Intermediate Statistics

By Junru Shao <junrushao1994@gmail.com>

https://github.com/junrushao1994/Cheat-Sheets

Misc

- Moment generating function: $M_x(t) = \mathbb{E}[\exp(tx)]$
- σ-sub-Gaussian: $M_{x-\mu} \le \exp(\sigma^2 t^2/2)$, 0-mean $[a, b]$ bounded RVs are $\frac{(b-a)}{2}$-sub-Gaussian
- Jensen's inequality: for g convex, $g(\mathbb{E}[X]) \le \mathbb{E}[g(X)]$
- Cauchy inequality: $(\mathbb{E}[XY])^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2]$
- maximin $\le$ minimax
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, $\mathrm{Var}\left(\sum_{i=1}^n \alpha_i X_i\right) = \sum_{i=1}^n\sum_{i=1}^n \alpha_i\alpha_j \mathrm{Cov}(X_i, X_j)$
- $\mathrm{Var}(XY) = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2\mathbb{E}[Y]^2$
- Gamma function $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)\,dx$. $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. $\Gamma(x+1) = x\Gamma(x)$
- Law of total expectation: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$, total variance: $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[\mathbb{E}(Y \mid X)]$
- Conditional gaussian: $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}$ ($\boldsymbol{\mu}_1 \in \mathbb{R}^q$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{N-q}$), $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$. Then
  - $\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_2)$, $\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$
- Posterior for μ in Gaussian, with known σ, and prior $\mu \sim \mathcal{N}(m, \tau^2)$
  - $\mathbb{E}[\mu \mid X^n] = \frac{\tau^2}{\tau^2+\sigma^2/n}\overline{X}_n + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}m$
  - $\mathrm{Var}(\mu \mid X^n) = \tau^2 \cdot \frac{\sigma^2}{n} / \left(\tau^2 + \frac{\sigma^2}{n}\right)$
- Berry-Essen Theorem: let $F_n(x) = P\left(\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu) \le x\right)$, then $\sup|F_n(x) - \Phi(x)| \le \frac{33}{4}\mathbb{E}|X_1 - \mu|^3/\sigma^3\sqrt{n}$
- KL divergence: $\mathrm{KL}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$
- Total variance distance: $\delta(P, Q) = \frac{1}{2}\int|P(x) - Q(x)|\,dx$

Inequalities

- Markov: for $X \ge 0$, $P[X \ge t] \le \mathbb{E}[X]/t$
- Chebyshev: $P[|X - \mathbb{E}X| \ge k\sigma] \le 1/k^2$
- Chernoff bound: $P[X - \mu \ge u] \le \inf_{0\le t\le b}\frac{M_{x-\mu}(t)}{\exp(tu)}$, where $M_{x-\mu}$ is finite $\forall|t| \le b$
  - (sub-)Gaussian tail bound: $P(|X - \mu| \ge k\sigma) \le 2\exp(-2k^2)$
- Hoeffding bound (bounded RV): $P(|x - \mu| \ge k(b-a)) \le 2\exp(-2k^2)$
  - for n RVs, $P(|x - \mu| \ge t) \le 2\exp\left(-2n^2t^2/\sum_{i=1}^n(b_i - a_i)^2\right)$
- Bernstein ($[a, b]$ bounded RV, $\mu = 0$, small σ, i.i.d.)
  - $P(|\mu - \hat{\mu}| \ge t) \le 2\exp\left(-\frac{nt^2}{2(\sigma^2 + (b-a)t)}\right)$
  - $P\left(|\mu - \hat{\mu}| \le 4\sigma\sqrt{\frac{\ln(2/\delta)}{n}} + \frac{4(b-a)\ln(2/\delta)}{n}\right) \ge 1 - \delta$
- Azuma's bound (difference bounded: $|f(x_1, \dots, x_k, \dots, x_n) - f(x_1, \dots, x_k', \dots, x_n)| \le L_k$)
  - $P[|f(x_1, \dots, f_n) - \mu| \ge t] \le 2\exp\left(-2t^2/\sum_{k=1}^n L_k^2\right)$
- U-statistics: $U(x_1, \dots, x_n) = \frac{1}{\binom{n}{2}}\sum_{j<k} g(x_j, x_k)$ and g is symmetric and $|g| \le B$
  - Azuma gives: $P[|U(x_1, \dots, x_n) - \mu| \ge t] \le 2\exp(-nt^2/8B^2)$
- χ^2 tail bound: $Y \sim X_n^2$ where $X_n \sim \mathcal{N}(0, 1)$. $P\left[\left|\frac{1}{n}Y - 1\right| \ge t\right] \le 2\exp(-nt^2/8)$
- Johnson-Lindenstrauss Lemma: for $X^n \in \mathbb{R}^d$, $m \ge 16\frac{\log(n/\delta)}{\varepsilon^2}$, $Z \in \mathbb{R}^{m\times d}$, $Z_{i,j} \sim \mathcal{N}(0, 1)$, $F(X_i) = \frac{Z}{\sqrt{n}}X_i$
  - $(1-\varepsilon)\|X_i - X_j\|_2^2 \le \|F(X_i) - F(X_j)\|_2^2 \le (1+\varepsilon)\|X_i - X_j\|_2^2$, holds w.p. $1 - \delta$

Convergence

Convergence: estimator $\hat{\theta}_n$ is consistent iff $\hat{\theta}_n \xrightarrow{p} \theta$

- a.s. $P\left(\lim_{n\to\infty} X_n = X\right) = 1$; i.p. $\lim_{n\to\infty} P(|X_n - X| \ge \varepsilon) = 0$; q.m. $\lim_{n\to\infty}\mathbb{E}(X_n - X)^2 = 0$; d $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$
  - a.s. $\Rightarrow$ p; q.m. $\Rightarrow$ p; q.m. $\Leftarrow$ p if $|X_n|$ bounded; p $\Rightarrow$ d; p $\Leftarrow$ d if $X = c$; p $\not\Rightarrow$ $\mathbb{E}[X_n] = c$
- Continuous mapping: if $X_n \xrightarrow{p} X$, then for any continuous function f, $f(X_n) \xrightarrow{p} f(X)$; if $X_n \xrightarrow{d} X$, then for any continuous function f, $f(X_n) \xrightarrow{d} f(X)$
- Slutsky's theorem: if $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$, and $X_nY_n \xrightarrow{p} XY$; if $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$, and $X_nY_n \xrightarrow{d} cX$
- WLLN: $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mu$ for i.i.d. X_is with $\mathbb{E}(|X|) < \infty$ and $\mathrm{Var}(X) < \infty$
- CLT: X^n independent RVs, mean and variance finite, then $\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, 1)$, works also with estimated variance
  - Lyapunov CLT: X^n are independent but not i.i.d. Define $s_n = \frac{1}{n}\sqrt{\sum_{i=1}^n \sigma_n^2}$. If $\lim_{n\to\infty}\frac{1}{s_n^3}\sum_{i=1}^n \mathbb{E}|X_i - \mu_i|^3 < \infty$, we have $\frac{1}{s_n}\sum_{i=1}^n (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1)$
  - multivariate CLT: $\sqrt{n}(\mu - \hat{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma)$
- Delta method (CLT for function of RV): if $\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu) \to \mathcal{N}(0, 1)$, g continuous differentiable, $g'(\mu) \ne 0$
  - $\frac{\sqrt{n}}{\sigma}(g(\hat{\mu}) - g(\mu)) \to \mathcal{N}\left(0, [g'(\mu)]^2\right)$
  - multivariate delta method: $\sqrt{n}(g(\hat{\mu}) - g(\mu)) \to \mathcal{N}\left(0, \nabla_g^T(\mu)\Sigma\nabla_g(\mu)\right)$

Empirical process

- Vapnik-Chervonenkis Theory
  - empirical CDF: $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \le x)$
    - Glivenko-Cantelli Theorem: $\Delta = \sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F_X(x)\right| \xrightarrow{p} 0$
  - empirical probability of set: $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \in A)$, and $\Delta(A) = \sup_{A\in\mathcal{A}}|P_n(A) - P(A)|$
  - empirical process: $\Delta(\mathcal{F}) = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f]\right|$
  - Glivenko-Cantelli class: $\Delta(\mathcal{F}) \xrightarrow{p} 0$ where X^n i.i.d.
- Shattering: n-th shattering coef of $\mathcal{A}$: $s(\mathcal{A}, n) = \max_{z^n} N_{\mathcal{A}}(z^n)$, where $N_{\mathcal{A}} = \#\{z^n \cap A\}$ (valid coloring)
- VC dimension: max d that $s(\mathcal{A}, d) = 2^d$. There exists $\{z_i\}_{i=1}^d$ each of whose colorings are valid, but for any $(d+1)$ points, there exists an invalid coloring
  - you select the best $x_1, \dots, x_n$, adversary assigns label $y_1, \dots, y_n$
  - if $\mathrm{VC}_{\mathcal{A}} \ge n$, you can find $f \in \mathcal{A}$ that is consistent with the labels
- VC dimension example:
  - $\mathcal{A} = \{A_1, \dots, A_N\}$, $V_{\mathcal{A}} \le \log_2 N$
  - intervals $[a, b]$ on the real line: 2
  - discs in $\mathbb{R}^2$: 3
  - closed balls in $\mathbb{R}^d$: $V_{\mathcal{A}} \le d + 2$
  - rectangles in $\mathbb{R}^d$: $2d$
  - half-space in $\mathbb{R}^d$: $d + 1$
  - convex polygons in $\mathbb{R}^2$: ∞
  - convex polygon with d vertices: $2d + 1$
- Empirical risk minimization: $\hat{f} = \arg\min_{f\in\mathcal{F}}\hat{R}_n(f)$, optimal $f^* = \arg\min_{f\in\mathcal{F}} R(f)$
  - minimize $\Delta(\mathcal{F}) = R\left(\hat{f}\right) - R(f^*)$, consider minimize $\Delta(\mathcal{A}) = \sup_{A\in\mathcal{A}}|P_n(A) - P(A)|$
  - if $|A|$ is finite: $P(|P_n(A) - P(A)| \ge t) \le 2\exp(-2nt^2)$, $P(\Delta(\mathcal{A}) \ge t) \le 2|A|\exp(-2nt^2)$
  - if $|A|$ is not finite: $P(\Delta(\mathcal{A}) \ge t) \le 8s(\mathcal{A}, n)\exp(-nt^2/32)$
- Sauer's Lemma: if $\mathrm{VC}(\mathcal{A}) = d < \infty$, then for $n > d$, $s(\mathcal{A}, n) \le (n+1)^d$
  - another form: $S_{\mathcal{F}}(n) \le \sum_{k=1}^{\mathrm{VC}_{\mathcal{F}}}\binom{n}{k}$
- Rademacher complexity: $\mathcal{R}(\mathcal{F}) = \mathbb{E}_\epsilon\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right]$
  - $\Delta(\mathcal{F}) = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f]\right|$
  - Rademacher Theorem: $\mathbb{E}[\Delta(\mathcal{F})] \le 2\mathcal{R}(\mathcal{F})$
  - finite class bound: if $|\mathcal{F}| = N$, $\forall f_i \in \mathcal{F}$ we have $\|f_i\|_\infty \le b$, then $\mathcal{R}(\mathcal{F}) \le 2b\sqrt{\frac{\log(2N)}{n}}$

Sufficiency

- Sufficient statistics: $T(X_1, \dots, X_n)$ is sufficient for θ if $p(X^n \mid T = t; \theta)$ does not depend on θ
  - factorization: $T(X_1, \dots, X_n)$ is sufficient for θ iff $p(X_1, \dots, X_n; \theta) = h(X_1, \dots, X_n)\,g(T; \theta)$
- Minimal sufficient statistics $\Rightarrow$ for RVs $\{x_n\}$ and $\{y_n\}$, $T(X^n) = T(Y^n) \Leftrightarrow \frac{p(Y^n; \theta)}{p(X^n; \theta)}$ does not depend on θ
- Rao-Blackwell Theorem: $\hat{\theta}$ an estimator, T SS, $\widetilde{\theta} = \mathbb{E}\left[\hat{\theta} \mid T\right]$, then $R\left(\widetilde{\theta}, \theta\right) \le R\left(\hat{\theta}, \theta\right)$ under squared loss

Parameter Estimation

- Method of moments (MOM) : solve for θ such that $\mathbb{E}[X^j] = \frac{1}{n}\sum_{i=1}^n X_i^j$
- MLE: $\hat{\theta}$ maximize $\mathcal{L}(X^n; \theta)$
  - equivariance: if we replace θ by $\eta = g(\theta)$, the MLE estimators satisfy $\hat{\eta} = g\left(\hat{\theta}\right)$ and $\mathcal{L}(\hat{\eta}) = \mathcal{L}\left(\hat{\theta}\right)$
    - profile likelihood: if g not invertible, let $\mathcal{L}^*(\eta) = \sup_{\theta:\, g(\theta)=\eta}\mathcal{L}(\theta)$, then $\hat{\eta} = g(\hat{\theta})$ and $\mathcal{L}^*(\eta) = \mathcal{L}\left(\hat{\theta}\right)$
  - MLE, ERM and KL: $R_n\left(\theta, \hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i; \theta)}{p(X_i; \hat{\theta})}$ = constant − MLE
    - population risk is $R\left(\theta, \hat{\theta}\right) = \mathbb{E}_{X;\theta}\log\frac{p(X; \theta)}{p(X; \hat{\theta})} = \mathrm{KL}\left(p(X; \theta)\,\|\,p\left(X; \hat{\theta}\right)\right)$
  - conditions for MLE to be consistent
    - strong identifiability: $\inf_{\hat{\theta}:\,|\widetilde{\theta}-\theta|\ge\varepsilon}\mathrm{KL}\left(p(X; \theta)\,\|\,p\left(X; \hat{\theta}\right)\right) > 0$
    - uniform LLN: $\sup_{\widetilde{\theta}}\left|R_n\left(\widetilde{\theta}, \theta\right) - R\left(\widetilde{\theta}, \theta\right)\right| \xrightarrow{p} 0$

- MLE asymptotics: $\sqrt{n}\left(\hat{\theta}-\theta\right)\xrightarrow{d}\mathcal{N}\left(0,\,I_1^{-1}\left(\theta\right)\right)$, or $\sqrt{n}\left(\hat{\tau}-\tau\right)\rightsquigarrow\mathcal{N}\left(0,\,\left(g'\right)^T I_1^{-1}g'\right)$ where $\tau=g\left(\theta\right)$, conditions:
  - 1) θ is identifiable. 2) $p\left(X;\,\theta\right)$ is thrice differentiable function of θ; 3) The range of X does not depend on $\theta \Leftarrow$ interchange diff w.r.t θ and int over X. 4) θ is in the interior of Θ. 5) dimension space does not change with n
- Influence functions: $\psi\left(x\right)=\frac{\nabla_\theta\log p\left(x;\theta\right)}{I\left(\theta\right)}$, $\hat{\theta}=\theta+\frac{1}{n}\sum_{i=1}^n\psi\left(X_i\right)+\text{Remainder}$. Robust if ψ is bounded
  - asymptotically linear estimators: satisfy $\hat{\theta}\approx\theta+\frac{1}{n}\sum_{i=1}^n\psi\left(X_i\right)$
  - any sufficiently well-behaved (regular) estimator is asymptotically linear
- Asymptotic Relative Efficiency (ARE): two estimators W_n and V_n estimating $\tau\left(\theta\right)$ where $\sqrt{n}\left(W_n-\tau\left(\theta\right)\right)\rightsquigarrow\mathcal{N}\left(0,\,\sigma_w^2\right)$, $\sqrt{n}\left(V_n-\tau\left(\theta\right)\right)\rightsquigarrow\mathcal{N}\left(0,\,\sigma_v^2\right)$. Then $\text{ARE}\left(V_n,\,W_n\right)=\sigma_w^2/\sigma_v^2$. In general, MLE estimator $\hat{\theta}$ satisfies $\text{ARE}\left(\widetilde{\theta},\,\hat{\theta}\right)\leq 1$ for any other estimator $\widetilde{\theta}$.

Decision Theory

- Fisher information
  - score function: $s\left(\theta\right)=\sum_{i=1}^n\nabla_\theta\log p\left(X_i;\,\theta\right)$. data dependent, but $\mathbb{E}_{X^n;\,\theta}\left[s\left(\theta\right)\right]=0$
  - fisher information: $I\left(\theta\right)=\mathbb{E}\left[s\left(\theta\right)s^T\left(\theta\right)\right]=\text{cov}\left(s\left(\theta\right)\right)$. data independent
  - single sample: $I_1\left(\theta\right)=\mathbb{E}\left[-\nabla_\theta^2\log p\left(X;\,\theta\right)\right]$; n samples: $I\left(\theta\right)=nI_1\left(\theta\right)$
  - Cramér-Rao bound: $\text{Var}\left(\hat{\theta}\right)\geq 1/nI_1\left(\theta\right)$. Multivariate: $\text{Var}\left(\hat{\theta}\right)\succeq\frac{1}{n}I_1^{-1}\left(\theta\right)$
  - efficient estimator: unbiased & achieve Cramér-Rao bound
- Risk w.r.t loss $L\left(\theta,\,\hat{\theta}\right)$ is $R\left(\theta,\,\hat{\theta}\right)=\mathbb{E}_{X^n;\,\theta}\left[L\left(\theta,\,\hat{\theta}\right)\right]$
  - MSE: L is squared error. $\text{MSE}=\mathbb{E}_\theta\left[\left(\hat{\theta}-\theta\right)^2\right]=\left(\mathbb{E}_\theta\left[\hat{\theta}\right]-\theta\right)^2+\text{Var}_\theta\left(\hat{\theta}\right)$
  - minimax risk: $R_n=\inf_{\hat{\theta}}\sup_{\theta\in\Theta}R\left(\theta,\,\hat{\theta}\right)=\inf_{\hat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}_{X^n;\,\theta}\left[L\left(\theta,\,\hat{\theta}\right)\right]$
- Minimax estimator: $\hat{\theta}$ that minimize the worst case $R\left(\hat{\theta},\,\cdot\right)$: $\sup_\theta R(\theta,\,\hat{\theta})=\inf_{\hat{\theta}}\sup_\theta R(\hat{\theta},\,\theta)$
  - maximum risk: $\overline{R}\left(\hat{\theta}\right)=\sup_{\theta\in\Theta}R\left(\theta,\,\hat{\theta}\right)=\sup_{\theta\in\Theta}\mathbb{E}_{X^n;\,\theta}\left[L\left(\theta,\,\hat{\theta}\right)\right]$
  - if Bayes estimator $\hat{\theta}$ for prior distribution π satisfies $\overline{R}\left(\hat{\theta}\right)\leq B_\pi\left(\hat{\theta}\right)$, then $\hat{\theta}$ is minimax. π is least favorable prior
    - Corollary: if risk $R\left(\theta,\,\hat{\theta}\right)$ is constant for a Bayes estimator $\hat{\theta}$, then $\hat{\theta}$ is also minimax
  - if $L\left(\theta,\,\hat{\theta}\right)=l\left(\theta-\hat{\theta}\right)$ where l is convex and bounded, symmetric about the origin, $X\sim\mathcal{N}\left(\theta,\,\Sigma\right)$, then X is the unique minimax estimator of θ
  - MLE estimator is approximately minimax as $n\to\infty$
- Bayes estimator (natural estimator): $\hat{\theta}$ that minimizes Bayes risk
  - Bayes risk: $B_\pi\left(\hat{\theta}\right)=\mathbb{E}_\pi\left[R\left(\theta,\,\hat{\theta}\right)\right]=\mathbb{E}_\pi\mathbb{E}_{X^n;\,\theta}\left[L\left(\theta,\,\hat{\theta}\right)\right]\leq\overline{R}\left(\hat{\theta}\right)$
  - marginal: $m\left(X^n\right)=\int p\left(X^n\,|\,\theta\right)\pi\left(\theta\right)\,d\theta=\frac{p\left(\theta,\,X^n\right)}{\pi\left(\theta\,|\,X^n\right)}$ (normalizer in Bayes)
  - Bayes estimator $\hat{\theta}_B$ minimizes posterior risk: $r\left(\hat{\theta}\,|\,X^n\right)=\mathbb{E}_{\theta\,|\,X^n}\left[L\left(\theta,\,\hat{\theta}\right)\right]$
    - squared loss: posterior mean $\mathbb{E}\left[\theta\,|\,X^n=x^n\right]$; absolute loss: median; 0/1 loss: mode
    - under squared loss, $r\left(\hat{\theta}\,|\,X^n\right)=\text{Var}\left(\theta\,|\,X^n\right)$
  - Bayes risk equivalence: $B_\pi\left(\hat{\theta}\right)=\int r\left(\hat{\theta}\,|\,X^n\right)m\left(X^n\right)\,dX^n$

Exponential family: $p\left(X;\,\theta\right)=\exp\left[\sum_{i=1}^s\eta_i\left(\theta\right)T_i\left(x\right)-A\left(\theta\right)\right]h\left(x\right)$

- Canonical parametrization: $\eta_i\left(\theta\right)=\theta_i$. $A\left(\theta\right)=\log\left[\int_X\exp\left[\sum_{i=1}^s\theta_iT_i\left(x\right)\right]h\left(x\right)\,dx\right]$
- Log partition: $\frac{\partial A(\theta)}{\partial\theta_i}=\mathbb{E}\left[T_i\left(X\right)\right]$, $\frac{\partial^2A(\theta)}{\partial\theta_i\partial\theta_j}=\text{cov}\left(T_i\left(X\right),\,T_j\left(X\right)\right)$. Therefore, A is convex
- Sufficient statistics: $T\left(X^n\right)=\left(\sum_{i=1}^nT_1\left(X_i\right),\,\ldots,\,\sum_{i=1}^nT_s\left(X_i\right)\right)$
- Concave log-likelihood: $\mathcal{LL}\left(\theta;\,X^n\right)\propto\sum_{i=1}^s\theta_i\sum_{j=1}^nT_i\left(x_j\right)-nA\left(\theta\right)$, $\mathcal{LL}\propto\langle\theta,\,T\rangle-nA\left(\theta\right)$
- Minimal representation: the T_is are linearly independent
  - over-complete exponential families are not statistically identifiable

Hypothesis testing: Null hypothesis: $H_0:\,\theta\in\Theta_0$. Alternative hypothesis: $H_1:\,\theta\in\Theta_1$

- TN = retain True; FP = reject True (type 1); FN = retain False (type 2); TP = reject False
  - type 1: the incorrect rejection of a true H_0
  - type 2: the failure to reject a false H_0
- Statistical significance (from Wikipedia)
  - definition: unlikely to have occurred under H_0, when $p<\alpha$
  - significance level α: $\Pr\left[\text{reject }H_0\,|\,H_0\text{ is true}\right]=\Pr\left[\text{type 1}\right]$
- Power (prob of rejection θ): should be large for $\theta\in\Theta_1$, small for $\theta\in\Theta_0$
  - probability of rejection when true parameter is θ (Type 1 error): $\beta\left(\theta\right)=P_\theta\left(X^n\in R\right)$
  - always favor the null hypothesis and only consider tests that control the Type-I error.
  - size $s=\sup_{\theta\in\Theta_0}\beta\left(\theta\right)$. level α: $s\leq\alpha$
  - power of a test: $\Pr\left(\text{reject }H_0\,\middle|\,H_1\text{ is true}\right)=1-\Pr\left[\text{type 2}\right]$
- p-value: defined for data x, $H_0:\,\theta\in\Theta_0$, smallest α at which we would reject H_0
  - definition: $p=\inf\left\{\alpha:\,T\left(x\right)\in R_\alpha\right\}=\sup_{\theta\in\Theta_0}P_\theta\left(T\left(X\right)\geq T\left(x\right)\right)$
  - $p\sim\text{Unif}\left(0,\,1\right)$ under Θ_0
  - a small p-value indicates strong evidence against H_0
- Neyman-Pearson Test ($H_0:\,\theta=\theta_0$ vs $H_1:\,\theta=\theta_1$)
  - $T=\frac{\mathcal{L}\left(\theta_1\right)}{\mathcal{L}\left(\theta_0\right)}$, $R_\alpha=\left\{T>k_\alpha\right\}$; set $k\in\left\{t:\,P_{\theta_0}\left(T>t\right)\leq\alpha\right\}$ to obtain tests of level α.
- Wald Test ($H_0:\,\theta=\theta_0$ vs $H_1:\,\theta\neq\theta_0$): MLE $\hat{\theta}$, $T=\frac{\hat{\theta}-\theta}{\hat{\sigma_0}}$, reject $\left|T_n\right|\geq z_{\alpha/2}$
  - Assumption: asymptotically normal estimator $\hat{\theta}$ that satisfies $\hat{\theta}\xrightarrow{d}\mathcal{N}\left(\theta_0,\,\sigma_0^2\right)$ under H_0. Can replace σ_0 with its plug-in estimation $\text{se}\left(\hat{\theta}\right)$
  - For MLE $\hat{\theta}$: $T=\sqrt{nI_1\left(\theta_0\right)}\left(\hat{\theta}-\theta_0\right)$, $R_\alpha=\left\{\left|T\right|\geq z_{\alpha/2}\right\}$
  - Power: let $\Delta=\sqrt{nI_1\left(\theta_0\right)}\left(\theta-\theta_0\right)$, $\beta\left(\theta\right)=1-\Phi\left(\Delta+z_{\alpha/2}\right)+\Phi\left(\Delta-z_{\alpha/2}\right)$. Large if $\left|\theta-\theta_0\right|$ or n large
- Likelihood Ratio Test ($H_0:\,\theta\in\Theta_0$, $H_1:\,\theta\notin\Theta_0$), p-value $=P\left(\chi^2_{\#\text{param}}>\lambda\right)$
  - $\lambda\left(X^n\right)=2\log\frac{\sup_{\theta\in\Theta}\mathcal{L}\left(\theta\right)}{\sup_{\theta\in\Theta_0}\mathcal{L}\left(\theta\right)}\sim\chi_d^2$, $R_\alpha=\left\{\chi^2_{d,\,\alpha}>\lambda\right\}$, where $d=\dim\left(\Theta\right)-\dim\left(\Theta_0\right)$
- Goodness-of-fit test:
  - χ^2 test for multinomial distributions
    - one sample testing: $H_0:\,\mathbf{p}=\mathbf{p}_0=\left(p_{01},\,\ldots,\,p_{0k}\right)$ vs $H_1:\,\mathbf{p}\neq\mathbf{p}_0$. Z^k is bin counts
      - $T=\sum_{i=1}^k\frac{\left(Z_i-np_{0i}\right)^2}{np_{0i}}\sim\chi^2_{k-1}$, $R_\alpha=\left\{\chi^2_{k-1,\alpha}>T\right\}$, p-value $=\chi^2_{k-1,\,T}$
    - two sample testing: $H_0:\mathbf{p}_x=\mathbf{p}_y$ vs $H_1:\,\mathbf{p}_x\neq\mathbf{p}_y$. Z_x, Z_y are bin counts for x, y
      - Let $\hat{p}_i=\frac{Z_{xi}+Z_{yi}}{n_x+n_y}$, then $T=\sum_{i=1}^k\frac{\left(Z_{xi}-n_x\hat{p}_i\right)^2}{n_x\hat{p}_i}+\frac{\left(Z_{yi}-n_y\hat{p}_i\right)^2}{n_y\hat{p}_i}\sim\chi^2_{k-1}$, $R_\alpha=\left\{\chi^2_{k-1,\alpha}>T\right\}$
  - permutation test: observe $\left\{X_i\right\}_{i=1}^n\sim P$, $\left\{Y_j\right\}_{j=1}^m\sim Q$. We have $H_0:\,P=Q$ vs $H_1:\,P\neq Q$
    - $T=\frac{1}{N!}\sum_L\mathbb{I}\left(g\left(X,\,Y,\,L\right)>g\left(X,\,Y,\,L^*\right)\right)\sim\text{Unif}\left(0,\,1\right)$, $R_\alpha=\left\{T<\alpha\right\}$, p-value $=\frac{1}{N!}\sum_{i=1}^{N!}\mathbb{I}\left(T_i>T_{\text{obs}}\right)$

Multiple testing

- Family-Wise Error Rate (FWER): $P\left[\text{exist false rej}\right]$. False Discovery Rate (FDR): $\mathbb{E}\left[\#\text{false rej}/\#\text{rej}\right]$
  - FWER $\geq$ FDR, control FWER is controlling FDR; under global null, FDR = FWER
- Multiple Testing: d tests in total
  - Sidak method: reject any test if its p-value $\leq 1-\left(1-\alpha\right)^{1/d}=\alpha_t$ (p-value independent), then FWER $\leq\alpha$
  - Bonferroni method: p-value $\leq\alpha/d$, then FWER $\leq\alpha$
  - Holm's procedure: $i^*=\min\left\{i:\,p_i>\frac{\alpha}{d-i+1}\right\}$, reject all H_i for $i<i^*$, then FWER $\leq\alpha$
  - BH procedure: $t_i=\frac{i\alpha}{d}$, $i^*=\max\left\{i:\,p_i<t_i\right\}$, reject all H_i for $i\leq i_{\max}$, then FDR $\leq\alpha$

Confidence set: a random set $C(X,\,\alpha)$ for $\alpha\in\left(0,\,1\right)$ that satisfies $P\left(\theta\in C\left(X\right)\right)\geq 1-\alpha$

- Probability inequalities: if a bounded $\hat{\theta}$ is an unbiased estimator of θ, we can apply Hoeffding's bound
- Inverting a test: $C\left(X,\,\alpha\right)=\left\{\theta:\,X\in A\left(\theta,\,\alpha\right)\right\}$, where A is the acceptance region of the test with "center" θ
  - Wald Interval: for MLE $\hat{\theta}$ we have $\text{se}\left(\hat{\theta}\right)=\frac{1}{\sqrt{nI_1\left(\hat{\theta}\right)}}$. Then $C\left(X,\,\alpha\right)=\left(\hat{\theta}-z_{\alpha/2}\text{se}\left(\hat{\theta}\right),\,\hat{\theta}+z_{\alpha/2}\text{se}\left(\hat{\theta}\right)\right)$
    - Note: can also be applied to another asympotically normal estimators
    - delta method: $\tau\left(\hat{\theta}_n\right)\pm z_{\alpha/2}\text{se}\left|\tau'\left(\hat{\theta}_n\right)\right|$
  - Likelihood Interval: for MLE $\hat{\theta}$ we have $C\left(X,\,\alpha\right)=\left\{\theta:\,\frac{\mathcal{L}\left(\theta\right)}{\mathcal{L}\left(\hat{\theta}\right)}>\exp\left(-\frac{1}{2}\chi^2_{d,\,\alpha}\right)\right\}$, where $d=\dim\Theta$
- Pivots: a function $Q\left(X,\,\theta\right)$ whose distribution does not depend on θ. Therefore, a interval for θ is $\left\{\theta:\,Q\left(X,\,\theta\right)\in C\left(Q,\,1-\alpha\right)\right\}$

Causal inference:

- Average treatment effect: $\tau=\mathbb{E}\left(Y\left(1\right)-Y\left(0\right)\right)$. Association: $\alpha=\mathbb{E}\left[Y\left(1\right)\,|\,W=1\right]-\mathbb{E}\left[Y\left(0\right)\,|\,W=0\right]$
  - α is easy to estimate: we have $\hat{\alpha}=\frac{1}{m}\sum_{i:W_i=1}Y^{obs}-\frac{1}{n-m}\sum_{i:W_i=0}Y^{obs}=\sum_{i=1}^n\frac{W_iY_i(1)}{m}-\frac{(1-W_i)Y_i(0)}{n-m}$
  - If $W\perp\left(Y\left(0\right),\,Y\left(1\right)\right)$, then $\alpha=\tau$ and $\hat{\alpha}=\hat{\tau}$ is an unbiased estimator for τ. Otherwise, we suffer from selection bias
- Exact p-values: we test hypothesis $H_0:\mathbb{E}[Y(1)]=\mathbb{E}[Y(0)]$ vs $H_1:\mathbb{E}[Y(1)]\neq\mathbb{E}[Y(0)]$
  - define L^* as the original treatment assignment and $\mathcal{L}$ as the set of possible assignments

– randomly assign treatments. For each assignment L we calculate $\hat{\tau}_L = \frac{1}{m}\sum_{i:W_{Li}=1} Y^{obs} - \frac{1}{n-m}\sum_{i:W_{Li}=0} Y^{\text{obs}}$
– we then have $p = \frac{1}{|\mathcal{L}|}\sum_L \mathbb{I}_{\{|\hat{\tau}_L|>|\hat{\tau}_{L^*}|\}}$
- No unmeasured confounding: we have $W \perp (Y(0), Y(1)|X$, where X is the confounding variable
  – $\tau = \mathbb{E}_X[\mathbb{E}(Y^{obs}|X, W=1)] - \mathbb{E}_X[\mathbb{E}(Y^{obs}|X, W=0)] = \mathbb{E}[\mu_1(X)] - \mathbb{E}[\mu_0(X)]$
  – Suppose we can estimate μ_1 and μ_0 by $\hat{\mu}_1$ and $\hat{\mu}_0$. Then we directly have the plug-in estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^n[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$
  – Alternative estimator (Horvitz-Thompson): $\hat{\tau} = \frac{1}{n}\sum_{i=1}^n[\frac{Y_i^{obs}W_i}{\pi(X_i)} - \frac{Y_i^{obs}(1-W_i)}{1-\pi(X_i)}]$, where $\pi(x) = \mathbb{E}(W|X = x)$ is the propensity score

Regression: estimate $r(x) = \mathbb{E}[Y \mid X = x]$, risk of $\hat{r}$: $R(\hat{r}) = \mathbb{E}[\mathcal{L}(Y, \hat{r}(X))]$
- If joint distribution of (X, Y) is known, under square, risk is minimized by $\hat{r}(x) = \mathbb{E}[Y \mid X = x]$
- Kernel regression: $r(x) = \sum_{i=1}^n w_i(x) Y_i$, where $w_i = K\left(\frac{x-x_i}{h}\right) / \sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)$
  – bandwidth h: larger h, smoother regression function, more bias, less variance
  – Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right)$
  – simple analysis
    ∗ assumption: 1) $y_i = r(x_i) + \epsilon_i$; 2) x_i is 1-dimensional, equally spaced in $[0, 1]$; 3) $r(x) = \mathbb{E}[Y \mid X = x]$ is L-Lipschitz $\left|\frac{d}{dx}r(x)\right| \le L$; 4) noise i.i.d: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$; 5) use spherical kernel $k(x) = \mathbb{I}(-1 \le x \le 1)$
    ∗ if bandwidth $h \ge 1/n$, then bias $b(x) \le Lh$, variance $v(x) \le \frac{\sigma^2}{nh}$.
    ∗ let bandwidth $h = \left(\frac{\sigma^2}{2nL^2}\right)^{1/3}$, then $\hat{R}(\hat{r}, r) \le 2\left(\frac{L\sigma^2}{n}\right)^{2/3}$
  – Extension: if r is β-smooth, then $b(x) \approx h^{2\beta}$, $v(x) \approx \frac{1}{nh^d}$. Consequently, $R(\hat{r}, r) \approx n^{-2\beta/(2\beta+d)}$
    ∗ the curse of dimensionality: the rate gets exponentially slow as d increases (we only get linearly slow in parametric regression).
    ∗ get rid: smoothness / parametric (e.g. linear) / sparsity assumption
- Gaussian sequence model: observe y^n where $y_i = \theta_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$
  – minimax estimator under squared loss: $\hat{\theta} = [y_1, \ldots, y_d]^T$, l_2 risk: $R\left(\hat{\theta}, \theta\right) = \frac{\sigma^2 d}{n}$. Not consistent when $d \gg n$
  – hard thresholding t: $\hat{\theta}_i = y_i\mathbb{I}(|y_i| \ge t)$. Solution to: $\hat{\theta} = \arg\min_\theta \frac{1}{2}\|y-\theta\|_2^2 + \frac{t^2}{2}\sum_{i=1}^d \mathbb{I}(\theta_i \ne 0)$
  – soft thresholding t: $\hat{\theta}_i = \text{sign}(y_i)\max(|y_i| - t, 0)$. Solution to: $\hat{\theta} = \arg\min_\theta \frac{1}{2}\|y-\theta\|_2^2 + t\sum_{i=1}^d |\theta_i|$
  – analyzing hard thresholding
    ∗ maximum of Gaussians: if $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then w.p. $\ge 1-\delta$, $\max_{i=1}^d |\epsilon_i| \le \sigma\sqrt{2\log(2d/\delta)}$
    ∗ let threshold $t = 2\sigma\sqrt{2\log(2d/\delta)/n}$, then w.p. $\ge 1-\delta$, $\left\|\hat{\theta}-\theta\right\|_2^2 \le 9\sum_{i=1}^d \min\left(\theta_i^2, \frac{t^2}{4}\right)$
    ∗ risk upper bound: $R\left(\hat{\theta}, \theta\right) \lesssim \sum_{i=1}^d \min\left(\theta_i^2, \frac{\sigma^2\log d}{n}\right)$
      · worst case: $R\left(\hat{\theta}, \theta\right) \lesssim \frac{\sigma^2 d\log d}{n}$, similar to minimax estimator
      · s sparse: $R\left(\hat{\theta}, \theta\right) \lesssim \frac{\sigma^2 s\log d}{n}$, consistent
      · l_1 sparse: if $\sum_{i=1}^d |\theta| \le R$, then $R\left(\hat{\theta}, \theta\right) \lesssim 2R\sigma\sqrt{\frac{\log d}{n}}$
- Linear regression: observe $\{(x_i, y_i)\}_{i=1}^n$, model $y_i = \langle x_i, \beta^*\rangle + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$
  – $\hat{\beta} = \arg\min_\beta \frac{1}{2}\sum_{i=1}^n (y_i - \langle x_i, \beta\rangle)^2 = \hat{\Sigma}^{-1}X^Ty = \left(X^TX\right)^{-1}X^Ty \sim \mathcal{N}\left(\beta^*, \sigma^2\left(X^TX\right)^{-1}\right)$
    ∗ consider $X \in \mathbb{R}^{n\times d}$ as random variables: random design matrix
    ∗ consider $X \in \mathbb{R}^{n\times d}$ as fixed: fixed design matrix
  – in-sample prediction error $\mathbb{E}\left[\left\|X\hat{\beta} - X\beta\right\|_2^2/n\right]$
    ∗ $X\hat{\beta} \sim \mathcal{N}\left(X\beta^*, \sigma^2 X\left(X^TX\right)^{-1}X^T\right)$
    ∗ $\mathbb{E}\left[\left\|X\hat{\beta} - X\beta\right\|_2^2/n\right] = \sigma^2\mathbb{E}\left[\text{tr}\left(X\left(X^TX\right)^{-1}X^T\right)\right]/n = \sigma^2 d/n$
  – l_2 error $\mathbb{E}\left[\left\|\hat{\beta}-\beta\right\|_2^2\right] = \frac{\sigma^2}{n}\text{tr}\left(\hat{\Sigma}^{-1}\right)$
    ∗ if eigenvalues of $\hat{\Sigma}^{-1}$ are lower bounded by c, then l_2 error $\le \frac{\sigma^2 d}{cn}$
    ∗ $I_n(\beta) = n\mathbb{E}\left[\frac{X^TX}{n\sigma^2}\right] = \frac{n\Sigma}{\sigma^2}$
    ∗ $\sqrt{n}\left(\hat{\beta} - \beta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\Sigma^{-1}\right)$
  – issues when d increases: 1) $\hat{\Sigma}$ not invertible; 2) error dominated by d; 3) too many solutions of $y = X\beta$ in high dimension
- High dimensional regression
  – hard thresholding type estimator
    ∗ penalized form: $\hat{\beta} = \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \frac{t^2}{2}\sum_{i=1}^d \mathbb{I}(\beta_i \ne 0)$
    ∗ constraint form (best subset regression): $\hat{\beta} = \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2$, subject to $\sum_{i=1}^d \mathbb{I}(\beta_i \ne 0) \le k$
  – soft thresholding type estimator (LASSO: least absolute selection and shrinkage operator)
    ∗ penalized form: $\hat{\beta} = \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + t\sum_{i=1}^d |\beta_i|$
    ∗ constraint form: $\hat{\beta} = \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2$, subject to $\sum_{i=1}^d |\beta_i| \le k$
    ∗ prediction error, w.p. $\ge 1-\delta$, $\frac{1}{n}\left\|X\hat{\beta} - X\beta\right\|_2^2 \le 4\sigma\|\beta^*\|_1\sqrt{\frac{2\log 2d/\delta}{n}}$

Bayesian inference
- Bayesian vs Frequentist
  – probability: subjective degree of belief; limiting frequency
  – goal: analyze belief; create procedures with frequency guarantee
  – θ: random variable; fixed
  – X: random variable; random variable
  – use Bayes: Yes, to update beliefs; Yes if it leads to procedure with good frequentist behavior, otherwise no
- Setup: given a collection of distributions $\{P_\theta : \theta \in \Theta\}$; prior: $\theta \sim \pi$; likelihood: $p(X^n \mid \theta) \sim P_\theta$
  – compute posterior belief: $\pi(\theta \mid X) = p(X \mid \theta)\pi(\theta) / \int_\theta p(X \mid \theta)\pi(\theta) \propto p(X \mid \theta)\pi(\theta)$
  – Bayesian: success if you can calculate posterior
  – frequentist: success if the posterior concentrates on the true parameter θ^*
- Credible set C_α for posterior distribution: $\int_{C_\alpha}\pi(\theta|X^n)\,d\theta = 1-\alpha$
- Frequestist view
  – Consistency: $\forall\epsilon > 0$, $\pi\left(\{\theta : \|\theta - \theta^*\| \ge \epsilon\} \,\middle|\, X^n\right) \to 0$
  – Convergence rate (based on consistency): for $\forall\delta > 0$, we define its convergence rate $\epsilon = \sup\{\epsilon' : \pi_\theta(\|\theta - \theta^*\| \ge \epsilon' \mid X) \le \delta\}$
- Bernstein-von Mises Theorem: if prior is continuous and strictly positive around θ^*, then for a fixed dimension d, as $n \to \infty$, the posterior is close to Gaussian:
  – $\left\|\pi(\theta; X^n) - \mathcal{N}\left(\hat{\theta}_n, 1/nI_1\left(\hat{\theta}_n\right)\right)\right\|_{\text{TV}} \to 0$, where $\hat{\theta}_n$ is MLE, and TV is total variance
  – in high-dimensional or non-parametric setting, strictly positive around θ^* is hard to be true
  – if it is true, can use credible sets like Wald test
- Ideas in choosing prior: 1) by convenience (minimax); 2) doesn't matter (low-dimensional / parametric); 3) based on data: empirical Bayes; 4) non-informative prior; 5) Joffrey's prior: $\pi(\theta) \sim \sqrt{I(\theta)}$: invariant under transformations; 6) hierarchical prior
- Regularizer: 1) LASSO: posterior mode with Laplace prior; 2) ridge regression: Gaussian prio; 3) estimating Bernoulli prob with laplace smoothing: posterior mean with Beta prior

Monte Carlo
- MC Integration: compute $\mu = \mathbb{E}_{X\sim P}[f(X)]$
  – directly sampling from P: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n f(X_i)$. If cannot sample from P
  – importance weighting: $\mu = \mathbb{E}_{X\sim Q}\left[\frac{p(X)}{q(X)}f(X)\right] = \mathbb{E}_{X\sim Q}[w(X)f(X)]$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n w(X_i)f(X_i)$
  – if doesn't work, use Markov Chain Monte Carlo
- Markov chain: transition probability $T(x_i, x_{i+1}) = P\left(x_{i+1} \mid x_i\right)$
  – limiting distribution always exists: $\pi(x) = P\left(\lim_{n\to\infty} x_n = x\right)$
  – detailed balance: condition for reaching limiting distribution: $\forall x, y$, $\pi(x)T(x, y) = \pi(y)T(y, x)$
  – Markov LLN: MC $\{X^n\}$, limiting distribution π, then $\frac{1}{n}\sum_{i=1}^n f(X_i) \to \mathbb{E}_\pi[f(x)]$ (weak dependence)
- Metropolis-Hastings: at step i, $y \sim q(y \mid X = x_i)$, accept w.p. $r = \min\left\{1, \frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}\right\}$
  – proposal distribution q, often $q(y \mid x_i = x) \sim \mathcal{N}(x, \sigma^2)$
  – if $q(\cdot \mid \cdot)$ is symmetric, we sample more from high-prob regions, and accept w.p. 1 if prob going up
  – limiting distribution of this Markov chain is f: $T(y, x) = q(y \mid x)r$, verify $f(x)T(y, x) = f(y)T(x, y)$

Bootstrap: test the variability of a point estimate (variance, confidence set)
- Given $X^n \sim P$, estimate empirical distribution $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \in A)$, then draw from P_n
- Bootstrap variance estimate
  – draw bootstrap sample $X^{*n} \sim P_n$, compute $\hat{\theta}_n^* = g(X^{*n})$
  – repeat B times, yielding $\hat{\theta}_{n,1}^*, \hat{\theta}_{n,2}^*, \ldots, \hat{\theta}_{n,B}^*$, $\overline{\theta} = \frac{1}{B}\sum_{j=1}^B \hat{\theta}_{n,j}^*$
  – bootstrap variance $\hat{s}^2 = \frac{1}{B}\sum_{j=1}^B\left(\hat{\theta}_{n,j}^* - \overline{\theta}\right)^2$
- Bootstrap confidence interval

- draw bootstrap sample $X^{*n} \sim P_n$, compute $\hat{\theta}_n^* = g\left(X^{*n}\right)$
- repeat B times, yielding $\hat{\theta}_{n,1}^*$, $\hat{\theta}_{n,2}^*$, ..., $\hat{\theta}_{n,B}^*$
- calculate Bootstrap CDF: $\hat{G}\left(t\right) = \frac{1}{B} \sum_{j=1}^{B} \mathbb{I}\left(\sqrt{n}\left(\hat{\theta}_{n,j}^* - \hat{\theta}_n\right) \le t\right)$
- $C_n = \left[\hat{\theta}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}},\, \hat{\theta}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right]$ where $g_{\alpha/2} = \hat{G}^{-1}\left(\alpha/2\right)$, $g_{1-\alpha/2} = \hat{G}^{-1}\left(1-\alpha/2\right)$

- Bootstrap theorem: $F_n\left(t\right) = P\left(\sqrt{n}\left(\hat{\theta}_n - \theta\right) \le t\right)$, $\hat{F}_n\left(t\right) = P\left(\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) \le t\right)$, where $\hat{\theta}_n$ is the estimator based on X^n.
  - Example: if $\mu_3 = E\left|X_i\right|^3 < \infty$, and $\hat{\theta}_n$ is sample mean, then $\sup_t \left|\hat{F}_n\left(t\right) - F_n\left(t\right)\right| = O_P\left(\frac{1}{\sqrt{n}}\right)$

Model selection

- Definition: A scheme S is model selection consistent: as sample size $n \to \infty$, $P\left(S \text{ selects wrong model}\right) \to 0$
- Cross validation:
  - prediction consistent: as test size $n_{\text{te}} \to \infty$, risk estimation converges to its expectation.
  - not model selection consistent
- Akaike Information Criterion (AIC): define for model space Θ
  - $\text{AIC}\left(\Theta\right) = 2l(\hat{\theta}) - 2\dim\Theta$, where parameter $\hat{\theta}$ is the MLE within Θ and $l\left(\hat{\theta}\right) = \log P\left(X \mid \hat{\theta}\right)$
  - not model selection consistent
- Bayesian Information Criterion (BIC): define for model space Θ with sample size n.
  - $\text{BIC}\left(\Theta\right) = 2l\left(\hat{\theta}\right) - \dim\Theta\log n$, where parameter $\hat{\theta}$ is the MLE within Θ and $l\left(\hat{\theta}\right) = \log P(X \mid \hat{\theta})$
  - compared to AIC, prefer sparser / simpler models
  - model selection consistent
- Application of AIC and BIC: compare among different model spaces and select the one that maximizes AIC / BIC.
- Main take-away
  - if goal is "prediction", and have reasonable sample size & computational budget ⇒ cross validation
  - if goal is "prediction", less samples / computational budget ⇒ AIC
  - if goal is "selecting true model" ⇒ BIC

	$\mathcal{N}\left(\mu,\, \Sigma\right)$	$\text{Gamma}\left(\alpha,\, \beta\right)$	$\text{Exp}\left(\lambda\right)$	$\text{Poisson}\left(\lambda\right)$
support	$\mu + \text{span}\left(\Sigma\right)$	$\left(0,\, +\infty\right)$	$\left(0,\, +\infty\right)$	$k \in \mathbb{N}^+ \cup \{0\}$
pdf	$\frac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{(2\pi)^{d/2}\lvert\det(\Sigma)\rvert^{1/2}}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp\left(-\beta x\right)$	$\lambda e^{-\lambda x}$	$\frac{\lambda^k \exp(-\lambda)}{k!}$
cdf	-	$\frac{1}{\Gamma(\alpha)}\gamma\left(\alpha,\, \beta x\right)$	$1 - e^{-\lambda x}$	$e^{-\lambda}\sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$
mean	μ	α/β	λ^{-1}	λ
variance	Σ	α/β^2	λ^{-2}	λ
mgf	$\exp\left(\mu^T t + \frac{1}{2}t^T\Sigma t\right)$	$\left(1 - t/\beta\right)^{-\alpha}$ for $t < \beta$	$\frac{\lambda}{\lambda - t}$ for $t < \lambda$	$\exp\left(\lambda\left(z-1\right)\right)$
conv	$\mathcal{N}\left(\mu_x + \mu_y,\, \sigma_x^2 + \sigma_y^2\right)$	$\text{Gamma}\left(\alpha_1 + \alpha_2,\, \beta\right)$	-	$\text{Poisson}\left(\lambda_1 + \lambda_2\right)$
fisher	$\begin{bmatrix} 1/\sigma^2 & \\ & 1/2\sigma^4 \end{bmatrix}$	-	λ^{-2}	λ^{-1}
MoM	$\hat{\mu} = \overline{X}_n,\ \sigma^2 = \frac{1}{n}\sum_{i=1}^n \left(X - \overline{X}_n\right)$			$\hat{\lambda} = \overline{X}_n$
MLE	same			same

	$\text{Ber}\left(p\right)$	Rademacher	$\text{Binom}\left(n,\, p\right)$	$\text{Geometric}\left(p\right)$	$\chi^2\left(k\right)$
support	$\{0,\, 1\}$	$\{-1, 1\}$	$k \in \mathbb{N} \cup \{0\}$	$\mathbb{N}^+$	$x \in \left(0,\, \infty\right)$
pdf	-	-	$\binom{n}{k} p^k \left(1-p\right)^{n-k}$	$\left(1-p\right)^{k-1} p$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-\frac{x}{2}}$
cdf	-	-	-	$1 - \left(1-p\right)^k$	-
mean	p	0	np	$1/p$	k
variance	$p\left(1-p\right)$	1	$np\left(1-p\right)$	$\left(1-p\right)/p^2$	$2k$
mgf	$q + pe^t$	$\exp\left(t^2/2\right)$	$\left(1 - p + pe^t\right)^n$	$\frac{pe^t}{1-(1-p)\exp(t)}$	$\left(1-2t\right)^{-k/2}$ for $t < \frac{1}{2}$
conv	-	-	$\text{Binom}\left(n+m,\, p\right)$	-	-
fisher	$1/p\left(1-p\right)$	-	$n/p\left(1-p\right)$	-	-
MoM			$\hat{p} = \overline{X}_n/n$	$\hat{p} = 1/\overline{X}_n$	
MLE			same	same	

	$\text{Beta}\left(\alpha,\, \beta\right)$
support	$\left(0,\, 1\right)$
pdf	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}\left(1-x\right)^{\beta-1}$
cdf	-
mean	$\frac{\alpha}{\alpha+\beta}$
variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
mgf	$1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
conv	-
fisher	$\begin{bmatrix} \text{Var}\left[\ln X\right] & \text{Cov}\left[\ln X,\, \ln\left(1-X\right)\right] \\ \text{Cov}\left[\ln X,\, \ln\left(1-X\right)\right] & \text{Var}\left[\ln\left(1-X\right)\right] \end{bmatrix}$
MoM	
MLE	

Table 1: Probability distribution