1 STA257 Basics

Set stuff

$$A, B \text{ disjoint} \Leftrightarrow A \cap B = \emptyset$$
$$A \cup B = B \cup A$$
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

Probability Measure:

$$P(\Omega) = 1 \quad (1)$$
$$A \subset \Omega \Rightarrow P(A) \geq 0 \quad (2)$$
$$A_i \text{ mutually disjoint} \Rightarrow P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \quad (3)$$

PMF $p(x_i)$ PDF $f(x_i)$ CDF $F(x_i)$

Conditional Probability

$$P(A|B) = P(A \cap B)/P(B)$$

Law of Total Probability ($\bigcup_{i=1}^{n} B_i = \Omega, B_i$ disjoint, $P(B_i) > 0$)

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Bayes': A, B_i disjoint, $\sum_{i=1}^{n} B_i = \Omega, P(B_i) > 0$, then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Multinomial Coefficient: group n objects into k classes, each of size n_i

$$\binom{n}{n_1, \cdots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

Binomial Theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

2 Distributions

Bernouilli: success (p) or failure ($1-p$) with $p \in [0,1]$. $X \sim Ber(p)$ then

$$p(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

Expectation p Variance $(1-p)p$ MGF $q + pe^t$

Binomial: n $Ber(p)$ trials with k successes. $X \sim Bin(n, p)$ then

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

E np V npq MGF $(1-p+pe^t)^n$

Geometric: k $Ber(p)$ trials until 1 success. $X \sim Geo(p)$ then

$$p(k) = p(1-p)^{k-1}$$

E $\frac{1}{p}$ V $\frac{1-p}{p^2}$ MGF $\frac{pe^t}{1-(1-p)e^t}$

$$\sum_{n=1}^{\infty} az^{n-1} = \sum_{n=0}^{\infty} az^n$$
$$\sum_{n=1}^{k} a_n r^{n-1} = a_n \left(\frac{1-r^k}{1-r}\right)$$

Neg. Binomial: $Ber(p)$ trials conducted until r successes. $X \sim NB(r, p)$ then

$$p(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

E $\frac{r}{p}$ V $\frac{rp}{(1-p)^2}$ MGF $\left(\frac{1-p}{1-pe^t}\right)^2$

Hypergeometric: $X \sim HG(n, m, r)$ where n is the total number of items, m is the number of items sampled, and r is the number of items with a property, then

$$p(k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}} \quad (k \in 0, \cdots, \min(r, m))$$

E $\frac{mr}{n}$

Poisson: # events. $X \sim Poi(\text{rate} = \lambda)$ then

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

E λ V λ MGF $e^{\lambda(e^t - 1)}$

Properties of Poisson:

- For $|S_i| = N_i$ independent $Poi(\lambda)$, $N_i \sim Poi(\lambda|S_i|)$
- For n large, $Bin(n, p) \sim Poi(np)$ can be approximated

Exponential ($Gamma(1, \lambda)$): X waiting time $\sim Exp(\lambda)$, then

$$f(x) = \lambda e^{-\lambda x} \quad F(x) = 1 - e^{-\lambda x}$$

E $\frac{1}{\lambda}$ V $\frac{1}{\lambda^2}$ M $\frac{\lambda}{\lambda - t}$

Gamma (sum a iid $Exp(\lambda)$): $X \sim \Gamma(a, \lambda)$ then

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-x\lambda}$$

E $\frac{a}{\lambda}$ V $\frac{a}{\lambda^2}$ M $\left(\frac{\lambda}{\lambda - t}\right)^a$

Beta: used to model proportions between 0 and 1 with $a, b > 0$ shape parameters. $X \sim Beta(a, b)$ then

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

E $\frac{a}{a+b}$ V $\frac{ab}{(a+b)^2(a+b+1)}$

Normal: $X \sim N(\mu, \sigma^2)$ then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$\text{MGF } e^{\mu t + \frac{\sigma^2 t^2}{2}} \quad f(\mu - x) = f(\mu + x)$$
$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

Std. Normal $Z \sim N(0,1)$ $\phi(z)$ given in table

$X_i \sim N(\mu_i, \sigma_i^2)$ for $i \in \{1, \cdots, n\}$. $Y = \sum_{i=1}^{n} a_i X_i + b$, then

$$Y \sim N\left(\sum_{i=1}^{n} a_i \mu_i + b, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Chi-Square distribution: $U = Z^2$ where $Z \sim N(0,1)$, then $U \sim \chi^2_{df=1}$. If $X \sim \chi^2_{(m)}$ and $Y \sim \chi^2_{(n)}$ then $X + Y \sim \chi^2_{(m+n)}$. If $X \sim \chi^2_{(m)}$ then $E(X) = m$.

t distribution: $Z \sim N(0,1) \perp U \sim \chi^2_{(m)}$ then $\frac{Z}{\sqrt{\frac{U}{m}}} \sim t_{(m)}$, the t distribution with m degrees of freedom

F distribution: $X \sim \chi^2_{(m)} \perp Y \sim \chi^2_{(n)}$ then $\frac{\frac{X}{m}}{\frac{Y}{n}} \sim F(m, n)$.

3 Inequalities, Expectation, Variance, MGFs

Markov's Inequality: $P(X \geq 0) = 1, E(X)$ exists then

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Chebyshev's Inequality: $P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$ (proof: set $Y = (x - \mu)^2$ and apply Markov))

$$\text{r-th moment} = E(X^r)$$
$$\text{r-th central moment} = E[(X - E(X))^r]$$
$$M(t) = E(e^{tX})$$
$$X \perp Y \Rightarrow M_{X+Y} = M_X M_Y$$
$$M_{XY}(s, t) = E(e^{sX+tY})$$
$$M^{(r)}(0) = E(X^r)$$
$$E(X) = \sum_i x_i p(x_i) \quad E(X) = \int_{\mathbb{R}} x f(x) \mathrm{d}x$$
(provided these converge)
$$Y = g(X) \Rightarrow E(Y) = \sum_i g(x_i) p(x_i)$$
$$E(Y) = \int_{\mathbb{R}} g(x) f(x) \mathrm{d}x$$
$$E(aX + b) = aE(X) + E(b) \quad E(XY) = E(X)E(Y) \text{ (if } X \perp Y)$$
$$Var(X) = E[(X - E(X))^2]$$
$$= E(X^2) - (E(X))^2 = \int_{\mathbb{R}} (x - \mu) f(x) \mathrm{d}x$$
$$sd = \sqrt{Var(X)}$$
$$Y = aX + b \Rightarrow Var(Y) = a^2 Var(X)$$

4 Conditional & Multivariate Stuff

Law of Total Expectation

$$E(Y) = E(E(Y|X))$$

Law of Total Variance

$$Var(Y) = Var(E(Y|X)) + E(Var(Y|X))$$
$$p_{xy}(x, y) = p_{X|Y}(x|y) p_y(y)$$
$$p_x(x) = \int_y p_{xy}(x, y) \mathrm{d}y$$
$$E(Y) = \int \cdots \int \underbrace{g(x_1, \cdots, x_n)}_{\text{may do nothing}} f(x_1, \cdots, x_n) \mathrm{d}\{x_i\}$$
$$E(Y|X = x) = \sum_y p_{Y|X}(y|x)$$
$$E(h(Y)|X = x) = \int_y h(Y) f_{Y|X}(y|x) \mathrm{d}y$$

$X_i \perp X_j$ then $Var(\sum x_i) = \sum Var(X_i)$ and $Cov(X_i, X_j) = 0$)

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$
$$= E(XY) - E(X)E(Y)$$
$$Cov(a + X, Y) = Cov(X, Y)$$
$$Cov(aX, bY) = abCov(X, Y)$$
$$Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$$
$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$
$$Cov(aW + bX, cY + dZ) = acCov(W, Y) + bcCov(X, Y) + adCov(W, Z) + bdCov(X, Z)$$
$$Var(a + \sum b_i x_i) = \sum_{i,j} b_i b_j Cov(x_i, x_j)$$

Posterior density:

$$f_{P|X}(p|x) = \frac{f_{X,P}(x,p)}{f_X(x)} = \frac{f_{X|P}(x|p)f_P(p)}{f_X(x)}$$

5 Limit Theorems

Law of Large Numbers: $X_1, X_2, \cdots, X_i, \cdots$ independent and $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Then $\forall \epsilon > 0, P(|\bar{X}_n - \mu| > \epsilon) \overset{n\to\infty}{\longrightarrow} 0$ by Chebyshev's inequality.

Convergence in Distribution: $X_1, \cdots$ are random variables with $F_1, \cdots$, and X has cdf F. $X_n \xrightarrow{D} X$ if $F_n(X) \overset{D}{\to} F(X)$ wherever F is continuous.

- the next outcome (as we get more and more X_is) converge closer and closer to some cdf
- to show converge in distribution, we usually use MGFs. Call $\{F_n\}$ a sequence of cdfs with MGFs $\{M_n\}$.

$$\overbrace{M_N(t) \to M(t)}^{\forall t \in I \text{ s.t. } 0 \in t} \Rightarrow F_n(x) \to F(x)$$

since the MGF uniquely determines the distribution of a RV.

Central Limit Theorem: $X_1, \cdots$ iid with mean μ, variance σ^2, cdf F, MGF M defined in a neighbourhood of 0. Let $S_n = \sum_{i=1}^n X_i$. Then $\bar{X}_n \overset{D}{\to} N(\mu, \frac{\sigma^2}{n})$ or $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{D}{\to} N(0,1)$ or $P\Big(\frac{S_n}{\sigma\sqrt{x}} \le x\Big) \to \phi(x)$.

Proof: $M_{S_n}(t) = (M_x(t))^n, M_{Z_n}(t) = \Big(M_x\Big(\frac{t}{\sigma\sqrt{x}}\Big)\Big)^n, M_X(s) = M_X(0) + SM'(0) + \frac{S^2}{2}M''(0) + \epsilon_S$ with $\frac{\epsilon_S}{S^2} \to 0$. This equals $1 + \frac{1}{2}\sigma^2 + \Big(\frac{t}{\sigma\sqrt{x}}\Big))^2 + \epsilon_n$, so $M_{Z_n}(t) = \Big(1 + \frac{t^2}{2n} + \epsilon_n\Big)^n \to e^{\frac{t^2}{2}}$ hence $Z_n \sim N(0,1)$.

6 Definitions

- Population: a collection of all the subjects that have something in common.

→ Parameter: a characteristic/summary of the population, represented by θ. Can be mean (μ), std. dev (σ), etc.

- Sample: a subset of the population. We use the sample to make an inference about the unknown parameters of our population.

→ Statistic: any summary of the sample; since statistics/estimators are a function of sample observations, we use T to represent them. Examples: sample total ($\sum X_i$), sample mean ($\bar{X}$), etc.

parameter	estimator	estimate
μ	$\bar{X} = \frac{\sum X}{n}$	$\bar{x} = \frac{\sum x}{n}$
σ	S	s

7 Method of Moments

$X_1, \cdots, X_n$ iid RVs. Define the k-th population moment to be $\mu_k = E(X^k)$ and the k-th sample moment to be $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. We use $\hat{\mu}_k$ as an estimator of μ_k using 3 steps:

1. express lower order population moments in terms of the parameters
2. invert the expressions to express the parameters in terms of the population moments
3. replace the population moments using the sample moments

8 Likelihood

$X_1, \cdots, X_n$ RVs with joint density/mass function $f(x_1, \cdots, x_n|\theta)$. Given a sample $(x_1, \cdots, x_n)$, the likelihood function of θ is defined as

$$L(\theta) := L(\theta|x_1, \cdots, x_n) = f(x_1, \cdots, x_n|\theta)$$

where the likelihood is intuitively the probabilility of the parameter being some value given the sample data. If $X_1, \cdots, X_n$ are iid, then we can express the joint as the product of the marginal densities, i.e.

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

Suppose we have θ with likelihood function $L(\theta)$. The best point estimate can be found by picking a $\hat{\theta}$ that maximizes $L(\theta)$, i.e. $\hat{\theta}$ satisfies $L(\hat{\theta}) \ge L(\theta) \quad \forall \theta \in \Omega$.

Usually, we compute the MLE by optimizing the log-likelihood $\ell(\theta)$ ($\ln x$ is one-to-one and increasing). Solve $\frac{\partial \ell(\theta)}{\partial \theta} = 0$ for θ and check that $\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} < 0$.

Invariance property: suppose $\hat{\theta}$ is the MLE of θ and let $\psi(\theta)$ be any 1-1 function of θ defined on Ω, then $\psi(\hat{\theta})$ is the MLE of $\psi(\theta)$.

Fisher Information

$$I(\theta) = E\Big[\frac{\partial}{\partial\theta}\ln f(X|\theta)\Big]^2$$

also, if f is sufficiently smooth (in order to bring operation in integration when finding expectation),

$$= -E\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\Big]$$

The large sample distribution of a MLE is approximately normal with mean θ_0 and variance $\frac{1}{nI(\theta_0)}$. Moreover, the asymptotic variance is given by

$$\underbrace{Var(\hat{\theta}) \ge \frac{1}{nI(\theta_0)}}_{\text{Cramér-Rao Bound}} = -\frac{1}{E\ell''(\theta_0)}$$

Proof: consider the correlation coefficient ρ between two variables Y, Z, which is bounded between -1 and 1. Then

$$\rho^2(T, S()\theta)) \le 1$$
$$\frac{(\text{cov}(T, S(\theta)))^2}{\text{var}(T)\text{var}(S(\theta))} \le 1$$
$$\Rightarrow \text{var(T)} \ge \frac{(\text{cov}(T, S(\theta)))^2}{\text{var}(S(\theta))}$$

9 Mean Squared Error & Bias

Let θ be a parameter, $\psi(\theta)$ be a real-valued function, and T be an estimator of $\psi(\theta)$. The Mean Squared Error is defined as

$$\begin{aligned} MSE_\theta(T) &= E_\theta[(T - \psi(\theta))^2 \\ &= Var_\theta(T) + (E_\theta(T) - \psi(\theta))^2 \quad (*) \\ &= Var_\theta(T) + (Bias(T))^2 \end{aligned}$$

Proof (*): Add $-E(T) + E(T)$ to inner term in definition of MSE, expand using squares.

$$\text{Bias} := E_\theta(T) - \psi(\theta)$$

When the bias of an estimator is 0, it is unbiased.

10 Quiz 1 Problems

a) (Rice E8Q4) Suppose X is a discrete random variable with $P(X = 0,1,2,3) = \frac{2}{3}\theta, \frac{1}{3}\theta, \frac{2}{3}(1-\theta), \frac{1}{3}(1-\theta)$ respectively where $\theta \in [0,1]$ and 10 observations were taken: $(3,0,2,1,3,2,1,0,2,1)$.

Method of moments estimate of θ: $E(X) = \sum_{k=0}^3 kP(X=k) = \frac{\theta}{3} + \frac{4}{3}(1-\theta) + (1-\theta) = \frac{7}{3} - 2\theta$. We rearrange for θ and write the sample mean $\bar{X}$ in place of $E(X)$: $\hat{\theta} = \frac{7}{6} - \frac{1}{2}\bar{X}$ which yields $\hat{\theta} = 0.417$.

Standard error: we need to calculate the variance of X ($Var(X) = E(X^2) - (E(X))^2$). The process yields $E(X^2) = \frac{17-16\theta}{3}$ so $Var(X) = -4\theta^2 + 4\theta + \frac{2}{9}$. $Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}Var(X_i) = \frac{1}{n}Var(X_1)$.

$Var(\hat{\theta}) = \frac{1}{4}Var(\bar{X}) = -\frac{1}{10}\theta^2 + \frac{1}{10}\theta + \frac{1}{180}$. We replace θ with $\hat{\theta} = 0.417$, yielding that $s^2_{\hat{\theta}} = 0.0299$ and the standard deviation is simply the square of this.

b) (Rice E8Q19) Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$. MLE of each of σ, μ, with the other one known:

$L(\theta) = \frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\left(\sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2\right)}$. Log likelihood: $\ell(\theta) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$ and then maximize this function for each parameter. We get that $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2}$ and $\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n}$.

c) (Rice E8Q7) Suppose $X \sim Geo(p)$ and we have an iid sample of size n. Method of moments estimate of p: we are looking to express p in terms of the moments; we have that $E(X) = p$ so $p = \frac{1}{E(X)}$. Hence, $\hat{p} = \frac{1}{\bar{X}}$.

MLE of p: $L(p) = p^n(1-p)^{\sum_{i=1}^n (x_i - 1)}$, $\ell(p) = n\ln(p) + \sum_{i=1}^n (x_i - 1)\ln(1-p)$ which we can use to maximize p: $p = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}}$, obtaining that $\bar{p} = \frac{1}{\bar{X}}$.

Variance of our MLE: $Var(\hat{p}) = \frac{-1}{E(\ell''(p))}$ asymptotically and we have that $E(\ell''(p)) = E\left(-\frac{n}{p^2} - \left[\sum_{i=1}^n X_i - n\right]\frac{1}{(1-p)^2}\right) = -\frac{n}{p^2} - \left(\sum_{i=1}^n E(X_i) - n\right)\frac{1}{(1-p)^2} = -\frac{n}{p^2(1-p)}$. Hence $Var(\hat{p}) \approx \frac{p^2(1-p)}{n}$.

If p has a uniform prior distribution on $[0,1]$, the posterior distribution of p is $f_{P|X}(p|x) = \frac{f_{X|P}(x|p)f_p(p)}{f_X(x)}$. $f_{X|P}(x|p)$ is simply the likelihood function, and $f_X(x)$ can be computed: $f_X(x) = \int_{\mathbb{R}} f_{X|P}(x|p)f_P(p)\text{d}p = \int_0^1 f_{X|P}(x|p)\text{d}p = \int_0^1 p^n(1-p)^{\sum_{i=1}^n (x_i - 1)}\text{d}p$. Relating this to a Beta distribution, we find a value for $f_X(x)$ so we plug it into $F_{P|X}(p|x)$.

The expected value of our posterior distribution is given by $E(P|X) = \frac{n}{n + \sum_{i=1}^n x_i - n}$ (again, using the mean of a Beta distribution).

11 Types of Convergence

- Convergence in distribution: $Z_n \xrightarrow{d} Z$ if $\lim_{n\to\infty} F_n(x) = F(x)$ everywhere F is continuous
- Convergence in probability: $Z_n \xrightarrow{P} Z$ if $\lim_{n\to\infty} P(|Z_n - Z)| > \epsilon) = 0$
- Almost-surely convergence: $Z_n \xrightarrow{a.s.} Z$ if $P(\lim_{n\to\infty} Z_n = Z) = 1$

12 Unbiased Estimator for σ^2

We have that the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

and we have may want to estimate it using $\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ or $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. The first is unbiased, and the other one (given by both the method of moments & MLE estimation) is biased. Proof:

$$\begin{aligned}\sum_i (X_i - \mu)^2 &= \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_i \big[(X_i - \bar{X})^2 + (\bar{X} - \mu)^2 \\ &\quad + 2(X_i - \bar{X})(\bar{X} - \mu)\big] \\ &= n\hat{\sigma}^2 + n(\bar{X} - \mu)^2 + 0(\bar{X} - \mu) \\ &= \sum_i (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\end{aligned}$$

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Take expectation on both sides:

$$\begin{aligned}E[\cdots] &= \sum_i Var(X_i) - nVar(\bar{X}) \\ &= \sum_i \sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 \\ &= (n-1)\sigma^2\end{aligned}$$

Another method for the normal distribution: it can be shown that $\bar{X}$ and S^2 are independent, and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$. Starting with the equation before we took the expectation, we can divide it by σ^2 to obtain

$$\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

where $\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$ and $\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_{(1)}$. So we can use MGFs and the independence of $\bar{X}$ and S^2 to find that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$. Since the expectation of a Chi Square variable is just its degrees of freedom we have that $E(S^2) = \sigma^2$ so S^2 is an unbiased estimator of σ^2 under the normal distribution.

13 Sufficient Statistic

Roughly, a sufficient statistic for a parameter is a summary of a sample which yields the same information about the parameter as the entire sample (data reduction). Formally, a statistic $T(X_1, \cdots, X_n)$ is sufficient for θ if the conditional distribution of $X_1, \cdots, X_n$ given $T = t$ doesn't depend on θ.

Factorization Theorem: $T(X_1, \cdots, X_n)$ is sufficient for θ if the joint probability function factors in the form

$$\begin{aligned}&f(x_1, \cdots, x_n|\theta) \\ &= g(T(x_1, \cdots, x_n), \theta)h(x_1, \cdots, x_n)\end{aligned}$$

where h is a function of sample observations and g involves θ and the sufficient statistic T.

It is also useful to identify distributions in the exponential family of distributions. We can express the density or frequency functions in the form

$$f(x|\theta) = \begin{cases} e^{c(\theta)T(x) + d(\theta) + S(x)}, & x \in A \\ 0, & x \notin A \end{cases}$$

Such distributions include the normal, exponential, log-normal, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, geometric, inverse Gaussian, von Mises and von Mises-Fisher distributions. Some distributions are exponential families only if some of their parameters are held fixed. The family of Pareto distributions with a fixed minimum bound x_m form an exponential family. The families of binomial and multinomial distributions with fixed number of trials n but unknown probability parameter(s) are exponential families. The family of negative binomial distributions with fixed number of failures (a.k.a. stopping-time parameter) r is an exponential family. However, when any of the above-mentioned fixed parameters are allowed to vary, the resulting family is not an exponential family. As mentioned above, as a general rule, the support of an exponential family must remain the same across all parameter settings in the family.

14 Consistent Estimators

In this course, consider only estimators being consistent in probability.

Immediate results: the LLN tells use that $\bar{X} = \frac{1}{n}\sum X_i \xrightarrow{P} E(X_i)$ for any distribution, so $\bar{X}$ is a consistent estimator of μ for $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $\bar{X}$ is a consistent estimator of λ for $X_i \overset{iid}{\sim} Poi(\lambda)$, etc.

Slutsky's lemma: X_n, Y_n are sequences.

$$X_n \xrightarrow{P} X, Y_n \xrightarrow{P} Y \Rightarrow X_n + Y_n \xrightarrow{P} X + Y$$
$$X_n \xrightarrow{P} X, Y_n \xrightarrow{P} Y \Rightarrow X_n Y_n \xrightarrow{P} XY$$

Continuous mapping theorem: $X_n \xrightarrow{P} X$ and g continuous. Then $g(X_n) \xrightarrow{P} g(X)$

MSE consistent: an estimator T_n is MSE consistent if $MSE(T_n) \xrightarrow{n\to\infty} 0$

We have that the MLE is consistent. Let θ_0 represent the true value of the parameter which produced the data (unknown, constant) let $X_i \overset{iid}{\sim} f(x|\theta_0)$, and let $\hat{\theta}$ be the MLE. Proof that $\hat{\theta} \xrightarrow{P} \theta_0$: start with the log likelihood $\ell(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$. Divide both sides by the sample size n and apply LLN.

Since $\frac{1}{n}\ell(\theta)$ gets closer to $E(\log f(X_i|\theta))$, the θ that maximizes $\frac{1}{n}\ell(\theta)$ should be closed to the θ that maximizes $E(\log f(X_i|\theta))$. We can show that $E(\log f(X_i|\theta))$ is maximized at θ_0.

15 Score and Fisher Information

The score function is the derivative of the log likelihood,

$$S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \ell'(\theta)$$

and the solution to the score equation

$$S(\theta) = 0$$

is the MLE. We say that $S(\theta)|_{\theta=\hat{\theta}} = 0$.

The score as a random variable for iid X_i has that

$$\begin{aligned}S(\theta|X_i) &= \frac{\partial}{\partial\theta}\sum_i \log f(X_i|\theta) \\ &= \sum_i \frac{\partial}{\partial\theta}\log f(X_i|\theta) = \sum_i S(\theta|X_i)\end{aligned}$$

They key takeaway is that $S(\theta)$ is random; for a different set of observations, the likelihood function varies.

The Fisher Information is defined by

$$\begin{aligned}I(\theta_0) &= Var\left(S(\theta|X)|_{\theta=\theta_0}\right) \\ &= E\left(\frac{\partial}{\partial\theta}\log f(X|\theta)|_{\theta=\theta_0}\right)^2 \\ &= -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)|_{\theta=\theta_0}\right)\end{aligned}$$

and is the amount of information that each observable random variable X contains about θ.

16 Quiz 2 Problems

a) (W4 Notes) Suppose $X_1, X_2, X_3 \overset{iid}{\sim} Poi(\lambda)$. Verify that $T = \sum_{i=1}^3 X_i$ is a sufficient statistic for λ.

$$\begin{aligned}&P(X_1 = x_1, \cdots, X_3 = x_3|\sum x_i = t) \\ &= \frac{P(X_1 = x_1, \cdots, X_3 = x_3, \sum x_i = t)}{P(\sum x_i = t)} \\ &= \frac{P(X_1 = x_1, \cdots, X_3 = t - (x_1 + x_2))}{P(\sum x_i = t)} \\ &= \frac{\frac{\cancel{e^{-\lambda}}\lambda^{x_1}}{x_1!}\frac{\cancel{e^{-\lambda}}\lambda^{x_2}}{x_2!}\frac{\cancel{e^{-\lambda}}\lambda^{t-(x_1+x_2)}}{(t-(x_1+x_2))!}}{\frac{\cancel{e^{-3\lambda}}(3\lambda)^t}{t!}} \\ &= \frac{\cancel{\lambda^t}}{x_1!x_2!}\frac{t!}{(3\cancel{\lambda})^t} = \frac{t!}{x_1!x_2!3^t}\end{aligned}$$

which doesn't depend on λ.

b) (W4 Notes) Factorization theorem for $Poi(\lambda)$:

$$\begin{aligned}L(\lambda) &= e^{n\lambda}\lambda^{\sum x_i}\frac{1}{\prod x_i!} \\ &= g\left(\sum_{i=1}^n x_i, \lambda\right)h(x_1, \cdots, x_n)\end{aligned}$$

so $T = \sum x_i$ is a sufficient statistic for λ.

c) (W4 Notes) Score & Fisher Information using Poisson Distribution

Let $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$ with λ_0 parameter.

$$\begin{aligned}L(\lambda) &= \prod f(X_i|\theta) \\ &= \prod \frac{e^{-\lambda}\lambda^{\sum X_i}}{\prod X_i!} \\ \ell(\lambda) &= -n\lambda + (\sum X_i)\ln\lambda - \sum \ln(X_i)! \\ \frac{\partial}{\partial\lambda}\ell(\lambda) &= -n + \frac{\sum X_i}{\lambda} \\ I(\lambda_0) &= Var(S(\lambda|X)|_{\lambda=\lambda_0}) \\ &= Var\left(-1 + \frac{X}{\lambda_0}\right) \\ &= \frac{1}{\lambda_0^2}Var(X) = \frac{1}{\lambda_0}\end{aligned}$$

17 Efficiency

For two estimators T_1, T_2 of a statistic θ, the efficiency of T_1 relative to T_2 is

$$\text{eff}(T_1, T_2) = \frac{\text{var}(T_2)}{\text{var}(T_1)}$$

A more efficient estimator has a smaller variance, and the most efficient unbiased estimator achieves the Cramèr-Rao Lower Bound.

18 Distribution of MLE

Under some conditions, as $n \to \infty$,

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\frac{1}{n} I(\theta_0)}} \xrightarrow{D} N(0,1)$$

As a consequence, for large n, $E(\hat{\theta}) = \theta_0$ and $V(\hat{\theta}) = \frac{1}{nI(\theta_0)}$.

Proof: we have that $\ell'(\hat{\theta}) = 0, E(\ell'(\theta_0)) = 0), \ell'(\theta_0) \xrightarrow{D} N(0, nI(\theta_0))$ so $\frac{1}{n}\ell'(I(\theta_0)) \xrightarrow{D} N(0, \frac{1}{n} I(\theta_0))$.

Additionally, by LLN, $\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} -I(\theta_0)$.

We then use the Taylor series expansion of $\ell'(\hat{\theta})$ around θ_0 to find that

$$\hat{\theta} - \theta_0 \approx -\frac{\ell'(\theta_0)}{\ell''(\theta_0)} = \frac{\frac{1}{n}\ell'(\theta_0)}{-\frac{1}{n}\ell''(\theta_0)}$$

And with this we obtain the result. Additionally, the MLE is: asymptotically unbiased, a function of sufficient statistics, consistent, and asymptotically efficient.

19 Confidence Intervals

Some sampling distributions:

- $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

- Distribution of MLE (above)

γ-Confidence interval: intuitively, some interval which will have at least γ chance of containing the true parameter. $C(X_1, \cdots, X_n) = (l(X_1, \cdots, X_n), u(X_1, \cdots, X_n))$ is a γ confidence interval for ψ if $P(l() \leq \psi(\theta) \leq u) \geq \gamma$ for every $\theta \in \Omega$.

Pivotal Quantity: a random variable defined in terms of the sample; involves the unknown parameters in its expression, but the distribution of this quantity doesn't depend on the parameters.

For each of the sampling distributions above, we can use the z, t, or χ^2 distributions to calculate the necessary quantiles which correspond to γ. We use the standard normal for μ when σ^2 is known, the t distribution for μ when σ^2 is unknown, and the χ^2 distribution for σ^2.

For two-sided confidence intervals, notice that the Z, t distributions are symmetric about their means so the shortest interval is also symmetric about the mean. For the chi^2 distribution, we use $\frac{1 \pm \gamma}{2}$ even though it may not yield the shortest interval (its shape is determined by the degrees of freedom).

For MLE-based confidence intervals, the confidence interval is given by

$$\hat{\theta} \pm z_{(\frac{1+\gamma}{2})} \sqrt{\frac{1}{nI(\theta_0)}}$$

where $\hat{\theta}$ is the MLE and we use either a plug-in estimate of the Fisher Information (replace θ_0 by $\hat{\theta}$) or the observed Fisher Information (replace θ_0 with summary of observed data).

For one-sided confidence intervals, we consider one side of each distribution.

Interpretation:

- For z, t intervals, the sample mean is the midpoint of the lower and upper bound. The width of the interval is given by $u - l$, and the margin of error is half of the width.
- The width of the interval increases as the confidence level increases or the standard deviation (σ, s) increases
- The width of the interval decreases as the sample size increases
- CIs are not fixed numbers but rather random variables; "there is a 95% chance that μ is between 4.442 and 5.318 is incorrect." Rather, we should say, "if we keep taking samples and construct 0.95-CIs, 95% of the confidence intervals will capture the true value of the parameter."