

# Statistics Oral Exam Notes 2008

Chaitanya Ekanadham

## 1 Sufficient statistics

### 1.1 Definitions/facts

1. **Observations/data:**  $X$  a (typically  $\mathbb{R}^n$ -valued) random variable on the space  $(\Omega, \mathcal{F})$ . The difference from classical random variables is that the probability measure  $P$  is not known, although it is typically assumed to be in some family of measures on  $(\Omega, \mathcal{F})$ . The  $\sigma$ -field generated by  $X$  is  $\sigma(X) = \{X^{-1}(A) \in \mathcal{F} : A \in \mathbb{B}^n\}$ .
2. **parameter space:** Given random observations  $X$  on  $(\Omega, \mathcal{F})$ , we define a parametric space  $\Theta$  (typically  $\mathbb{R}^d$ ) where each  $\theta \in \Theta$  specifies a probability measure on  $(\Omega, \mathcal{F})$  or equivalently a measure  $P$  on  $\mathbb{R}^n$  induced by  $X$ . We define the parametric family of induced measures to be  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ . In most cases, every  $P \in \mathcal{P}$  is absolutely continuous w.r.t. a  $\sigma$ -finite measure  $\nu$  on  $(\mathbb{R}^n, \mathcal{B}^n)$ , and so  $\mathcal{P}$  can equivalently be specified by the densities  $\{\rho_\theta(x) = \frac{dP_\theta}{d\nu}(x) : \theta \in \Theta\}$ . We say  $\mathcal{P}$  is *dominated* by  $\nu$ .
3. **statistic:** Any measurable function  $T$  of the observations  $X$ . Typically  $T : \mathbb{R}^n \rightarrow \mathbb{R}^d$  for some  $d \leq n$ . The  $\sigma$ -field generated by  $T$  is  $\sigma(T) = \{T^{-1}(A) \in \mathcal{B}^n : A \in \mathcal{B}^d\}$ .
  - (a) **complete statistic:**  $T(X)$  is complete w.r.t. a family  $\mathcal{P}$  if  $\forall g : \mathbb{R}^d \rightarrow \mathbb{R}$  measurable,  $\mathbb{E}[g(T)] = 0 \forall P \in \mathcal{P} \Rightarrow g = 0$  a.s. w.r.t.  $\mathcal{P}$ .
  - (b) **sufficient statistic:** We define a sub  $\sigma$ -field  $\Sigma \subseteq \mathcal{B}^n$  to be sufficient for  $\mathcal{P}$  (or equivalently,  $\Theta$ ) if  $\forall A \in \mathcal{B}^n, \exists f_A \Sigma$ -measurable s.t.  $\forall P \in \mathcal{P}, f_A(x) = \mathbb{E}_P[1_A | \Sigma]$  a.s. wrt  $P$  (i.e. the conditional expectations given  $\Sigma$  is independent of the law of  $X, P \in \mathcal{P}$ ). We say  $T(X)$  is sufficient for  $\mathcal{P}$  if its corresponding  $\sigma$ -field is sufficient for  $\mathcal{P}$ .

Equivalently, we can define  $T(X)$  to be sufficient if  $\exists r(A, t) : \mathcal{B}^n \times \mathbb{R} \rightarrow \mathbb{R}$  s.t.  $r(\cdot, t)$  is a measure on  $\mathbb{R}^n$  for each  $t$  and  $r(A, \cdot)$  is a measurable fn of  $t$  for each  $A$ , and  $\mathbb{E}_{P_\theta}[1_B | T = t] = r(B, t)$  a.e.  $P_\theta \forall t, \theta$ . Here  $r(A, t)$  is just the 'common' Radon Nikodym derivative of the  $1_A$  wrt all the  $P_\theta$ 's, evaluated at some  $x \in T^{-1}(t)$ .

- (c) **pairwise sufficient statistic:** We define a sub  $\sigma$ -field  $\Sigma$  to be pairwise sufficient for  $\mathcal{P}$  if  $\forall P, Q \in \mathcal{P}$ , we have that  $R_{Q/P} = \left(\frac{dQ}{d(P+Q)}\right)\left(\frac{dP}{d(P+Q)}\right)$  is  $\Sigma$ -measurable. In otherwords, the likelihood ratio is measurable wrt the subfield. A statistic is pairwise sufficient if its generated subfield is. Note that we can WLOG replace  $P + Q$  with a sigma-finite measure  $\nu$  that dominates  $\mathcal{P}$ , if one exists.
- (d) **minimal sufficient statistic:** is a sufficient statistic  $T(X)$  w.r.t.  $\Theta$  s.t. for any suff. stat.  $S(X)$  w.r.t.  $\Theta$ ,  $T(x) = \psi(S(x))$  a.s. w.r.t. all  $P \in \mathcal{P}$ . In otherwords, a minimal sufficient statistic can be thought of as 'weeding out all unnecessary information' by consecutively applying measurable functions to a statistic while still maintaining sufficiency.

4. **likelihood function:** probability function  $\rho_\theta(x)$  viewed as a function of  $\theta$ .

## 1.2 Useful Theorems

**Theorem 1.** Domination implies domination by a weighted combination of measures in the family. Let  $\mathcal{P}$  be a family of measures on  $(\mathbb{R}^n, \mathcal{B}^n)$  that is dominated by a  $\sigma$ -finite measure  $\nu$ . Then  $\exists (c_j) \geq 0$  and  $(P_j) \in \mathcal{P}$  s.t.  $\sum_j c_j = 1$  and  $\mathcal{P}$  is equivalent by the measure  $Q = \sum_j c_j P_j$ . That is,  $(Q(B) = 0) \Leftrightarrow (P(B) = 0 \forall P \in \mathcal{P})$ .

**Depends on:** Radon-Nikodym derivative, sup constructions

**Proof idea:** 4 steps:

1. Construct a *probability* measure  $\mu'$  dominating  $\mathcal{P}$  by 'normalizing'  $\nu$
2. Construct a prob. measure  $\tilde{\mu}$  that is *equivalent* to  $\mathcal{P}$  by 'modding out' the set  $C$  of maximal  $\mu'$ -measure but for which  $P(C) = 0 \forall P \in \mathcal{P}$ .
3. Construct a sequence of measures  $\{P_j\}_{j \geq 1}$  from  $\mathcal{P}$  that is equivalent to  $\tilde{\mu}$  by a inf sup construction over countable sequences in  $\mathcal{P}$ .
4. Take  $Q = \sum_j 2^{-j} P_j$ .

□

**Theorem 2.** Neyman-Fisher factorization theorem

Let  $X$  be observations on  $(\Omega, \mathcal{F})$  that induces measure  $P_\theta$  on  $(\mathbb{R}^n, \mathcal{B}^n)$  for some  $\theta \in \Theta$ . Suppose  $\Theta$  is dominated by a  $\sigma$ -finite measure  $\nu$  on  $(\mathbb{R}^n, \mathcal{B}^n)$  and let the associated densities be  $\{\rho_\theta(x) : \theta \in \Theta\}$ . Let  $T(X)$  be a statistic. Then:

1.  $T$  is sufficient for  $\Theta \Leftrightarrow$
2.  $\exists h : \mathbb{R}^n \rightarrow \mathbb{R}$  measurable and integrable, and for each  $P_\theta$  a  $\sigma(T)$ -measurable function  $g_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$  s.t.  $\forall \theta \in \Theta, \rho_\theta(x) = g_\theta(x)h(x)$  (a.s. wrt  $P_\theta$ ).  $\Leftrightarrow$
3.  $T$  is pairwise-sufficient, i.e.  $\forall P_{\theta_1}, P_{\theta_2}$ , the function  $R_{\theta_2, \theta_1}(x) \equiv \frac{dP_{\theta_1}}{d\nu}(x) \frac{d\nu}{dP_{\theta_2}}(x)$  is  $\sigma(T)$ -measurable. (In general sufficiency implies pairwise sufficiency even when there is no dominating measure).

**Depends on:** dominating measures, Radon nikodym derivative, properties of conditional expectation

**Proof idea:** Let  $Q = \sum_j c_j P_j$  be the dominating measure of  $\mathcal{P}$  from the previous theorem.

1. (2)  $\Rightarrow$  (3): This is easy since the  $h$ 's cancel in the numerator and denominator, leaving only the ratio  $\sigma(T)$ -measurable  $g$ 's.
2. (3)  $\Rightarrow$  (1): Take the  $Q = \sum 2^{-j} Q_j$  from previous theorem and let  $f_A(x) = \mathbb{E}_Q[1_A | \sigma(T)]$ . Show that  $f_A$  satisfies the desired property by looking at  $\mathbb{E}_{P_\theta}[f_A 1_B]$  for  $B \in \sigma(T)$ , using  $dP_\theta = \frac{dP_\theta}{dQ} dQ$ , and then bringing the first term into the conditional expectation defining  $f_A$ , using properties of conditional expectation, and changing measure of integration back to  $dP_\theta$ .
3. (3)  $\Rightarrow$  (2): Easy - take  $Q$  from previous theorem, and show that  $\frac{dP_\theta}{dQ} = 1/R_{Q/P_\theta} = \sum 2^{-k} R_{Q_k/P_\theta}$  is  $\sigma(T)$ -measurable. Then write  $\frac{dP}{d\nu}$  as a product of radon-nikodym derivatives.
4. (1)  $\Rightarrow$  (2): The basic idea is the same as above use  $g_\theta(x) = \frac{dP_\theta}{dQ}$  and  $h(x) = \frac{dQ}{d\nu}$  where  $Q = \sum 2^{-j} Q_j$  is the dominating (and equivalent) measure derived from the previous theorem, and use sufficiency to show that  $g_\theta(x)$  is  $\sigma(T)$ -measurable. To do this it STS that  $\frac{dP_\theta}{dQ}$  is equal to its own conditional expectation  $\mathbb{E}_Q[\frac{dP_\theta}{dQ} | \sigma(T)]$ . Thus it STS their integral is the same over any measurable set. For  $B \in \mathcal{B}^n$ , this is basically messing around with conditional expectations:

$$\begin{aligned} \mathbb{E}_Q\left[\frac{dP_\theta}{dQ} 1_B\right] &= \mathbb{E}_{P_\theta}[1_B] = \mathbb{E}_{P_\theta}[\mathbb{E}_{P_\theta}[1_B|\sigma(T)]] = \mathbb{E}_{P_\theta}[f_B] = \mathbb{E}_Q[f_B \frac{dP}{dQ}] = \\ &= \mathbb{E}_Q[\mathbb{E}_Q[f_B \frac{dP}{dQ}|\sigma(T)]] = \mathbb{E}_Q[f_B \mathbb{E}_Q[\frac{dP}{dQ}|\sigma(T)]] = \mathbb{E}_{P_\theta}[\mathbb{E}_Q[1_B|\sigma(T)] \mathbb{E}_Q[\frac{dP}{dQ}|\sigma(T)]] = \\ &= \mathbb{E}_Q[1_B \mathbb{E}_Q[\frac{dP}{dQ}|\sigma(T)]]. \end{aligned}$$

5. (1)  $\Rightarrow$  (3) when there is no dominating measure: for any  $P, Q \in \mathcal{P}$ , take the dominating measure as  $(P + Q)/2$  and apply the previous case.

□

### 1.3 Important examples

1. For  $n$  i.i.d. *Bernoulli*( $\theta$ ) trials  $X_j$ , a sufficient statistic is  $T(X) = \sum X_j$ , since  $\rho_\theta(x) = \theta^{\sum x_j} (1 - \theta)^{n - \sum x_j} = g(T(x), \theta)$
2. For i.i.d  $N(\mu, \sigma^2)$  trials  $X_j$ , a sufficient statistic is  $T(X) = (\bar{X}, \sum (X_j - \bar{X})^2)^T$ , since:  
 $\rho_{\mu, \sigma}(x) = (2\pi)^{-n/2} \sigma^{-n} \exp(-\frac{1}{2\sigma^2} \sum (x_j - \bar{x})^2) \exp(-\frac{1}{2\sigma^2} n(\bar{x} - \mu)^2) = g(T(x), \mu, \sigma^2)$
3. The sample mean is also sufficient statistics for i.i.d *Poisson*( $\lambda$ ) trials.
4. For i.i.d *Uniform*( $\theta_1, \theta_2$ ) trials, the min/max trial values are sufficient statistics.

## 2 Estimation theory

### 2.1 Definitions

For a point estimator  $T(X)$  (statistic) of a parameter  $\theta \in \Theta$ , we consider the following properties:

1. **bias:**  $b_T(\theta) = \mathbb{E}[T(X) - \theta] \neq 0$ .
2. **variance:**  $Var[T(X)]$
3. **mean squared error:**  $mse_T(\theta) = \mathbb{E}[(T(X) - \theta)^2] = b_T(\theta)^2 + Var[T(X)]$
4. **consistency:**  $\forall \theta \in \Theta, (T_n(X_n)) \rightarrow \theta$  in probability as  $n \rightarrow \infty$ , where  $T_n(X_n)$  is a statistic of  $n$  samples from  $P_\theta$ .
5. **Fisher Information (Information number):** If  $\mathcal{P}$  is dominated by the Lebesgue measure so that  $P_\theta$  has density  $\rho_\theta(x)$  that is differentiable as a function of  $\theta$ , then  $I : \Theta \rightarrow \mathbb{R}$  is defined by  $I(\theta) = \mathbb{E}_{P_\theta}[(\frac{d}{d\theta} \log \rho_\theta(X))^2]$ . One can also define the *conditional* Fisher information given a statistic  $T$  by replacing  $\rho_\theta(X)$  by  $\rho_\theta(X|T)$ .

6. **Maximum likelihood estimator (MLE):**  $\hat{\theta}_{MLE} \equiv \arg \max_{\theta} \rho_{\theta}(X)$  (may not be unique, or even exist).

## 2.2 Useful theorems

**Theorem 1.** Rao-Blackwell theorem: sufficient statistics can't hurt

Let the statistic  $\hat{\theta} = \hat{\theta}(X)$  be an unbiased estimator for  $\theta \in \Theta$  and let  $T(X)$  be a sufficient statistic. Let  $\varphi(T) = \mathbb{E}_{P_{\theta}}[\hat{\theta}(X)|\mathcal{F}_T]$ . Then:

1. The distribution of  $\varphi(T)$  does not depend on  $\theta$
2.  $\varphi(T)$  is an unbiased estimator for  $\theta$
3.  $Var_{P_{\theta}}[\varphi(T)] \leq Var_{P_{\theta}}[\hat{\theta}]$ .

**Depends on:** factorization theorem, conditional Jensen's inequality

**Proof idea:** The facts that  $\varphi(T)$  is unbiased and does not depend on  $\theta$  follow from iterated conditional expectation and the factorization theorem, respectively. The variance condition follows from Jensen's inequality for conditional expectation:  $Var[\varphi(T)] = \mathbb{E}[\mathbb{E}[\hat{\theta}|\mathcal{F}_T]^2 - \theta^2] \leq \mathbb{E}[\mathbb{E}[\hat{\theta}^2|\mathcal{F}_T] - \theta^2] = \mathbb{E}[\hat{\theta}^2] - \theta^2 = Var[\hat{\theta}]$ .  $\square$

**Theorem 2.** (Alternate form of Fisher information)

Let  $X$  be an  $\mathbb{R}$ -valued r.v. on  $(\Omega, \mathcal{F}, P)$  and let  $\{f_{\theta}(x) : \theta \in \Theta\}$  be a family of densities on  $\mathbb{R}$  corresponding to  $\{\mathcal{P}_{\theta}\}$ . Define  $Y = \frac{d}{d\theta} \log(f_{\theta}(X))$ . Suppose that:  $\frac{d}{d\theta} \int (\cdot) dP = \int \frac{d}{d\theta} (\cdot) dP$ . Then  $\mathbb{E}_{\mathcal{P}_{\theta}}[Y] = 0$  and  $I(\theta) \equiv \mathbb{E}_{\mathcal{P}_{\theta}}[Y^2] = -\mathbb{E}_{\mathcal{P}_{\theta}}[\frac{d^2}{d\theta^2} \log f_{\theta}(X)]$ .

**Depends on:** differentiation under the integral sign, manipulations

**Proof idea:** 1. To show  $\mathbb{E}[Y] = 0$ , write LHS as an integral w.r.t.  $f_{\theta}(x)dx$ .

Take the derivative of the log inside the integral and let the densities cancel. Then bring  $\frac{d}{d\theta}$  outside the integral, which evaluates to 1 regardless of  $\theta$ .

2. Write  $\frac{d^2}{d\theta^2} \log f = \frac{\frac{d^2}{d\theta^2} f}{f} - \{\frac{df}{d\theta}\}^2$ . The second integrates to what we want. The first integrates to 0 if we pull out  $\frac{d^2}{d\theta^2}$ .  $\square$

**Remark:**

1. The original interpretation of  $I(\theta)$  is the amount of information the r.v.  $X$  has about the parameter  $\theta$ . One can think of this alternative interpretation as the 'expected curvature' of the likelihood function  $f_\theta(X)$ , as a function of  $\theta$ . This condition does not hold in general. Consider the case where  $f_\theta(x) = 1/\theta$  is the uniform density on  $[0, \theta]$ .
2. The multivariate form is:  $I(\theta) = \mathbb{E}[(\nabla_\theta \log f_\theta(X))(\nabla_\theta \log f_\theta(X))^T] = -\mathbb{E}[\nabla_{\theta\theta}^2 \log f_\theta(X)]$

**Theorem 3.** (Fisher information is additive and is decreased for a statistic) Suppose  $X$  and  $Y$  are i.i.d r.v.'s on  $(\Omega, \mathcal{F}, P)$ . Then  $I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta)$ . Let  $T(X)$  be a statistic of the data  $X$  with density  $f_\theta(X)$ . Then  $I_{T(X)}(\theta) \leq I_X(\theta)$  with equality iff  $T(X)$  is sufficient. (In multivariate terms,  $I_X(\theta) - I_{T(X)}(\theta)$  is positive definite  $\forall \theta \in \Theta$ .)

**Depends on:** basic properties

**Proof idea:** Showing that Fisher information is additive over i.i.d.'s observations follows directly from the joint-product distribution turning into a sum with log. To show the other part, we define a family of 'joint' distribution on  $(X, T)$ ,  $Q_\theta(C) = P_\theta(\{x : (x, T(x)) \in C\})$  for  $C \in \mathbb{R}^n \times \mathbb{R}^d$ . Taking Radon-Nikodym derivative w.r.t dominating measure  $\nu$  we get a joint distribution  $f_{X,T|\theta}(x, t) = f_{X|\theta}(x) = f_{T|\theta}(t) f_{X|T,\theta}(x)$ . Taking the log and  $\frac{d}{d\theta}$  gives the result. The last term is 0 iff  $\log f_{X|T,\theta}(x)$  does not depend on  $\theta$ , i.e. iff  $T$  is sufficient. □

**Theorem 4.** Cramer-Rao lower bound for unbiased estimators

Let  $\{X_j\}_{j=1}^n$  be  $n$  i.i.d. samples of a random variable  $X$  with density in the family  $\{f_\theta(x)\}$ . Suppose  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  is an unbiased estimator of  $\theta \in \Theta$ , and suppose also we can differentiate under the integral sign as in the previous theorem. Then  $Var[\hat{\theta}] = \frac{1}{nI(\theta)}$ .

**Depends on:** Previous theorem, Cauchy-Schwartz inequality

**Proof idea:** Start with  $n = 1$  and the fact that  $\frac{d}{d\theta} \mathbb{E}_{\mathcal{P}_\theta}[\hat{\theta}] = 1$ . Take  $\frac{d}{d\theta}$  inside the expectation (integral) and rewrite  $\frac{df}{d\theta} = \frac{d}{d\theta} \log f \cdot f$  to get that:  $\mathbb{E}[\hat{\theta}Y] = \mathbb{E}[(\hat{\theta} - \theta)Y] = 1 \Rightarrow 1 \leq \mathbb{E}[(\hat{\theta} - \theta)^2] \mathbb{E}[Y^2] = Var[\hat{\theta}] I(\theta)$ . For  $n > 1$ , take  $f$  as the product density and use the same calculations. □

**Theorem 5.** Slutsky's theorem

Let  $\{X_n\}_{n \geq 1}, \{Z_n\}_{n \geq 1}$ , and  $X$  be r.v.'s on  $(\Omega, \mathcal{F}, P)$ . If  $X_n \Rightarrow X$  in distribution and  $Z_n \rightarrow 0$  in probability, then  $(X_n + Z_n) \Rightarrow X$  in distribution.

**Depends on:** distribution functions, continuity points, inclusion of events

**Proof idea:** STS pointwise convergence  $P(X_n + Z_n \leq x) \rightarrow F_X(x)$  for all continuity points  $x$  of  $F_X$ . Show  $\leq$  and  $\geq$  separately by conditioning on the event  $|Z_n| \leq \epsilon$  and taking  $\epsilon \rightarrow 0$  using continuity of  $F$ :

1.  $\leq$ :  $\limsup P(X_n + Z_n \leq x) \leq \limsup P(X_n - |Z_n| \leq x) \leq \limsup P(X_n \leq x + \epsilon) + P(|Z_n| \geq \epsilon) = F(x)$
2.  $\geq$ :  $\liminf P(X_n + Z_n \leq x) \geq \liminf P(X_n + |Z_n| \leq x) \geq \liminf P(X_n \leq x - \epsilon, |Z_n| < \epsilon) = \liminf P(Z_n \leq x - \epsilon) - P(X_n \leq x - \epsilon, |Z_n| \geq \epsilon) = F(x)$

**Remark:** This can be extended to the case where  $Z_n \rightarrow c$  in probability. The  $(X_n + Z_n) \Rightarrow X + c$  in distribution.  $\square$

**Theorem 6.** ('Asymptotic' consistency of the MLE)

Let  $X$  be an observation with distribution  $P_{\theta_0}$  for some  $\theta_0 \in \Theta$ . Suppose the following assumptions hold:

1.  $S_{\theta}(X) = \frac{\rho_{\theta}(X)}{\rho_{\theta_0}(X)}$  is well-defined nonconstant random variable on  $(\Omega, \mathcal{F})$ .
2.  $\theta_0$  is an interior point of  $\Theta$
3.  $f_{\theta}(x)$  is continuous in  $\theta$ .
4.  $\log S_{\theta}(X)$  has finite first moment  $\forall \theta \in \Theta$ .

Then  $\forall \epsilon > 0, P_{\theta_0}(\exists \text{ a local max of } \rho_{\theta}(X) \in B(\theta_0, \epsilon)) \rightarrow 1$ .

**Depends on:** WLLN, Jensen's inequality

**Proof idea:** Consider the sequence of indep. r.v.'s  $Y_j = \log \frac{\rho_{\theta}(X_j)}{\rho_{\theta_0}(X_j)}$ . Apply

the WLLN to show that the average  $\frac{1}{n} \sum_{j=1}^n Y_j \rightarrow \mathbb{E}_{P_{\theta_0}}[Y_1]$  in probability, Apply

Jensen's inequality to this limit and manipulate to get that  $\mathbb{E}_{P_{\theta_0}}[Y_1] < 0$ . Thus the sequence has a negative limit in probability. Taking the exponential gives that  $P_{\theta_0}(\rho_{\theta}(X) < \rho_{\theta_0}(X)) \rightarrow 1$ . Using continuity of the likelihood function gives the result.  $\square$

**Theorem 7.** (Asymptotic normality of the MLE with minimum variance)  
 Let  $X$  be an observation with distribution  $P_{\theta_0}$  for some  $\theta_0 \in \Theta$ . Suppose the assumptions of the previous theorem as well as those for the Cramer-Rao inequality hold. Suppose also that the MLE  $\hat{\theta}_n$  is uniquely defined  $\forall n$  (i.e. the likelihood function has unique maximum). Then  $\hat{\theta}_n \Rightarrow N(\theta_0, nI^{-1}(\theta_0))$  in distribution where  $I(\theta)$  is the Fisher information matrix.

**Depends on:** CLT, WLLN, Slutsky, Taylor expansions

**Proof idea:** Assume  $n = 1$  - generalization to  $n$  dimensions is straightforward. Consider the i.i.d  $\theta$ -dependent r.v.'s  $S_n(\theta) = \frac{1}{\sqrt{nI(\theta_0)}} \sum_{j=1}^n \frac{d}{d\theta} \log \rho_{\theta}(X_j)$ .

1. Expand  $S_n(\theta)$  about  $\theta = \theta_0$  to order  $o|\theta - \theta_0|$ . Evaluate this expansion at  $\hat{\theta}_{MLE}$  (the LHS must be 0 by def. of MLE).
2. In the first term on RHS,  $S_n(\theta_0)$ , apply CLT to get the limiting standard normal dbn.
3. Factor the second term on RHS as  $\left\{ \frac{1}{nI(\theta_0)} \sum_{j=1}^n \frac{d^2}{d\theta^2} \log \rho_{\theta}(X_j) \Big|_{\theta=\theta_0} \right\} \left\{ \sqrt{nI(\theta_0)} (\hat{\theta}_n - \theta_0) \right\}$ . BY WLLN to the first braced term goes to  $-1$  in probability.
4. Apply Slutsky's theorem to get the result.

□

## 2.3 Important examples

1. Fisher information for various distributions:
  - (a) Binomial( $\theta$ ):  $I(\theta) = \frac{1}{\theta(1-\theta)}$
  - (b) Poisson( $\lambda$ ):  $I(\lambda) = 1/\lambda$
  - (c) Normal( $\mu, \sigma^2$  known):  $I(\mu) = 1/\sigma^2$
2. Unbiased estimator for the uniform density is  $\frac{n+1}{n} \max\{X_1, \dots, X_n\}$ . One can derive the CDF easily and get the density. The variance is  $O(1/n^2)$ .
3. MLE for various distributions:
  - (a) Normal( $\mu, \sigma^2$  known):  $\hat{\mu} = \bar{X}$

- (b) Normal( $\mu$  known,  $\sigma^2$ ):  $\hat{\sigma}^2 = \frac{1}{n} \sum (\mu - X_i)^2$
- (c) Normal( $\mu, \sigma^2$ ):  $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \frac{1}{n} \sum (X_i - \bar{X})^2)$  (biased!)
- (d) Poisson( $\lambda$ ):  $\hat{\lambda} = \bar{X}$
- (e) Binomial( $\theta$ ):  $\hat{\theta} = \bar{X}$
- (f) Unif( $0, \theta$ ):  $\hat{\theta} = \max\{X_1, \dots, X_n\}$  (biased!)

### 3 Distributions and their derivation/properties

#### 1. Chi-square:

- (a) **Definition:**  $\chi^2$  with  $n$  degrees of freedom with  $\rho_n(x) = \frac{x^{n/2-1} e^{-x/2}}{\Gamma(n/2) 2^{n/2}}$  (i.e.  $\text{Gamma}(n/2, 2)$ ).
- (b) **Where it comes up:**
  - i. *Sum of squares of standard normals:*  $X_j \sim N(0, 1)$  i.i.d.  $\Rightarrow Y = X_1 + \dots + X_n \sim \chi_n^2$  (look at char fns, or alternative differentiate the CDF of  $Y$  and factor out the derivative of  $f(x) = \text{Vol}(B(0, \sqrt{x}))$ ). As a result, we have  $\sum_{j=1}^n \left[\frac{X_j - \mu}{\sigma}\right]^2 \sim \chi_n^2$  if  $X_j \sim N(\mu, \sigma^2)$  i.i.d.
  - ii. *Scaled sample variance:* If  $X_j \sim N(\mu, \sigma^2)$  i.i.d. then  $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$  where  $s^2 = \sum_{j=1}^n \frac{(X_j - \bar{X})^2}{\sigma^2}$ . In addition,  $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$  and  $\bar{X}, s^2$  are independent. Show this by integrating the joint CDF  $F(u, v)$  and applying an orthogonal transformation  $y = Px$  where the first row of  $P$  is a row of  $1/\sqrt{n}$ 's. Factor out the normal CDF and the CDF for sum of squares of standard normals.
  - iii. *Normal linear regression:* Assume  $x \in \mathbb{R}^n$  with  $x \sim N(A\theta, \sigma^2 I)$  for some  $A \in \mathbb{R}_{n \times k}$ ,  $\theta \in \mathbb{R}^k$ . Let  $\hat{\theta} = (A^T A)^{-1} A^T x$  be the LSE for  $\theta$ . Then  $\frac{\|x - A\hat{\theta}\|_2^2}{\sigma^2} \sim \chi_{n-k}^2$  and is  $\perp$  of  $\hat{\theta}$  (think of  $A\hat{\theta}$  as the projection of  $x$  onto the colspace of  $A$ ).
  - iv. *ANOVA:* The setup is  $k$  normally distributed populations each with  $n$  samples with fixed variance  $\sigma^2$ .  $H_0$  represents hypothesis that means are all equal to  $\mu$ . Let:
    - A.  $SSW$  be the 'unit-wise' sample variance  $\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$

B.  $SSB$  be the 'block-wise' sample variance  $n \sum_{i=1}^k (\bar{X}_i - \bar{X})^2$ . Then  $\frac{SSB}{\sigma^2} = \chi_{k-1}^2$  (use property (2) by noticing that  $\bar{X}_i = \mathcal{N}(\mu, \sigma^2/n)$ ) and  $\frac{SSW}{\sigma^2} = \chi_{k(n-1)}^2$  (use property (2) on the inner sum), and  $SSB \perp SSW$  (use fact that sample variance is indep. of sample mean).

v. *Chi-square for goodness-of-fit*: Consider  $n$  samples from a multinomial dbn with  $k$  categories  $A_i$  each w.p.  $p_i$ . Let  $f_i$  be the observed frequency of category  $A_i$ . Consider the statistic

$$\chi^{(l)} = \sum_{i=1}^k \frac{f_i - np_i}{np_i}^2 \quad (\text{measure how well data fits with hypothesized } p_i).$$

Then the distribution of  $\chi^{(l)}$  tends to  $\chi_{k-1}^2$ . (To see this, interpret  $\chi^{(l)}$  as the norm of a vector  $v$ . Take an orthogonal transformation  $P^{-1/2}v$  where  $P$  has diagonals  $\{p_i\}$  and show this vector has a limiting distribution  $N(0, I_k - \sqrt{p}\sqrt{p}^T)$  using multivariate CLT).

## 2. t-distribution:

(a) **Definition**:  $t$  dbn with  $n$  degrees of freedom has  $\rho_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$ . 'Normal with heavier tails.'

(b) **Where it comes up**:

- i. *Relation with  $\chi^2$ , Gaussian*:  $T = \frac{X}{\sqrt{Y/n}}$  has a  $t$ -dbn with  $n$  DOF if  $X \perp Y$ ,  $X \sim N(0, 1)$ , and  $Y \sim \chi_n^2$ .
- ii. *Conf. interval for mean of normal sample, variance unknown*: Take  $(X_j)_{j=1}^n$  i.i.d. samples from  $N(\mu, \sigma^2)$ . Then  $P(\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}) = 1 - \alpha$ , independent of  $\sigma$ . (use previous property to  $t = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$ ).
- iii. *Prediction error*: Suppose  $x \in \mathbb{R}^n$  with  $x = \mathcal{N}(A\theta, \sigma^2 I)$ ,  $x_{n+1} \in \mathbb{R}$  with  $x_{n+1} = \mathcal{N}(\langle \alpha, \theta \rangle, \sigma^2)$ , indep. of  $x$ . Let  $\hat{x}_{n+1} = \langle \alpha, (A^T A)^{-1} A^T x \rangle$ . Then  $\frac{x_{n+1} - \hat{x}_{n+1}}{s \sqrt{1 + \langle \alpha, (A^T A)^{-1} \alpha \rangle}} = t_{n-k}$  (in distribution). Furthermore the interval  $(\langle \alpha, \hat{\theta} \rangle \pm t_{\beta/2, n-k} s \sqrt{1 + \langle \alpha, (A^T A)^{-1} \alpha \rangle})$  contains  $x_{n+1}$  w.p.  $(1 - \beta)$ . (Manipulate into form of (1), using relationship (3) between  $\xi^2$  and normal regression).

## 3. F-distribution:

(a) **Definition**:  $F_{m,n}$  is the quotient of 2 indep.  $\chi^2$  r.v.'s with  $m, n$  DOF normalized by their respective DOFs, i.e.  $\frac{\chi_m^2/m}{\chi_n^2/n}$ .

(b) **Where it comes up:**

- i. *ANOVA*: (see above for setup). Consider the ratio  $f = \frac{SSB/(k-1)}{SSW/k(n-1)}$ , of which the  $\lambda$ -ratio test is a monotonically decreasing function. Then  $f \sim F_{k-1, k(n-1)}$ . This follows from the property (4) for  $\chi^2$  distribution.

4. **Exponential family:**

(a) **Definition:**

(b) **Where it comes up:**

## 4 Order statistics

### 4.1 Definitions

order statistic

### 4.2 Useful theorems

**Theorem 1.** (The distribution of the order statistics)

Let  $X_1, \dots, X_n$  be i.i.d. samples with distribution function  $F(x)$  and density  $f(x) = F'(x)$ . Let  $(X_{(1)}, \dots, X_{(n)})$  denote the order-statistics. Then:

1. The density  $\rho(y_1, \dots, y_n)$  of  $(X_{(1)}, \dots, X_{(n)})$  is simply  $n!f(y_1)\dots f(y_n)$ ,  $\forall y_1 < \dots, y_n$  (and 0 otherwise). As a result, the orderstatistics are sufficient.

2. 
$$P(X_{(r)} \leq x) = \sum_{j=r}^n \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$$

3. If  $r(n)$  is a sequence of indices with  $\frac{r}{n} \rightarrow p$  faster than  $\frac{1}{\sqrt{n}}$ , then  $\sqrt{n}(X_{(r(n))} - \xi_p) \Rightarrow \mathcal{N}(0, \frac{p(1-p)}{f(\xi_p)^2})$  where  $\xi_p = F^{-1}(p)$ .

**Depends on:** Binomial CDF's, normal approximation to the binomial

**Proof idea:** 1. This follows directly from the fact that since the  $X_i$ 's are i.i.d., they are exchangeable - just look at the CDF of the orderstats vector and differentiate. Sufficiency follows directly from this since the density of  $f(x_1, \dots, x_n) = \frac{1}{n!}\rho(y_1, \dots, y_n)$  factorizes.

2. The event  $\{X_{(r)} \leq x\}$  means at least  $r$  observations were  $\leq x$ , so we can interpret each sample  $X_j$  as a Bernoulli trial with probability of success  $F(x)$  so that  $P(X_{(r)} \leq x) = P(\text{Binom}(n, F(x)) \geq r)$ . We can also derive the density of  $X_{(r)}$  using this approach by considering the probability of the outcome that there is exactly 1 observation falling in  $[x, x + h]$ ,  $r - 1$  observations in  $(-\infty, x)$ , and  $n - r$  observations in  $(x + h, \infty)$ .
3. Use the normal approx. to the binomial to argue that if  $\frac{r - n\theta}{\sqrt{n\theta(1-\theta)}} \rightarrow x$  then  $P(\text{Binom}(n, \theta) \geq r(n)) \rightarrow 1 - \Phi(x)$ . Therefore,  $P(\frac{\text{Binom}(n, \theta) - n\theta}{\sqrt{n\theta(1-\theta)}} \geq \frac{r(n) - n\theta}{\sqrt{n\theta(1-\theta)}}) \rightarrow 1 - \Phi(x)$ . Now consider  $P(X_{(r)} \leq \xi_p + x/\sqrt{n})$ .
4. Show first when  $F(x) = x$  (the *Unif*[0, 1] dbn). Using the previous formula for  $P(X_{(r)} \leq x')$  with  $x' = \xi_p + x/\sqrt{n}$ , and the result from the previous step (and the fact that  $\xi_p + x/\sqrt{n} \rightarrow \xi_p$ , it STS that:  $\frac{r - n(\xi_p + x/\sqrt{n})}{\sqrt{np(1-p)}} \rightarrow \frac{-x}{\sqrt{p(1-p)}}$ , which is easily verified by manipulations and using the condition on  $r(n)$ .
5. TO extend to arbitrary distributions, we observe that  $F(X_{(1)}), \dots, F(X_{(n)})$  are the order statistics of a uniform distribution (by monotonicity of  $F$  and the property that  $F(X)$  is *Unif*(0, 1)). So  $\sqrt{n}(F(X_{(r)}) - F(\xi_p)) \Rightarrow \mathcal{N}(0, p(1-p))$  in distribution. Taylor expanding  $F(X_{(r)})$  on the LHS about  $\xi_p$  gives the result.

□

**Theorem 1.** (Using order stats to estimate p-quantiles)

Let  $X_1, \dots, X_n$  be an i.i.d. sample with distribution  $F$  and let  $\xi_p = F^{-1}(p)$ . Then if we set  $i = np - z_{\alpha/2}\sqrt{np(1-p)}$  and  $j = np + z_{\alpha/2}\sqrt{np(1-p)} + 1$ , the interval  $[X_{(i)}, X_{(j)}]$  contains  $x_{i_p}$  with probability approximately  $1 - \alpha$ .

**Depends on:** normal approximation to binomial

**Proof idea:** Notice that  $X_{(i)} < \xi_p \Leftrightarrow$  at least  $i$  observations are less than  $\xi_p$ , and  $X_{(j)} > \xi_p \Leftrightarrow$  at most  $j - 1$  observations are less than  $\xi_p$ . Interpreting a 'succes' of  $k$ th trial as the event  $X_k < \xi_p$ , we get that  $P(X_{(i)} < \xi_p <$

$X_{(j)}) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}$ . By the normal approximation this sum is approximately  $\Phi(\frac{j-1-np}{\sqrt{np(1-p)}}) - \Phi(\frac{i-np}{\sqrt{np(1-p)}}) = 1 - \alpha$  by construction.

□

## 5 Other stuff

**Theorem 1.** (Propagation of errors/Delta-method)

Let  $Y_n$  be a sequence of random variables and let  $a_n \rightarrow 0$  be a sequence of numbers s.t.  $a_n^{-1}(Y_n - \mu) \rightarrow \mathcal{N}(0, \Sigma)$  in dbn. Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be  $C^1$  at  $x = \mu$ . Then  $a_n^{-1}(f(Y_n) - f(\mu)) \rightarrow \mathcal{N}(0, \langle \nabla f, \Sigma \nabla f \rangle)$ .

**Depends on:** Slutsky, Taylor expansion

**Proof idea:** Take  $M$  arbitrarily large and show that  $P(|Y_n - \mu| > a_n M) \rightarrow 2(1 - \Phi(M))$ . Sending  $M \rightarrow \infty$  gives that  $Y_n \rightarrow \mu$  in probability. Then Taylor expand  $\frac{f(Y_n) - f(\mu)}{a_n}$  about  $\mu$  to get the result.  $\square$

**Theorem 2.** (Properties of sample correlation)

The sample correlation  $r$  of a bivariate normal distribution with correlation  $\rho$  has the same distribution as  $[\sum_{j=2}^n X_i Y_i] / [\sum_{i=2}^n X_i^2 \sum_{i=2}^n Y_i^2]^{1/2}$  where  $(X_i, Y_i)$  are indep. samples from the standard normal distribution with correlation  $\rho$ . Furthermore, the  $\sqrt{n}(r - \rho) \rightarrow \mathcal{N}(0, (1 - \rho^2)^2)$  in distribution.

**Depends on:** Orthogonal transformations, CLT, and propagation of errors

**Proof idea:** To show the first, assume mean 0 and unit variance (need to verify that sample correlation is invariant to linear transformations). The joint density is then:

$(\frac{1}{2\pi(1-\rho^2)})^{n/2} \exp(-\frac{1}{2(1-\rho^2)}(\|x\|_2^2 - 2\rho\langle x, y \rangle + \|y\|_2^2))$ , which is invariant to orthogonal transformations. Take an orthogonal matrix  $Q$  with first row of  $1/\sqrt{n}$ 's and let  $(U, V) = (QX, QY)$ . Show that  $\|X - \bar{X}e\|_2^2 = \sum_{j=2}^n U_j^2$  and analogously for  $Y$  and  $V$ . The rest follows.

To show the second part, define 3D vector  $(X_i^2, Y_i^2, X_i Y_i)^T$  and apply the multivariate CLT to this sequence (the mean is  $(1, 1, \rho)^T$ ). Let  $f(u_1, u_2, u_3) =$

$\frac{u_3}{\sqrt{u_1 u_2}}$  and take  $u_1 = \frac{1}{n} \sum_{j=1}^n X_j^2, u_2 = \frac{1}{n} \sum_{j=1}^n Y_j^2, u_3 = \frac{1}{n} \sum_{j=1}^n X_j Y_j$ . Apply propagation of errors.  $\square$

**Remark:** This motivates Fisher's z-transformation: if we take  $\eta = \tanh^{-1}(\rho)$ ,  $z = \tanh^{-1}(r)$  we can use propagation of errors to show that  $\sqrt{n}(\eta - z) \rightarrow \mathcal{N}(0, 1)$ . This can in turn be used to construct confidence intervals for  $\tanh^{-1}(\rho)$  around  $\tanh^{-1}(r)$ . Applying  $\tanh$  to the endpoints of the interval gives a confidence interval for  $\rho$ .

## 6 Hypothesis testing

### 6.1 Definitions

1. **2-hypothesis test:** Consider a r.v.  $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^n, \mathcal{B}^n, \mu)$  where  $\mu \in \{P, Q\}$ .
  - (a) **(Randomized) Test:** a measurable function of the observations  $f : \mathbb{R}^n \rightarrow [0, 1]$  where we reject the hypothesis  $\mu = P$  with probability  $f(X)$ .  $f$  is a non-randomized test if  $f = 1_A$  for some Borel set  $A \subseteq \mathbb{R}^n$ .
  - (b) **Size:**  $\int f dP = \text{Prob of Type I error or false negative}$
  - (c) **Power:**  $\int f dQ = (1 - \text{Prob of Type II error or false positive})$ .
2. **Likelihood ratio:** Given above test,  $R(x) = h(x)/(1 - h(x))$  where  $h = \frac{dP}{d(P+Q)}$  (and  $(1 - h) = \frac{dQ}{d(P+Q)}$ ).
3. **Admissible test:** a test  $f$  for  $P, Q$  s.t.  $\nexists$  a test  $g$  dominating  $f$ , i.e.  $size(f) \geq size(g)$  and  $power(f) \leq power(g)$ .

**Theorem 1.** Neyman-Pearson lemma: the best randomized tests are simple threshold tests

Let  $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^n, \mathcal{B}^n, \mu)$  be an r.v. with  $\mu \in \{P, Q\}$ , both of which have densities wrt Lebesgue measure  $dx$ . Let  $\lambda > 0$  be any number and define  $S^{(+, -, 0)} = \{x : \lambda \rho_P(x) (>, <, =) \rho_Q(x)\}$ , respectively.

Then for any  $A \subseteq S^0$ , the test  $f_A^*$  which rejects  $P$  iff  $x \in S_1 = S^- \cup A$  (i.e.  $f_A^*(x) = 1_{S_1}(x)$ ) is a most powerful test with size  $\alpha = \int_{S_1} \rho_P(x) dx$  in the sense that any randomized test  $f$  of size at most  $\alpha$  has power at most  $\int_{S_1} \rho_Q(x) dx$ .

**Depends on:** Simple computation

**Proof idea:** We want to show that the power  $\mathbb{E}_Q[f(X)] \leq \mathbb{E}_Q[1_{S_1}]$ . The proof for  $f = 1_B$  for some  $B \in \mathcal{B}^n$  is as follows (the proof for randomized tests is more complicated):

$$\begin{aligned}
power(f) &= Q(B) = Q(B \cap S_1) + Q(B \cap S_1^c) \\
&\leq Q(B \cap S_1) + \lambda P(B \cap S_1^c) = Q(B \cap S_1) + \lambda[P(B) - P(B \cap S_1)] \text{ (since } \\
&\lambda \rho_P \geq \rho_Q \text{ outside of } S_1) \\
&\leq Q(B \cap S_1) + \lambda[P(S_1) - P(B \cap S_1)] \text{ (since the size is at most } \alpha) \\
&= Q(B \cap S_1) + \lambda P(S_1 \cap B^c) \\
&\leq Q(B \cap S_1) + Q(S_1 \cap B^c) = Q(S_1) \text{ (since } \lambda \rho_P \leq \rho_Q \text{ on } S_1) \\
&= power(f_A^*). \quad \square
\end{aligned}$$