Probability Inequalities

Thm 1 (Gaussian Tail Inequality):
Let $X \sim \mathcal{N}(0,1)$. Then
Additionally:
Thm 2 (Markov Inequality): Let X be a non-negative random variable s.t. $\mathbb{E}(X)$ exists.
Then $\forall\ t > 0$
Thm 3 (Chebyshev's Inequality): Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$.
Then:
Lemma 4: Let $\mathbb{E}(X) = 0$ and $a \le X \le b$.
Then
Lemma 5: Let X be any random variable.
Then
Thm 6 (Hoeffding's Inequality): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$, $a \le X_i \le b$.
Then $\forall \epsilon > 0$
Thm 9 (McDiarmid): $X_1, \ldots, X_n$ indep't. If $\sup_{x_1,\ldots,x_n,x_i'} |g(x_1,\ldots,x_n) - g_i^*(x_1,\ldots,x_n)| \le c_i\ \forall i, \implies$

$$\mathbb{P}\left(g(X_1,\ldots,X_n) - \mathbb{E}(g(X_1,\ldots,X_n)) \ge \epsilon\right) \le e^{-2\epsilon^2/\sum_i c_i^2} \quad (1)$$

where $g_i^* = g$ with x_i replaced by x_i'.
Thm 12 (Cauchy-Schwartz inequality):
Thm 13 (Jensen's inequality):
Ex 15 (Kullback Leibler distance):
Thm 18:
O_p and o_p: $X_n = o_p(1)$ if $\forall\ \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0$.
$X_n = O_p(1)$ if $\forall\ \epsilon > 0$, $\exists\ C > 0$ s.t. $\lim_{n\to\infty} \mathbb{P}(|X_n| > C) \le \epsilon$.
$X_n = o_p(a_n)$ if $X_n/a_n = o_p(1)$ and $X_n = O_p(a_n)$ if $X_n/a_n = O_p(1)$.

Shattering

Note: remember uniform bounds and union bound.
F a finite set, $|F| = n$, and $G \subset F$. $\mathcal{A}$ is a class of sets.
$\mathcal{A}$ picks out G if $\exists A \in \mathcal{A}$ s.t. $A \cap F = G$.
Let $S(\mathcal{A}, F) = |\{G \subset F \text{ picked out by } \mathcal{A}\}| \le 2^n$.
F is **shattered** by $\mathcal{A}$ if $S(\mathcal{A}, F) = 2^n$ (ie if $\mathcal{A}$ picks out all $G \subset F$).
Let $\mathcal{F}_n$ be all finite sets with n elements.
The **shatter coefficient** $s_n(\mathcal{A}) = \sup_{F\in\mathcal{F}_n} s(\mathcal{A}, F) \le 2^n$.
The **VC dimension** $d(\mathcal{A})$ = the largest n s.t. $s_n(\mathcal{A}) = 2^n$.
Thm 5: $\forall \epsilon > 0$, $\mathbb{P}(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon) \le 8 s_n(\mathcal{A}) e^{-n\epsilon^2/32}$

Random Samples

For $X_1, \ldots, X_n \sim F$ a **statistic** is any $T = g(X_1, \ldots, X_n)$.
E.g. $\overline{X}_n$, $S_n = \sum_i (X_i - \overline{X}_n)^2/(n-1)$, $\left(X_{(1)}, \ldots, X_{(n)}\right)$
Notes: $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i)$, $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(X_i)/n$, $\mathbb{E}(S_n)^2 = \mathrm{Var}(X_i)$, $X_{1,\ldots,n} \sim \mathrm{Bern}(p) \implies \sum_i X_i \sim \mathrm{Bin}(n,p)$, $X_{1,\ldots,n} \sim \mathrm{Exp}(\beta) \implies \sum_i X_i \sim \Gamma(n,\beta)$, $X_{1,\ldots,n} \sim \mathcal{N}(0,1) \implies \sum_i X_i^2 \sim \chi_n$.
Thm. 1: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2) \implies \overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.

Convergence

$X, X_1, X_2, \ldots$ random variables.
(1) X_n converges **almost surely** $X_n \xrightarrow{a.s.} X$ if $\forall \epsilon > 0$

(2) X_n converges **in probability** $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$

(3) X_n converges **in quadratic mean** $X_n \xrightarrow{qm} X$ if

(4) X_n converges **in distribution** $X_n \rightsquigarrow X$ if

$\forall t$ on which F_X is continuous.

Thm 7:

Thm 10a: X,X_n,Y,Y_n random variables. Then

Thm 10b (Slutzky's Thm): X,X_n,Y_n random variables. Then

Thm 12 (Law of Large Numbers): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu \implies \overline{X}_n \xrightarrow{\text{qm}} \mu$.
Thm 14 (CLT): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\mathrm{Var}(X_i) = \sigma^2$
$\implies \sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$
$\implies \overline{X}_n \rightsquigarrow \mathcal{N}(\mu, \sigma^2/n)$
$\implies \sqrt{n}(\overline{X}_n - \mu)/S_n \rightsquigarrow \mathcal{N}(0,1)$
Thm 18 (delta method): If $\sqrt{n}(Y_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$, $g'(\mu) \neq 0$
$\implies \sqrt{n}(g(Y_n) - g(\mu))/|g'(\mu)|\sigma \rightsquigarrow \mathcal{N}(0,1)$
ie $Y_n \approx \mathcal{N}(\mu, \sigma^2/n) \implies g(Y_n) \approx \mathcal{N}(g(\mu), g'(\mu)^2\sigma^2/n)$
Thm 18b (2nd order delta method):

Sufficiency

If $X_1, \ldots, X_n \sim p(x;\theta)$, T **sufficient** for θ if $p(x^n|t;\theta) = p(x^n|t)$.
Thm 9 (factorization): for $X^n \sim p(x;\theta)$, $T(X^n)$ sufficient for θ if the joint probability can be factorized as.

T is a **minimal sufficient statistic (MSS)** if T is sufficient and $T = g(U)$ for all other sufficient stats U.
Thm 15: T is a MSS if:

Parametric Point Estimation

Method of Moments: Define equations

And solve for $\hat{\theta}$.
Maximum Likelihood (MLE): The MLE is

Often suffices to solve for θ in $\frac{\partial l(\theta)}{\partial \theta} = 0$. The MLE is **equivariant** $\implies$ if $\eta = g(\theta)$ then $\hat{\eta} = g(\hat{\theta})$.
Bayes Estimation: For prior $\pi(\theta)$, choose

Mean Squared Error (MSE): The MSE is

$$\mathrm{MSE} = \mathbb{E}(\hat{\theta} - \theta)^2 = \int (\hat{\theta} - \theta)^2 p(x^n;\theta) dx^n = \mathrm{bias}(\hat{\theta})^2 + Var(\hat{\theta}) \quad (2)$$

Defs: **bias**$(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. We say $\hat{\theta}$ is **consistent** if $\hat{\theta} = \hat{\theta}_n \xrightarrow{p} \theta$.
The **standard error** of $\hat{\theta}$, $\mathrm{se}(\hat{\theta})$, is the standard deviation of $\hat{\theta}$.

Risks and Estimators

$L(\theta, \hat{\theta})$ is the **loss** of an estimator $\hat{\theta} = \hat{\theta}(x^n)$ for $x^n \sim p(x^n;\theta)$.
The **risk** of this $\hat{\theta}$ is

When $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, the risk is the MSE.
The **max risk** of $\hat{\theta}$ over a set $\theta \in \Theta$ is

The **minimax risk** is

The **minimax estimator** is

The **Bayes risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

The **posterior risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

where $\pi(\theta|x^n) = \frac{\mathbb{P}(x^n;\theta)\pi(\theta)}{m(x^n)}$ is the posterior over θ.
The **Bayes estimator** is

which equals the posterior mean $\mathbb{E}(\theta|x^n)$ when $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$, the posterior median when $L(\theta,\hat{\theta}) = |\theta-\hat{\theta}|$, and the posterior mode when $L(\theta,\hat{\theta}) = \mathbb{I}[\theta \neq \hat{\theta}]$.
Thm 10: If $\hat{\theta}$ is a Bayes estimator for some prior π and $R(\theta,\hat{\theta})$ is constant, then $\hat{\theta}$ is a minimax estimator.
Note: The MLE is approximately minimax (as n increases, if dimension of the parameter is fixed).

Distributions

Discrete distributions: (a) Bernoulli
(b) Binomial
(c) Poisson
Continuous distributions: (b) Normal

Expected Values

The **mean** or **expected value** of $g(X)$ is
Related properties and definitions:
(g) Cov(X,Y) =
(h) Cov(X,Y) =
(i) $\rho(X,Y) =$
The **conditional expectation** of Y given X is the random variable $g(X) = \mathbb{E}(Y|X)$, where

The *Law of Total/Iterated Expectation* is
The *Law of Total Variance* is
The *Law of Total Covariance* is

Aymptotic (Large Sample) Theory

A random sequence A_n is:
1.
2.
3.
4.
If $Y_n \rightsquigarrow Y \implies Y_n = O_p(1)$
If $\sqrt{n}(Y_n - c) \rightsquigarrow Y \implies Y_n = O_p(1/\sqrt{n})$

Distances Between Distributions

For distributions P and Q with pdfs p and q:
$K(P,Q) = \int p\log(p/q)$ **Kullback-Leibler** divergence
A model is **identifiable** if: $\theta_1 \neq \theta_2 \implies K(\theta_1,\theta_2) > 0$.

Consistency

$\hat{\theta}_n = T(X^n)$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ (ie if $\hat{\theta}_n - \theta = o_p(1)$).
To show consistency, can show: $\text{Bias}^2(\hat{\theta}_n) + \text{Var}(\hat{\theta}_n) \to 0$.
The MLE is consistent under regularity conditions.
MLE not consistent when number of params (or support?) grows.

Score and Fisher Information

The **score function** is $S(\theta) = \frac{\partial}{\partial\theta} l(\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^n \log p(x_i|\theta)$.
The **Fisher information** is defined as

$$I_n(\theta) = \mathbb{E}_\theta\left[S(\theta)^2\right] = \text{Var}_\theta\left[S(\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right] \qquad (3)$$

and $I_n(\theta) = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X_1;\theta)\right] = nI_1(\theta)$.
The **observed information** $\hat{I}_n(\theta) = -\sum_i \frac{\partial^2}{\partial\theta^2}\log p(X_i;\theta)$.
Vector case: $S(\theta) = \left[\frac{\partial l(\theta)}{\partial\theta_i}\right]_{i=1,\ldots,K}$ $\quad I_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2 l(\theta)}{\partial\theta_i\partial\theta_j}\right]_{i,j=1,\ldots,K}$

Efficiency and Robustness

For an estimator $\hat{\theta}_n(X^n)$ of θ, where $X^n \overset{\text{iid}}{\sim} p(x|\theta)$:
If $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0,v^2)$, then v^2 is the **asymptotic-Var($\hat{\theta}_n$)**.
E.g. for $\hat{\theta}_n = \overline{X}_n$: $v^2 = \sigma^2 = \text{Var}(X_i) = \lim_{n\to\infty} n\text{Var}(\overline{X}_n)$.
In general, asymptotic-Var($\hat{\theta}_n$) $v^2 \neq \lim_{n\to\infty} n\text{Var}(\hat{\theta}_n)$.
We will use approx: $\text{Var}(\hat{\theta}_n) \approx v^2/n$.
For param $\tau(\theta)$, $v(\theta) = \frac{|\tau'(\theta)|^2}{I_1(\theta)}$ is the **Cramer-Rao lower bound**.
for most estimators $v^2 \geq v(\theta)$.
If $\sqrt{n}(\hat{\theta}_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0,v(\theta))$ (ie if $v^2 = v(\theta)$) $\implies \hat{\theta}_n$ **efficient**.
usually, $\sqrt{n}(\tau(\hat{\theta}_{\text{mle}}) - \tau(\theta)) \rightsquigarrow \mathcal{N}(0,v(\theta)) \implies$ MLE efficient.
The **standard error** of **efficient** $\hat{\theta}_n$ is $se = \sqrt{\text{Var}(\hat{\theta}_n)} \approx \sqrt{\frac{1}{I_n(\theta)}}$.
The **estimated standard error** of **efficient** $\hat{\theta}_n$ is $\hat{se} \approx \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$.
For efficient $\hat{\theta}_n$, $\hat{\tau} = \tau(\hat{\theta}_n)$, $se \approx \sqrt{\frac{|\tau'(\theta)|^2}{I_n(\theta)}}$, and $\hat{se} \approx \sqrt{\frac{|\tau'(\hat{\theta}_n)|^2}{I_n(\hat{\theta}_n)}}$.
In general, **asymptotic normality** is when:
$\frac{\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)}{\sqrt{\text{Var}(\hat{\theta}_n)}} \rightsquigarrow \mathcal{N}(0,1) \implies \hat{\theta}_n \rightsquigarrow \mathcal{N}(\mathbb{E}(\hat{\theta}_n), \text{Var}(\hat{\theta}_n))$.
If $\sqrt{n}(W_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0,\sigma_W^2)$ and $\sqrt{n}(V_n - \tau(\theta)) \rightsquigarrow \mathcal{N}(0,\sigma_V^2)$
$\implies$ **asymptotic relative efficiency** $\text{ARE}(V_n,W_n) = \sigma_W^2/\sigma_V^2$.
Often there is a tradeoff between efficiency and robustness. (?)

Hypothesis Testing

Null hypothesis $H_0 : \theta \in \Theta_0$, **alternative** $H_1 : \theta \in \Theta_1$.
Type I error: If H_0 true but we reject H_0.
To construct a test:

1. Choose a test statistic $W = W(X_1,\ldots,X_n)$
2. Choose a rejection region R (4)
3. If $W \in R$, reject H_0 otherwise retain H_0

For rejection region R, the **power function** $\beta(\theta) = \mathbb{P}_\theta(X^n \in R)$.
Want **level-**α test ($\sup_{\theta\in\Theta_0}\beta(\theta) \leq \alpha$) that maximizes $\beta(\theta \in \Theta_1)$.
A level-α test with power fn β is **uniformly most powerful** if:
$\beta(\theta) \geq \beta'(\theta)\ \forall\theta \in \Theta_1\ \forall\beta' \neq \beta$.

Neyman-Pearson Test

For simple $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, reject H_0 if $\frac{L(\theta_1)}{L(\theta_0)} > k$.
where k chosen s.t. $\mathbb{P}(\frac{L(\theta_1)}{L(\theta_0)} > k) = \alpha$.

Wald Test

For $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, reject H_0 if $\left|\frac{\hat{\theta}_n - \theta_0}{se}\right| > z_{\alpha/2}$.
where $z_{\alpha/2}$ is the inverse standard-normal CDF of $1 - \frac{\alpha}{2}$.
and $\hat{\theta}_n$ is an unbiased estimator for θ.
and $se = \sqrt{\text{Var}(\hat{\theta}_n)}$. Can also use $\hat{se} =_{\text{eg.}} \sqrt{S_n^2/n}$.
and if $\hat{\theta}_n$ efficient, can approx: $se \approx \sqrt{\frac{1}{I_n(\theta)}}$ or $\hat{se} \approx \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$.

Likelihood Ratio Test

For $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \notin \Theta_0$, reject H_0 if $\lambda(x^n) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \leq c$.
where $L(\hat{\theta}_0) = \sup_{\theta\in\Theta_0} L(\theta)$ and $L(\hat{\theta}) = \sup_{\theta\in\Theta} L(\theta)$.
and c chosen s.t. $\mathbb{P}(\lambda(x^n) \leq c) = \alpha$.
Thm: under $H_0 : \theta = \theta_0 \implies W_n = -2\log\lambda(X^n) \rightsquigarrow \chi_1^2$
$\implies$ reject H_0 if $W_n > \chi_{1,\alpha}^2$.
Also: for $\theta = (\theta_1,\ldots,\theta_k)$, if H_0 fixes some of the parameters
$\implies -2\log\lambda(X^n) \rightsquigarrow \chi_\nu^2$, where $\nu = \dim(\Theta) - \dim(\Theta_0)$.

P-Values

The **p-value** $p(x^n)$ is the smallest α-level s.t. we reject H_0.
Thm: For a test of the form: reject H_0 when $W(x^n) > c$,
$\implies p(x^n) = \sup_{\theta\in\Theta_0}\mathbb{P}_\theta(W(X^n) \geq W(x^n)) = \sup_{\theta\in\Theta_0}[1 - F(W(x^n)|\theta)]$.
Thm: Under $H_0 : \theta = \theta_0$, $p(x^n) \sim \text{Unif}(0,1)$.

Permutation Test

$X^n \sim F$, $Y^m \sim G$, $H_0 : F = G$, $H_1 : F \neq G$
Let $Z = (X^n, Y^m)$ and $L = (1,\ldots,1,2,\ldots,2)$.
Let $W = g(L,Z) = |$(ave of 1 labeled pts) $-$ (ave of 2 labeled pts)$|$.
Let $p = \frac{1}{N!}\sum_\pi \mathbb{I}(g(L_\pi,Z) > g(L,Z)) \implies$ reject H_0 when $p < \alpha$.