sigma(\theta) \left[\frac{\partial M^{-1}}{\partial \theta} M(\theta)\right]$

Confidence Interval

CI $: \mathcal{I} = [L(X_1, ..., X_n), U(X_1, ..., X_n))]$

CI of level $1-\alpha$

$$\mathbb{P}_\theta\left[\mathcal{I} \ni \theta\right] \geq 1-\alpha, \quad \forall \theta \in \Theta$$

CI of asymptotical level

$$\lim_{n\to\infty} \mathbb{P}_\theta\left[\mathcal{I} \ni \theta\right] \geq 1-\alpha, \quad \forall \theta \in \Theta.$$

Procedures to a CI

1. Start from a pivot statistic (non-asymptotic) or CLT (asymptotic)
2. Solve for $\mathbb{P}(\theta \in [\widehat{\theta} - s, \widehat{\theta} + s]) = 1-\alpha$
3. Two side symmetrical

$$\mathcal{I} = [\widehat{\theta} - \frac{\sigma q_{\alpha/2}}{\sqrt{n}}, \widehat{\theta} + \frac{\sigma q_{\alpha/2}}{\sqrt{n}}]$$

   (a) Conservative bound: use known bound on σ
   (b) Solve: solve equation
   (c) Plug-in: plug a consistent estimator of σ

Hypotheses Testing

$$\psi = \mathbb{1}\{|T_n| > q_{\alpha/2}\} = \mathbb{1}\{T_n > q_\alpha\} = \mathbb{1}\{T_n < -q_\alpha\}$$

Yes or No answer against 2 disjoint regions (both should be subsets of parameter space)

- **Rejection region** of a test ψ: $R_\psi = \{x \in E^n : \psi(x) = 1\}$.
- **Power Function**: $\beta(\theta) = \mathbb{P}_\theta[\psi = 1]$
- **Type I Error**: If $\theta \in \Theta_0$ (Given Null Reject Null; Reject wrongly)
$$\mathbb{P}_\theta[Type I of \psi] = \beta(\theta)$$
- **Type II Error**: If $\theta \in \Theta_0$ (Given Alter not Reject Null)
$$\mathbb{P}_\theta[Type II of \psi] = 1 - \beta(\theta)$$
- **Level** (upper bound on Type I error): A test ψ has level α if
  1. Non-Asymptotic: $\max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi = 1] <= \alpha$
  2. Asymptotic: $\lim_{n\to\infty} \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi = 1] <= \alpha$
- **Test from CI** Given a CI at level $1-\alpha$ I = [A, B], $\psi = 1[\theta_0 \notin [A, B]]$ is a test at level $1-\alpha$
- **p-value** The (asymptotic) p-value of a test ψ is the smallest (asymptotic) level α at which ψ rejects H_0
p-value = $\mathbb{P}(Z > T_n^{obs})$

Parametric Test

Wald's Test

If an estimator is both consistent and asymptotically normal. Then we can define test with following test statistic $W = \frac{\widehat{\theta} - \theta_0}{\sqrt{\widehat{var}(\widehat{\theta})}}$.

- require Slusky for replacing σ
- $\widehat{var}(\widehat{\theta})$ can be any consistent variance estimator of $\widehat{\theta}$
- For MLE estimator it equals $W = \sqrt{nI(\widehat{\theta}^{MLE})}(\widehat{\theta}^{MLE} - \theta_0)$
- 2-sample Wald-Test
$$\frac{(\widehat{\mu_1} - \widehat{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\widehat{\sigma_1}^2}{n_1} + \frac{\widehat{\sigma_2}^2}{n_2}}}$$

Likelihood Test

Consider an i.i.d. sample $X_1, \ldots, X_n$ with statistical model $\left(E, (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$, where $\Theta \subseteq \mathbb{R}^d$ ($d \geq 1$). Suppose the null hypothesis has the form

$$H_0 : (\theta_{r+1}, \ldots, \theta_d) = \left(\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)}\right),$$

for some fixed and given numbers $\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)}$.

Let

$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \ell_n(\theta) \qquad \text{(MLE)}$$

and

$$\widehat{\theta}_n^c = \operatorname*{argmax}_{\theta \in \Theta_0} \ell_n(\theta) \qquad \textit{(constrained MLE)}$$

where
$\Theta_0 = \left\{\theta \in \Theta : (\theta_{r+1}, \ldots, \theta_d) = \left(\theta_{r+1}^{(0)}, \ldots, \theta_d^{(0)}\right)\right\}$

Test statistic:

$$T_n = 2\left(\ell_n\left(\widehat{\theta}_n\right) - \ell_n\left(\hat{\theta}_n^c\right)\right).$$

Wilk's Theorem Assume H_0 is true and the MLE technical conditions are satisfied. Then,

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{d-r}$$

Likelihood ratio test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\left\{T_n > q_\alpha\right\},$$

where q_α is the $(1-\alpha)$-quantile of χ^2_{d-r}.

On sample T test Works for expected value of Gaussian X_i and small sample. In general, Wald test leads to smaller p-value

For a positive integer d, the Student's T distribution with d degrees of freedom (denoted by t_d) is the law of the random variable $\frac{Z}{\sqrt{V/d}}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$ and $Z \perp\!\!\!\perp V$.

$$T_n = \sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{\widetilde{S}_n}} = \frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{\widetilde{S}_n}{\sigma^2}}} \sim t_{n-1}$$

, where $\widetilde{S}_n$ is the unbiased estimator

Two sample T test

$$\frac{\overline{X}_n - \overline{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\widehat{\sigma}_d^2}{n} + \frac{\widehat{\sigma}_c^2}{m}}} \sim t_N$$

$$N = \frac{\left(\frac{\widehat{\sigma}_d^2}{n} + \frac{\widehat{\sigma}_c^2}{m}\right)^2}{\frac{\widehat{\sigma}_d^4}{n^2(n-1)} + \frac{\widehat{\sigma}_c^4}{m^2(m-1)}} \geq \min(n,m)$$

Multiple Test

- Family-wise error rate (FWER) = prob of making at least one false discovery (type I)
- False discovery rate (FDR) = expected fraction of false discovery among all significant results
- FDR <= FWER
- Bonferroni Correction to control FWER
$$p^i < \frac{\alpha}{m}$$
- BH to control FDR
  1. order p-value $P_1 < P_2 < ... < P_N$
  2. find max k such that $P_i <= \frac{k}{m}\alpha$
  3. reject all of $H_0^1, ..., H_0^k$

Nonparametric Testing

χ **Test** when H_0 hold

$$T_n = n\sum_{j=1}^K \frac{\left(\widehat{\mathbf{p}}_j - \mathbf{p}_j^0\right)^2}{\mathbf{p}_j^0} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-1}$$

χ **Test for Family of Distribution**

$$T_n = n\sum_{j=1}^K \frac{\left(\widehat{\mathbf{p}}_j - f_{\widehat{\theta}}(j)\right)^2}{f_{\widehat{\theta}}(j)} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-d-1}$$

d is the dim of parameter space

Empirical CDF

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left\{X_i \leq t\right\} = \frac{\#\{i = 1, \ldots, n : X_i \leq t\}}{n}, \quad \forall t \in \mathbb{R}.$$

Consistency $F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$.

Glivenko-Cantelli Theorem

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$$

Asymptotic Normality

$$\sqrt{n}\left(F_n(t) - F(t)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, F(t)\left(1 - F(t)\right)\right)$$

Donsker's Theorem

$$\sqrt{n}\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} \sup_{0 \leq t \leq 1} |\mathbf{B}(t)|,$$

where $\mathbf{B}(t)$ is a Brownian bridge on $[0,1]$.

Kolmogorov-Smirnov Test

Let $T_n = \sup_{t \in \mathbb{R}} \sqrt{n}\,|F_n(t) - F(t)|$. By Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n\to\infty]{(d)} Z$, where Z has a known distribution (supremum of the absolute value of a Brownian bridge).

KS test with asymptotic level α:

$$\delta_\alpha^{\text{KS}} = \mathbb{1}\left\{T_n > q_\alpha\right\}$$

where q_α is the $(1-\alpha)$-quantile of Z.

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the reordered sample. The expression for T_n reduces to

$$T_n = \sqrt{n}\max_{i=1,\ldots,n}\left\{\max\left(\left|\frac{i-1}{n} - F^0\left(X_{(i)}\right)\right|, \left|\frac{i}{n} - F^0\left(\right.\right.\right.\right.$$

KS table is for $\frac{T_n}{\sqrt{n}}$

Kolmogorov-Lilliefors Test

We want to test if X has a Gaussian distribution with unknown parameters. In this case, Donsker's theorem is *no longer valid*. Instead, we compute the quantiles for the test statistic

$$\sqrt{n}\sup_{t\in\mathbb{R}}\left|F_n(t)-\Phi_{\hat{\mu},\hat{\sigma}^2}(t)\right|$$

where $\hat{\mu}=\overline{X}_n$, $\hat{\sigma}^2=S_n^2$ and $\Phi_{\hat{\mu},\hat{\sigma}^2}(t)$ is the CDF of $\mathcal{N}\left(\hat{\mu},\hat{\sigma}^2\right)$.

They do not depend on unknown parameters.

should compare $\frac{F_n}{\sqrt{n}}$ with the table for both tests

Kolmogorov-Smirnov has a greater prob of rejection

Both Kolmogorov-Smirnov and Kolmogorov-Lilliefors test are non-asymptotic (statistics are pivot even for small n)

QQ plot

- Check if the points $\left(F^{-1}(\frac{1}{n}),F_n^{-1}(\frac{1}{n})\right),\ldots,\left(F^{-1}(\frac{n-1}{n}),F_n^{-1}(\frac{n-1}{n})\right)$ are near the line $y=x$.
- F_n is not technically invertible but we define $$F_n^{-1}(\tfrac{i}{n})=X_i,$$ the i^{th} largest observation.
- Right heavy tail (above). Left heavy tail (below)

Bayesian Stat

$$\pi\left(\theta|X_1,\ldots,X_n\right)\propto\pi(\theta)L_n(X_1,\ldots,X_n|\theta),\quad\forall\theta\in\Theta$$

Maximum a posteriori probability (MAP) The MAP estimate, $\hat{\theta}$, is the value at which the posterior distribution is maximum:

$$f_{\Theta|X}(\theta^*|x)=\max_{\theta}f_{\Theta|X}(\theta|x).$$

Least Mean Squares (LMS) The LMS estimate is the conditional expectation of the posterior distribution:

$$\hat{\theta}=\mathbb{E}\left[\Theta|X=x\right].$$

Linear Least Mean Squares LLMS In some cases, the conditional expectation $\mathbb{E}[\Theta|X]$ may be hard to compute or implement. In that case, we can restrict our attention to estimators of the form $\hat{\Theta}=aX+b$. Then,

$$\hat{\Theta}_{\text{LLMS}}=\mathbb{E}[\Theta]+\frac{\text{Cov}(\Theta,X)}{\text{Var}(X)}\left(X-\mathbb{E}[X]\right)$$
$$=\mathbb{E}[\Theta]+\rho\frac{\sigma_\Theta}{\sigma_X}\left(X-\mathbb{E}[X]\right)$$

Gaussian Distribution $\mu=-\dfrac{\beta}{2\alpha},\sigma^2=\dfrac{1}{2\alpha}$

$$f_X(x)=ce^{-(\alpha x^2+\beta x+\gamma)}$$

Bayes Rule

Discrete Θ, Discrete X

$$p_{\Theta|X}(\theta|x)=\frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$$
$$p_X(x)=\sum_{\theta'}p_\Theta(\theta')p_{X|\Theta}(x|\theta')$$

Discrete Θ, Continuous X

$$p_{\Theta|X}(\theta|x)=\frac{p_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$$
$$f_X(x)=\sum_{\theta'}p_\Theta(\theta')f_{X|\Theta}(x|\theta')$$

Continuous Θ, Continuous X

$$f_{\Theta|X}(\theta|x)=\frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$$
$$f_X(x)=\int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'$$

Continuous Θ, Discrete X

$$f_{\Theta|X}(\theta|x)=\frac{f_\Theta(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$$
$$p_X(x)=\int f_\Theta(\theta')p_{X|\Theta}(x|\theta')d\theta'$$

Jeffreys Prior Gives more weight to values of θ where

1. MLE estimate has less uncertainty
2. Data has more information towards deciding the parameter
3. Potential outcomes are more sensitive to slight changes in θ

$$\pi_J(\theta)\propto\sqrt{\det I(\theta)}$$

Reparametrization invariance principle: If η is a reparametrization of θ (i.e., $\eta=\phi(\theta)$ for some one-to-one map ϕ), then the PDF $\tilde{\pi}(\cdot)$ of η satisfies:

$$\tilde{\pi}(\eta)\propto\sqrt{\det\tilde{I}(\eta)},$$

where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

For $\theta=f(\theta_1)$, $I(\theta)d\theta=I(f(\theta_1))df(\theta_1)=I^{(1)}(\theta_1)d\theta_1$

Linear Regression

Regression Function Give Join Prob Distribution $\mathbb{P}$, the regression function of Y with respect to X is

$$v(x)=\mathbb{E}[Y|X=x]=\sum_{\Omega_Y}y\mathbb{P}(Y=y|X=x)$$

$$m(x),\int_{-\infty}^{m(x)}h(y|x)dy=\frac{1}{2}$$
$$m(x),\int_{-\infty}^{m(x)}h(y|x)dy=1-\alpha$$

Probabilistic Analysis

$$b^*=\frac{\text{Cov}\left(X,Y\right)}{\text{Var}\left(X\right)}$$
$$a^*=\mathbb{E}[Y]-b^*\mathbb{E}[X]=\mathbb{E}[Y]-\frac{\text{Cov}\left(X,Y\right)}{\text{Var}\left(X\right)}\mathbb{E}[X]$$

LSE

$$\hat{b}=\frac{\overline{XY}-\overline{X}\,\overline{Y}}{\overline{X^2}-\overline{X}^2}$$
$$\hat{a}=\overline{Y}-\hat{b}\overline{X}$$

Property of LSE

- LSE = MSE
- Distribution of $\widehat{\boldsymbol{\beta}}$: $$\widehat{\boldsymbol{\beta}}\sim\mathcal{N}_p\left(\boldsymbol{\beta}^*,\sigma^2\left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\right)$$
- Quadratic Risk of $\widehat{\boldsymbol{\beta}}$: $$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2\right]=\sigma^2\mathbf{tr}\left(\left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\right)$$
- Prediction Error: $$\mathbb{E}\left[\|\mathbf{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2\right]=\sigma^2\left(n-p\right)$$
- Unbiased estimator of σ^2: $$\widehat{\sigma}^2=\frac{\|\mathbf{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2}{n-p}=\frac{1}{n-p}\sum_{i=1}^n\widehat{\varepsilon}_i^2$$
- Fisher Info $I(\beta)=\dfrac{X^TX}{\sigma^2}$
- Heterosckedasticity $\widehat{\beta}=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$

T test (Non-asym)

$$\frac{\mu^T\widehat{\beta}-\mu^T\beta^0}{\widehat{\sigma}\sqrt{\mu^T(X^TX)^{-1}\mu}}\sim t_{n-p}$$

Generalized Linear Model

Generalization

1. Random component: $Y|X=x\sim$ some distribution
2. Regression function: $(\mu(x))=x^\intercal\beta$, g is the link function

Exponential Family A family of distribution with the following format

- $\eta_1,\ldots,\eta_k$ and $B(\theta)$
- $T_1,\ldots,T_k$, and $h(y)\in\mathbb{R}^q$

such that the density function of $\mathbb{P}_\theta$ can be written as

$$f_\theta(y)=\exp\left[\sum_{i=1}^k\eta_i(\theta)T_i(y)-B(\theta)\right]h(y)$$

One Param Canonical Exponential Family

$$f_\theta(y)=\exp\left(\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right)$$

for some known functions $b(\theta)$ and $c(y,\phi)$.

- **Expected value** $\mathbb{E}\left[Y\right]=b'(\theta)$.
- **Variance** $\text{Var}(Y)=b''(\theta)\cdot\phi$

GLM

- **Link function** Relate $X^T\beta$ to μ $$X^T\beta=g(\mu)=g(\mu(X))$$ $$\mu=g^{-1}(X^T\beta)$$
- **Canonical Link** Function that link μ to the canonical parameter θ $$g(\mu)=\theta=(b')^{-1}(\mu)$$
- **Full Model**

  Let $(X_i,Y_i)\in\mathbb{R}^p\times\mathbb{R}$, $i=1,\ldots,n$ be independent random pairs such that the conditional distribution of Y_i given $X_i=x_i$ has density in the canonical exponential family:

  $$f_{\theta_i}(y_i)=\exp\left[\frac{y_i\theta_i-b(\theta_i)}{\phi}+c(y_i,\phi)\right]$$

  Back to β: Given a link function g, note the following relationship between β and θ:

  $$\theta_i=(b')^{-1}(\mu_i)=(b')^{-1}\left(g^{-1}(X_i^\intercal\beta)\right)\equiv h\left(X_i^\intercal\beta\right)$$

  where h is defined as

  $$h=(b')^{-1}\circ g^{-1}=(g\circ b')^{-1}.$$

  If g is the *canonical link function*, h is the **identity** $g=(b')^{-1}$.

  Log-likelihood The log-likelihood is given by

  $$\ell_n\left(\mathbf{Y},\mathbb{X},\beta\right)=\sum_i\frac{Y_i\theta_i-b(\theta_i)}{\phi}+\text{constant}$$
  $$=\sum_i\frac{Y_ih\left(X_i^\intercal\beta\right)-b\left(h\left(X_i^\intercal\beta\right)\right)}{\phi}+\text{cons}$$

  When we use the *canonical link function*, we obtain the expression

  $$\ell_n\left(\mathbf{Y},\mathbb{X},\beta\right)=\sum_i\frac{Y_iX_i^\intercal\beta-b\left(X_i^\intercal\beta\right)}{\phi}+\text{constant}$$

Counting

Selection For a selection that can be done in r stages, wight n_i choices at each stage i, the number of possible selection is: $n_1 n_2 ... n-_r$

mutation # of ways of ordering n distinct elements: $n! = 1 * 2 * 3 ... n$

binations Give a set of n elements, number of ways of constructing an **ordered** sequence of k **distinct** element (result order does not matter): $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Subsets for subset of $\{1, ..., n\}$: $\sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \cdots + \binom{n}{n} = 2^n$.

Partitions $\frac{n!}{n_1! n_2! ... n_r!}$.

Bernoulli Process

Def. A sequence of Bournoulli trials X_i (independence + time-homogeneity)

First Arrival Time of first arrival

- $T_1 = min\{i : X_i = 1\}$
- $\mathbb{P}(T_1 = k) = (1-p)^{k-1} p$
- $\mathbb{E}(T_1) = \frac{1}{p}$
- $var(T_1) = \frac{1-p}{p^2}$

Memoryless Fresh start after a random time N. $X_{N+1}, X_{N+2}, ...$ is a Bernoulli Process independent of N, $X_1, ..., X_N$

K-th Arrival

- i-th inter-arrival time $T_i = Y_i - Y_{i-1} \sim Gep(p)$ and independent with T_j
- $\mathbb{E}[Y_k] = \frac{k}{p}, var(Y_k) = \frac{k(1-p)}{p^2}$
- $\mathbb{P}_{Y_k} = \binom{t-1}{k-1} p^k (1-p)^{t-k}$

Possion appro Total number of arrivals converge to Poisson distribution for large n, small p and moderate $\lambda = np$

Poisson Process

Prob of k arrivals in duration δ (λ is arrival rate, $\tau = n\delta$, N_τ is binomial and converge to Poisson) Then $\mathbb{P}(k, \delta) =$

$$\begin{cases} 1 - \lambda\delta + O(\delta^2) & \text{if k=0} \\ \lambda\delta + O(\delta^2) & \text{if k=1} \\ 0 + O(\delta^2) & \text{if k>1} \end{cases}$$

$$\mathbb{P}(k, \tau) = \mathbb{P}(N_\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

Time until first arrival $T_1 \sim Exp(\lambda)$

Time Y_k of the kth arrival

- Sum of independent Exp $Y_k = T_1 + T_2 + ... + T_k$
- $\mathbb{P}(Y_k \leq y) = \sum_{n=k}^{\infty} \mathbb{P}(n, y)$
- $Y_k \sim Erlang(k), f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$
- $\mathbb{E}[Y_k] = \frac{k}{\lambda}, var(Y_k) = \frac{k}{\lambda^2}$

Memoryless Starting from a constant time t or a certain arrival T_k (k is a constant), the following is a poisson process independent with history. Can divide time line and conque separately

Merge

- Sum of two Poisson process is a poisson process of parameter $\mu + v$
- $\mathbb{P}(kth - 1st) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$
- k of N arrivals are first = $\binom{N}{k} (\frac{\lambda_1}{\lambda_1 + \lambda_2})^k (\frac{\lambda_2}{\lambda_1 + \lambda_2})^{N-k}$
- Assume X, Y, Z are time until first arrival of 3 poisson process $min(X, Y, Z) \sim Poisson(\lambda_1 + \lambda_2 + \lambda_3)$

Split A poisson process can be splited in to Poisson(λq) and Poisson($\lambda(1-q)$). And the resulting 2 proceses are independent

Random Incidence Arrival at constant $t^\star$ U, V are time of last and next arrival.

$$(V - t^\star), (t^\star - U) \sim Exp(\lambda)$$

Markov Process

Given curent state, past does not matter

N step transition prob

- $r_{ij}(n) = \mathbb{P}(X_{n+s} = j | X_s = i)$
- recursion: $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1) p_{kj}$
- random initial state: $\mathbb{P}(X_n = j) = \sum_{i=1}^{m} \mathbb{P}(X_0 = i) r_{ij}(n)$
- convergence to π_j (not depend on n and i; only one recurrent class and it is not periodic)

Recurrent

- States: starting from i, and from wherever you can go, there is a way of returning to i
- Class: a collection of recurrent states communicating only between each other
- Periodic states in recurrent class:can be grouped in to d>1 groups so that all transitions from one group lead to the next group

Steady-stae Prob

- Converge to π_j if recurrent states are all in single class and is not periodic
- $\pi_j = \sum_k \pi_k p_{kj}, \sum_{j=1}^{m} \pi_j = 1$
- can be interpreted as: long run frequency in j, frequency of transition intoj

birth-death process

- $\pi_i p_i = \pi_{i+1} q_{i+1}$
- $\pi_0 + \pi_0 \frac{p_0}{q_1} + \pi_0 \frac{p_0 p_1}{q_1 q_2} + ... = 1$
- For fixed p <q: $\mathbb{E}(X_n)[\frac{\rho}{1-\rho}]$
- for p=q, all π equal

Absorption state

- a_i is the prob that eventually settle in absorb state a starting from i $a_i = \sum_{j=1}^{m} p_{ij} a_j$
- μ_i is the expected number of transitions reaching absorb state a starting from i $\mu_i = 1 + \sum_j p_{ij} \mu_j$

First passage and recurrence times

- Mean first passage time from i to s: $t_i = \mathbb{E}[min\{n \geq 0 such X_n = s\} | X_0 = i]$
- $t_s = 0, t_i = 1 + \sum_j p_{ij} t_j$ for all $i <> s$
- $t_s^\star = \mathbb{E}[min\{n \geq 1 such X_n = s\} | X_0 = s]$
- $t_s^\star = 1 + \sum_j p_{sj} t_j$

Derived Distribution

Given distribution of x what is the distribution of $y = g(x)$

Discrete case

- $\mathbb{P}_Y(y) = \mathbb{P}(g(x) = y) = \sum_{x : g(x) = y} \mathbb{P}_X(x)$
- When $g(x) = ax + b$

$$\mathbb{P}_Y(y) = \mathbb{P}_X(\frac{y-b}{a})$$

Continuous case

- **CDF**: $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(x) \leq y)$
- **Take derivative** $f_Y(y) = \frac{dF_Y}{dy}(y)$
- when $g(X) = aX + b$, then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$
- when g is mononic.

$$f_Y(y) = f_X(g^{-1}(y)) |\frac{g^{-1}(y)}{dy}|$$

Sum of RV

- Z=X+Y, $P_Z(z) = \sum_x P_X(x) P_Y(z-x)$
- $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$
- $var(X_1 + ... + X_n) = \sum_i^n var(X_i) + \sum_{\{(i,j):i<>j\}} cov(X_i, X_j)$
- for $Y = X_1 + ... + X_N$ where N is also a RV

$$E[Y] = E[N] E[X]$$

$$Var[Y] = E[N] var[X] + (E[X])^2 var(N)$$