0.1 Calculus cheat sheet

Logs: $log_b(M*N) = log_bM + log_bN$ • $log_b(\frac{M}{N}) = log_bM - log_bN$ • $log_b(M^k) = klog_bM$ • $e^n e^m = e^{n+m}$
Derivatives: $(x^n)' = nx^{n-1}$ • $(e^x)' = e^x$ • $(e^{u(x)})' = u'(x)e^x$ • $(log_e(x))' = (lnx)' = \frac{1}{x}$ • $(f(g(x)))' = f'(g(x))g'(x)$
Integrals: $\int_a^b f(x)dx = \int_{g(a)}^{g(b)} f(g(u))g'(u)du$ where $g(u) = x$ • $\int_a^b u(x)v'(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx$
Infinite series and sums: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ • $(1+\frac{a}{n})^n \longrightarrow e^a$
• $ln(1+x) = 1 - x + \frac{x^2}{2} - \frac{x^3}{3} + \cdots = \sum_{n=0}^{\infty}(-1)^n \frac{x^n}{n}$ • $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{n=0}^{\infty} a^x$ for $|x| < 1$
L'Hopitale: $\lim_{n\to c} f(x)/g(x) = \lim_{n\to c} f'(x)/g'(x)$ if $\lim_{n\to c} f(x) = \lim_{n\to c} g(x) = 0/\infty/-\infty$

0.2 Expectation

Conditional expectation: $p_{X|Y}(x|y) = \frac{p_{x,y}(x,y)}{p_y(y)}$
Bayes theorem: $P(E_i \mid B) = \frac{P(B|E_i)P(E_i)}{\sum_{j=1}^{\infty} P(B|E_j)P(E_j)} = \frac{P(B|E_i)P(E_i)}{P(B)}$, where $E_1, E_2, \dots$ form a partition of the sample space.
Expectation: $E(X) = \sum_x xP(X = x)$, also written $E(X) = \sum_{x\in S} X(s)p(s)$, where $p(s)$ is the probability that element $s \in S$
• $E(g(X)) = \sum_i g(x_i)p_X(x_i)$ • $E(aX + b) = aE(X) + b$ • $E(X + Y) = E(X) + E(Y)$
• $P_{Y,X}(Y > X) = E_{Y,X}\mathbb{I}\{Y > X\} = E_X E_Y[\mathbb{I}\{Y > X\} \mid X] = E_X P_Y(Y > X \mid X) = \int_{\mathbb{R}} P_Y(Y > x \mid X = x)F_X(dx)$
Variance: $Var(X) = E((X - E(X)))^2) = \sigma^2$
• $Var(X) = E(X^2) - \mu^2$ • $Var(aX + b) = a^2Var(X)$ • • $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
Covariance: $Cov(X, Y) = E((X - E(X)(Y - E(Y)) = E(XY) - E(X)E(Y)$
• $Cov(X, X) = Var(X)$ • $Cov(aX, bY) = abCov(X, Y)$ • $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$
Law of iterated expectation: $E(E(Y \mid X)) = E(Y)$
Proof: $E(Y \mid X) = \sum_y y\frac{f_{X,Y}(X,y)}{f_X(X)}$, $E(E(Y \mid X)) = \sum_x \sum_y \left(y\frac{f_{X,Y}(x,y)}{f_X(x)}\right) f_X(x) = \sum_x \sum_y yf_{X,Y}(x, y) = \sum_y yf_Y(y) = E(Y)$
Law of total probability: $P(E) = \sum_{i=-\infty}^{\infty} P(E \mid X = x)P(X)$ and $P(E) = \int_{-\infty}^{\infty} P(E \mid X = x)f(x)dx$
Variance decomposition formula: $Var(Y) = E(Var(Y \mid X)) + Var(E(Y \mid X))$
Cauchy-Schwartz inequality: $E(UV)^2 \le E(U^2)E(V^2)$, with equality if $P(cU = U) = 1$ for some constant, c
Transformations of random variables: For X with density f_X and $Y = g(X)$
$F_Y(y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$ • $f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$
Stationarity: $\{X_1, \dots, X_n\} \overset{d}{=} \{X_{m+1}, \dots, X_{m+n}\}$ • $EX_1 = EX_n$ • $Cov(X_1, X_n) = Cov(X_{m+1}, X_{m+n})$, called $c(n)$ where n is lag. We can prove $c(n) \xrightarrow{p} 0$ using Chebychev and solving for $Var\overline{X}_n$ as a function of $c(n)$: $P(|\overline{X}_n - EX_1| > \epsilon) \le \frac{2}{n}\sum_i c(i) \longrightarrow 0$

0.3 Inequalities

Jensen inequality: $E(g(x)) \ge g(E(x))$ for $g(x)$ convex
Markov inequality: For $X \ge 0$, $P(X \ge t) \le \frac{E(X)}{t}$ $\forall t > 0$. **Proof:**
Let $Y = \begin{cases} 1 & X \ge t \\ 0 & \text{else} \end{cases}$, then $tY \le X$ since $\begin{cases} X \ge t & t1 \le X \\ X < t & t0 < X \end{cases} \implies E(tY) \le E(X) \implies tP(X \ge t) \le E(X)) \implies P(X \ge t) \le \frac{E(X)}{t}$
Chebyshev inequality: $P(|X - E(X)| \ge t) \le \frac{Var(X)}{t^2}$ $\forall t > 0$.
Proof: $P(|X - E(X)| \ge t) = P((X - E(X))^2 \ge t^2) \le \frac{E((X-E(X))^2)}{t^2} = \frac{Var(X)}{t^2}$, by Markov inequality

Exponential inequality: $P(X > a) \le e^{-\theta a}E(e^{\theta X})$ for all $\theta > 0$.
Proof: $P(X > a) = P(\theta X > \theta a)$ for $\theta > 0 \implies P(e^{\theta X} > e^{\theta a}) \le e^{-\theta a}E(e^{\theta X})$, by Markov inequality
(Corrolary) Upper bound on large deviations: $P(S_n < na) \le e^{-nI(x)}$.
Proof: $P(S_n > a) \le e^{-\theta a}E(e^{\theta S_n}) = e^{-\theta a}\prod_i E(e^{\theta X_i}) = e^{-\theta a}E(e^{\theta X_1})^n = e^{-\theta a + n\psi(\theta)}$, by exponential inequality, iid, where $\psi(\theta) = \log Ee^{\theta X_1}$
$P(S_n > na) = e^{-n(\theta(x)a - n\psi(\theta(x))}$; minimizing RHS w.r.t $\theta \implies e^{-nI(x)}$ where $I(x) = \theta(x)a - n\psi(\theta(x))$

0.4 Generative functions

• $\phi_X(t) = E\exp(itX)$ • $N(\mu, \sigma^2)$: $\exp(it\mu - \frac{1}{2}\sigma^2t^2)$ • $Exp(\lambda)$: $(1 - it\lambda^{-1})^{-1}$ • $Poisson(\lambda)$: $\exp\lambda(e^{it} - 1)$ • $E[X^k] = i^{-k}E[X^k]$
• $\phi_{a_1X_1+\cdots+a_nX_n}(t) = \phi_{X_1}(a_1t)\dots\phi_{X_n}(A_nt)$ for X_i independent • $M_X(t) = E\exp(tX)$

0.5 Weak law of large numbers

For $X_1, X_2, \dots, X_n$ i.i.d. with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then for any $\epsilon > 0$

$$P(|\overline{X}_n - \mu| > \epsilon) \le \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty, \text{ by Chebyshev inequality,} \quad X_1 + \cdots + X_n = S_n \approx ES_n, \text{ the "meta result"}$$

0.5.1 Convergence in probability

Convergence in probability: $X_n \xrightarrow{p} X$ when $P(|X_n - X| > \epsilon) \longrightarrow 0$ as $n \longrightarrow \infty$
Continuous mapping theorem: if $X_n \xrightarrow{p} X$ and g a continuous function then $g(X_n) \xrightarrow{p} g(X)$
Consistent estimator: $T_n = T_n(X_1, \dots, X_n)$ converges in probability to $g(\theta)$, a function of the model parameter
Bounded convergence theorem: $Z_n \xrightarrow{p} Z_\infty, |Z_n| \le c < \infty \implies EZ_n \xrightarrow{p} EZ_\infty$
Proof starts with $|E(Z_n - Z_\infty)|$ and uses i) triangle inequality, ii) indicator functions for the case when difference is $> \epsilon, < \epsilon$
Dominated convergence theorem: $Z_n \xrightarrow{p} Z_\infty, E\beta < \infty, |Z_n(\omega)| \le \beta(\omega)\forall\omega \implies EZ_n \xrightarrow{p} EZ_\infty$
Fatous Lemma: for $Z_n > 0$, $E\lim_{n\to\infty} Z_n \le \lim_{n\to\infty} EZ_n$
Almost sure convergence: $P(\omega : \lim_{n\to\infty} X_n(\omega) \overset{p}{\to} X_\infty(\omega)) = 1$ where ω is element in set of all sequences • $P(\limsup_{n\to\infty}\{|X_n - X_\infty|\} > \epsilon) = 0$

0.6 Central limit theorem

$$\sqrt{n}\frac{(\overline{X}_n-\mu)}{\sigma} \longrightarrow N(0,1) \Longleftrightarrow \sqrt{n}(\overline{X}_n-\mu) \longrightarrow N(0,\sigma^2)=\sigma N(0,1),\quad X_i+\cdots+X_n=S_n\approx N(ES_n,VarS_n),\text{ the "meta result"}$$

$$\lim_{n\to\infty}P(\frac{S_n}{\sigma\sqrt{n}}\le x)=\Phi(x),\text{ for }S_n=\sum_{i=1}^{n}X_i,\quad X_1,X_2,\ldots,X_n\text{i.i.d. with }E(X_i)=0\text{ (WLOG)},\ Var(X_i)=\sigma^2$$

Proof sketch: Start with $M_{S_n}(t)$, plug in $t/(\sigma\sqrt(n))$, use Taylor expansion to show convergence to MGF of a normal, $e^{\frac{t^2}{2}}$
Monte Carlo: • Sample $Y\in\mathbb{R}^d$ • Compute $X=g(Y)$ • Repeat n times • form $\overline{X}_n$ and use CLT for asymptotic behavior.
Generating random data: with $X=F_X^{-1}(U)$ since $P(F_X^{-1}(U)\le x)=P(F_X(F_X^{-1}(U))\le F_X(x))=P(U\le F_X(x))=F_x(x)$

0.6.1 Delta method

If g is a differentiable function at μ, $\sqrt{n}(g(\bar{X}_n)-g(\mu))\xrightarrow{d}N(0,g'(\mu)^2\sigma^2)=g'(\mu)\sigma N(0,1)$. **Proof sketch:** Start with Taylor expansion $g(\bar{X}_n)\approx g(\mu)+g'(\mu)(\bar{X}_n-\mu)$, rearrange to get $\sqrt{n}(g(\bar{X}_n)-g(\mu))\approx g'(\mu)\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}N(0,g'(\mu)^2\sigma(2))$. **Note:** if we find $g'(\mu)=0$, then repeat this process with the second derivative, $g''(\mu)$.

0.6.2 Convergence in distribution (a.k.a. weak convergence)

Convergence in distribution: $X_n\xrightarrow{d}X$ when $\lim_{n\longrightarrow\infty}F_{X_n}(x)=F_X(x)$ at all continuity points in F_X
Equalities: $X_n\xrightarrow{d}X\Longleftrightarrow Eh(Z_n)\xrightarrow{p}Eh(Z_\infty)\forall h$, bounded/continuous $\Longleftrightarrow\phi_{Z_n}(t)\xrightarrow{p}\phi_{Z_\infty}(t)\forall t$
Confidence intervals: $P(Z_{\alpha/2}\le Z\le Z_{1-\alpha/2})=P(\frac{\hat{\sigma}}{\sqrt{n}}Z_{\alpha/2}\le\bar{X}_n-\mu\le\frac{\hat{\sigma}}{\sqrt{n}}Z_{1-\alpha/2})=P(\mu\in\left[\bar{X}_n\pm\frac{\hat{\sigma}}{\sqrt{n}}Z_{1-\alpha/2}\right])=1-\alpha$ for $\hat{\sigma}\xrightarrow{p}\sigma$
Slutsky's lemma $A_nX_n+B_n\xrightarrow{d}aX+b$ if $\{X_n\}$ sequence, $X_n\xrightarrow{d}X$, $\{A_n\}$ sequence, $A_n\xrightarrow{d}A$, $\{B_n\}$ sequence, $B_n\xrightarrow{p}b$

0.7 Theory of large deviations

Variance reduction: $E_ph(X)$ where $X\sim P=\int_{-\infty}^{\infty}h(x)p(x)dx=\int_{-\infty}^{\infty}h(x)p(x)q(x)/q(x)=E_q[h(x)p(x)/q(x)]$ where $X\sim Q$.
$\sqrt{n}(\overline{X}-E_pX)\xrightarrow{d}N(0,Var_p(X))$, $\sqrt{n}(\overline{\hat{X}}-E_q\hat{X})\xrightarrow{d}N(0,Var_q(\hat{X}))$ where $\hat{X}=Xp(x)/q(X)$
Importance sampling: Choose $h(x)$ to minimize variance. Minimal $H(dx)$ turns out to be the conditional probability of the event happening on event happening: $H^*(dx)=\mathbb{I}\{A\}(x)F(dx)/F(A)$
Exponential tilting: $Ef(X_1,\ldots,X_n)=E_\theta\exp(-\theta S_n+n\psi(\theta))f(X_1,\ldots,X_n)$, where $\psi(\theta)=\log M_x(\theta)$
Large deviations: • Use exponential tilting with $f(X_1,\ldots,X_n)=\mathbb{I}(S_n>an)$: $E\mathbb{I}(S_n>an)=E_\theta[\exp(-\theta S_n+n\psi(\theta))\mathbb{I}(S_n>an)]$ • Choose optimal $\theta^*=\theta(a)$ which satisfies $\psi'(\theta(a))=a$, which guarantees $E_{\theta(a)}X_i=a$ • E.g., Gaussian: $M_X(\theta)=\exp(\mu\theta+\sigma^2\theta^2/2)\Longleftrightarrow\psi(\theta)=\mu\theta+\sigma^2\theta^2/2\Longleftrightarrow\psi'(\theta)=\mu+\sigma^2\theta\leftarrow$ evaluate at $\psi'(\theta)=a$

0.8 Method of moments estimator

• $E(X^k)=g(\theta)$: Calculate moment with MGF, lower moments typically lead to estimators with lower asymptotic variance • $g^{-1}(E(X^k))=\theta$: Invert this expression to create an expression for the parameter(s) in terms of the moment • $\hat{\theta}=g^{-1}(\frac{1}{n}\sum X_i^k)$: Insert the sample moment into this expression, thus obtaining estimates of the parameters in terms of data • $\sqrt{n}(g^{-1}(\frac{1}{n}\sum X_i^k)-\theta)\xrightarrow{d}N(0,f'(E(X_i^k))^2Var(X_i^k)^2)$: Use the delta method • If multiple parameters characterize the distribution, use multiple moments and a system of equations

0.9 Maximum likelihood estimator

• $L(\theta)=\prod_{i=1}^n f(X_i,\theta)$: Construct the likelihood function • $log(L(\theta))=l(\theta)=\sum_{i=1}^n log(f(X_i,\theta))$: Take the log of the likelihood • Find critical points of this function (e.g., $0=\sum_{i=1}^n\frac{d}{d\theta}log(f(X_i,\hat{\theta})))$ and determine that one is a maximum
Approach to constructing MLE when indicators, $\mathbb{I}\{U\}$, are present: Logs of indicators and derivatives of indicators are very difficult to work with • Simplify likelihood function (splitting indicators when possible) • Make an argument for why the function is increasing or decreasing • Determine the value at the bounds of the function
Estimating equations: $E_{\theta_1}g(\theta_2,X_1)=0\Longleftrightarrow\theta_1=\theta_2$

0.10 Fisher Information

Fisher information: $I(\theta)=E_\theta\left[\left(\frac{d}{d\theta}log(\mathcal{L}(X_1,\theta))\right)^2\right]$
Properties: $E_\theta\left[\left(\frac{d}{d\theta}log(\mathcal{L}(X_1,\theta))\right)\right]=0$ • $I(\theta)=Var\left(\frac{d}{d\theta}log(\mathcal{L}(X_1,\theta))\right)$ • $I(\theta)=-E_\theta\left[\frac{d^2}{d\theta^2}log(\mathcal{L}(X_1,\theta))\right]$ • $I_{X_1,X_2}(\theta)=I_{X_1}(\theta)+I_{X_2}(\theta)$ for X_1,X_2 independent (Information increases with larger sample!)
Cramer-Rau bound: $Var(\hat{\theta})=Var(T(X_1,\ldots,X_n))\ge 1/I_{X_1,\ldots,X_n}(\theta)$ where $I_{X_1,\ldots,X_n}(\theta)=nE_\theta[(\frac{d}{d\theta}\log\mathcal{L}(X_1,\theta))^2]$ for $E(\hat{\theta})=g(\theta)$

0.11 Examples

Newsvendor problem: [Setup] $D\sim Exp(\lambda)$, price p, cost c, buy-back price b, order quantity q, earnings W; [Solution] $W=pq\min(q,D)-cq+b(q-min(q,D))\Longrightarrow EW=(p-b)E(min(q,D))-(c-b)q=(p-b)(xP(D>q)+\int_0^D yf_d(y)dy)-(c-b)q\Longrightarrow\frac{\partial}{\partial q}E(W)|_{q=q^*}$
Cramer-Rao bound: [Setup] $X\sim Exp(\lambda)$, MLE $=1/\overline{X}_n$, does $1/\hat{\lambda}_{MLE}$ achieve lower bound?; [Solution] i) Unbiased: $E(1/\hat{\lambda})=E(\overline{X}_n)=1/\lambda$, ii) $Var(1/\hat{\lambda})=Var(\overline{X}_n)=(1/n^2)nVar(X_1)=1/(\lambda^2n)$, iii) $I(1/\lambda)=I(\theta)=-E_X(\frac{d^2}{d\theta^2}log\mathcal{L}(\theta)):\mathcal{L}(\theta)=1/\theta\exp(-x/\theta)\Rightarrow\log\mathcal{L}(\theta)=-\log\theta-x/\theta\Rightarrow\frac{d^2}{d\theta^2}log\mathcal{L}(\theta)=1/\theta^2-2X/\theta^3\Rightarrow-E_X(\frac{d^2}{d\theta^2}log\mathcal{L}(\theta))=1/\theta^2$