Statistics 583: Final Exam Solution

Wellner; 6/5/2000

1. (30 points). **Define** the following terms, providing an appropriate context for your definition:
   A. A continuous functional $T(F)$ with respect to a metric d on distributions.
   B. A Fréchet differentiable functional $T(F)$ with respect to a metric d.
   C. The cumulative hazard function corresponding to a distribution function F.
   D. A maximal invariant with respect to a group G.
   E. An invariant test with respect to a group G.
   F. A metric d between distribution functions which is "compatible" with respect to the empirical measure.

   Solution: See the Chapter 6,7, and 8 notes for A,B,D,E, and F.
   For C, the cumulative hazard function corresponding to a distribution function F on $R^+ = [0, \infty)$ is defined by

   $$\Lambda(x) \equiv \int_0^x \frac{1}{1 - F(t-)} dF(t) \qquad \text{for} \qquad x \geq 0\,.$$

2. (27 points). **State** any *three* of the following four theorems, providing an appropriate context in each case:
   A. The Neyman-Pearson lemma.
   B. A Theorem about UMP $G-$invariant tests in the presence of monotone likelihood ratio of the $G-$maximal invariant.
   C. A theorem about asymptotically normality of the natural empirical estimator $T_n = T(\mathbb{F}_n)$ of a Fréchet - differentiable functional $T(F)$.
   D. Some version of *Hoeffding's formula* for the distribution of the ranks under alternatives.

   Solution: See the Chapter 6,7, and 8 notes.

3. (40 points). A simple test for asymmetry of a distribution function F is based on the difference of the mean and median: $T(F) = \int x dF(x) - F^{-1}(1/2)$. Note that $T(F) = 0$ if F is symmetric about some point, while $T(F)$ is positive for F skewed to the right, and negative for F skewed to the left. Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F, and let $\mathbb{F}_n$ be the empirical distribution of the sample, $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1_{(-\infty,x]}(X_i)$.
   A. Show that T is invariant under location shifts: If F_θ is defined by $F_\theta(x) = F(x - \theta)$, then $T(F_\theta) = T(F)$.
   B. Under appropriate assumptions on F (make these explicit), find the influence function of $T(F)$. [Hint: we already essentially did this in class, since we calculated the influence functions of $F^{-1}(1/2)$ and $\int x dF(x)$.]

C. Under appropriate assumptions on F (make these explicit), state a limit theorem for $\sqrt{n}(T(\mathbb{F}_n) - T(F))$ and compute the asymptotic variance $V^2(F)$.
D. Consider the bootstrap and jackknife estimators of $V_n(F) \equiv Var_F(\sqrt{n}(T(\mathbb{F}_n) - T(F)))$. Would either or both of these "work" for estimation of $V^2(F)$? Explain why or why not.

Solution: A. For $F_\theta(x) = F(x-\theta)$, $F_\theta^{-1}(u) = F^{-1}(u) + \theta$, so that

$$\begin{aligned} T(F_\theta) &= \int x dF_\theta(x) - F_\theta^{-1}(1/2) \\ &= \int x dF(x-\theta) - (F^{-1}(1/2) + \theta) \\ &= \int (y+\theta) dF(y) - (F^{-1}(1/2) + \theta) \\ &= \int x dF(x) + \theta - (F^{-1}(1/2) + \theta) \\ &= \int x dF(x) - F^{-1}(1/2) = T(F)\,. \end{aligned}$$

Thus $T(F)$ is invariant under location shifts.
B. To compute $IC(x; T, F)$ we calculate first the Gateaux derivative $\dot{T}(F; G-F)$. For $F_t \equiv (1-t)F + tG$, and assuming that $E_F X^2 < \infty$ and that F has a positive density f at $F^{-1}(1/2)$,

$$\begin{aligned} \dot{T}(F, G-F) &= \frac{d}{dt} T(F_t)|_{t=0} \\ &= \frac{d}{dt}\int x dF_t(x)|_{t=0} - \frac{d}{dt} F_t^{-1}(1/2)|_{t=0} \\ &= \int (x-\mu_F) dG(x) - \left(\frac{-(G-F)(1/2)}{f(F^{-1}(1/2))}\right)\,. \end{aligned}$$

Taking $G = 1_{[x,\infty)}$ corresponding to point mass at x yields the influence function for $T(F)$ at F:

$$IC(x; T, F) = (x - \mu_F) + \frac{1}{f(F^{-1}(1/2))}(1_{(-\infty, F^{-1}(1/2)]}(x) - 1/2)\,.$$

C. Thus assuming that $E_F(X^2) < \infty$ and that F has a density f at $F^{-1}(1/2)$ with $f(F^{-1}(1/2)) > 0$, the asymptotic variance $V^2(F)$ of $\sqrt{n}(T_n - T(F))$ will be

$$\begin{aligned} V^2(F) &= Var_F(X) + \frac{1/4}{f^2(F^{-1}(1/2))} \\ &\quad + 2E_F\left\{(X-\mu_F)(1_{(-\infty, F^{-1}(1/2)]}(X) - 1/2)\right\}/f(F^{-1}(1/2)) \\ &= Var_F(X) + \frac{1/4}{f^2(F^{-1}(1/2))} \\ &\quad + 2E_F\left\{(X-\mu_F)1_{(-\infty, F^{-1}(1/2)]}(X)\right\}/f(F^{-1}(1/2)) \\ &= Var_F(X) + \frac{1/4}{f^2(F^{-1}(1/2))} \\ &\quad + 2(E_F\left\{X 1_{(-\infty, F^{-1}(1/2)]}(X)\right\} - \mu_F/2)/f(F^{-1}(1/2))\,. \qquad (0.1) \end{aligned}$$

Thus we expect to have

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, V^2(F))$$

where $V^2(F)$ is given by (0.1). For example, if F is exponential(1), then $T(F) = E_F(X) - F^{-1}(1/2) = 1 - \log 2 \approx 0.30685$, $Var_F(X) = 1$, $f(F^{-1}(1/2)) = 1/2$, and

$$\begin{aligned} E_F(X 1_{[X \le F^{-1}(1/2)]}(X)) &= \int_0^{\log 2} x e^{-x} dx \\ &= -xe^{-x}|_{x=0}^{\log 2} + \int_0^{\log 2} e^{-x} dx \\ &= -\frac{1}{2}\log 2 + (1 - e^{-\log 2}) \\ &= 1 - \frac{1}{2} - \frac{1}{2}\log 2 \,. \end{aligned}$$

Thus

$$\begin{aligned} V^2(\exp(1)) &= 1 + \frac{1/4}{(1/2)^2} + \frac{2}{1/2}\left(\frac{1}{2} - \frac{1}{2} - \frac{1}{2}\log 2\right) \\ &= 2(1 - \log 2) \approx 0.61371 \,. \end{aligned}$$

Note that the contribution of the third term (minus twice the covariance between the influence functions for the mean and median) is $-2\log 2 \approx -1.386$, and the asymptotic correlation between the sample mean and sample median is

$$\begin{aligned} &\frac{-(E_F\left\{X 1_{(-\infty, F^{-1}(1/2)]}(X)\right\} - \mu_F/2)/f(F^{-1}(1/2))}{\sqrt{Var_F(X)\cdot 1/(4f^2F^{-1}(1/2))}} \\ &= \frac{-2(E_F\left\{X 1_{(-\infty, F^{-1}(1/2)]}(X)\right\} - \mu_F/2)}{\sqrt{Var_F(X)}} \\ &= \log 2 \approx 0.693 \qquad \text{for} \quad F = exp(1) \,. \end{aligned}$$

D. The jackknife estimator of the variance of $T(\mathbb{F}_n)$ will fail, because the jackknife estimator of the variance of the median $\mathbb{F}_n^{-1}(1/2)$ is not consistent. On the other hand, under the assumptions of parts B and C, the bootstrap estimator of the variance $V_n(F) \equiv Var_F(T(\mathbb{F}_n))$ should "work" in the sense of $nV_n(\mathbb{F}_n) \to_p V^2(F)$.

Remarks: This statistic for testing symmetry was suggested by Edgeworth (1887). See the discussion on pages 105 and 106 of Stigler (1999), which also indicates that the joint asymptotic distribution of the mean and median was known to Laplace in the early 1800's.]

References:

Edgeworth, F. Y. (1887). The empirical proof of the law of error. *Philosophical Magazine* **24**, 330 - 342.

Stigler, S. M. (1999). *Statistics on the Table.* Harvard University Press, Cambridge.

4. (40 points). Consider testing $H : F = G$ versus $K : (1-G) = (1-F)^\theta$ where $\theta > 1$ based on $X_1, \ldots, X_m$ i.i.d. F and $Y_1, \ldots, Y_n$ i.i.d. G. We want a test which is invariant with respect to monotone transformations of the data.
A. Draw a picture to show the relationship of F and G under the alternative hypothesis K. What is the relationship of the corresponding cumulative hazard functions under K? Draw a picture showing this.
B. Give an appropriate version of Hoeffding's formula for the distribution of the order Y-ranks $Q_1 < \ldots < Q_n$ under the alternative K.
C. Use the formula in B (or any other way) to compute the complete distribution of $\underline{Q}$ when $m = n = 2$ (so $N = m + n = 4$) and $\theta = 2$. For testing H versus $\theta = 2$, what observed value of $\underline{Q}$ would lead you to reject H at size $\alpha = 1/6$?
D. What statistic would you use for testing H versus K? In the case $m = n = 2$ in part C, what are its possible values? What observed value of $\underline{Q}$ leads you to reject H at size $\alpha = 1/6$?

Solution: A. For this alternative

$$\psi_\theta(u) = G \circ F^{-1}(u) = 1 - (1-u)^\theta$$

where $\psi_\theta(u) > u$ for $0 < u < 1$ when $\theta > 1$. Thus $G(x) > F(x)$ for values of x for which $0 < F(x) < 1$; i.e. $G <_s F$. Since

$$\Lambda_G(x) = \int_0^x (1 - G_-)^{-1} dG = -\log(1 - G(x))$$

if G is continuous and similarly for F, it follows that

$$\Lambda_G(x) = \theta \Lambda_F(x) ;$$

i.e. the hazard functions are proportional with constant of proportionality θ.

B. One version of Hoeffding's formula for this two-sample situation is

$$P_\theta(\underline{Q} = \underline{q}) = \frac{1}{\binom{N}{n}} E \prod_{j=1}^n \psi'_\theta(U_{(q_j)})$$

where $0 < U_{(1)} < \ldots < U_{(N)} < 1$ are the order statistics of N i.i.d. uniform$(0,1)$ random variables.
C. To compute the distribution of $\underline{Q}$ when $m = n = 2$ using the version of Hoeffding's formula given in B, we first calculate $\psi'_\theta(u) = \theta(1-u)^{\theta-1}$. Then, since $\binom{4}{2} = 6$,

$$\begin{aligned} P_\theta(\underline{Q} = (1,2)) &= \frac{\theta^2}{6} E[(1 - U_{(1)})^{\theta-1}(1 - U_{(2)})^{\theta-1}] \\ &= \frac{4!\theta^2}{6} \int\int\int\int_{0 \le u_1 \le u_2 \le u_3 \le u_4 \le 1} (1-u_1)^{\theta-1}(1-u_2)^{\theta-1} du_1 du_2 du_3 du_4 \end{aligned}$$

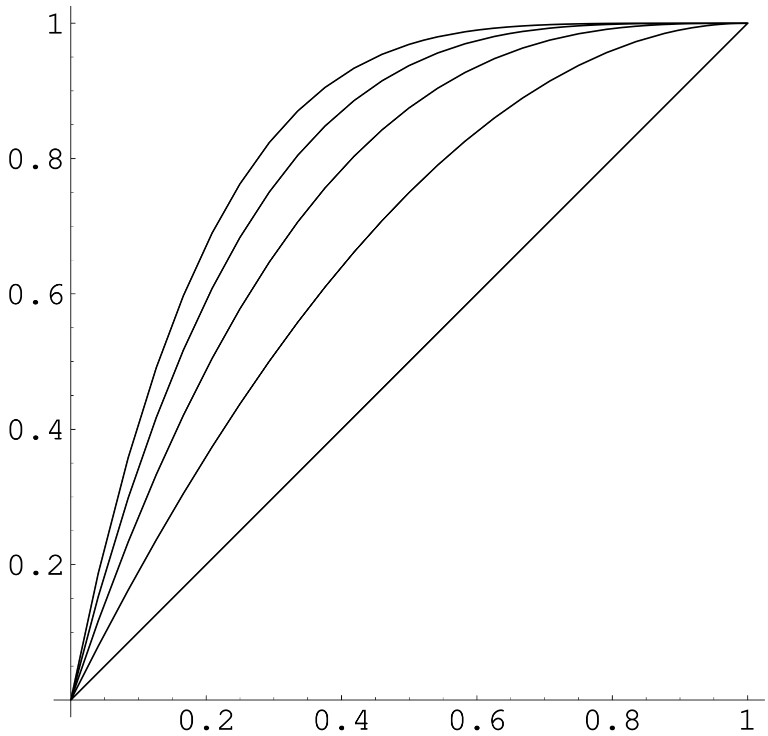

Figure 1: Plots of ψ_θ for $\theta = 1, 2, 3, 4, 5$

$$
\begin{aligned}
&= 4\theta^2 \int\int\int_{0\le u_1\le u_2\le u_3\le 1} (1-u_1)^{\theta-1}(1-u_2)^{\theta-1}\left\{\int_{u_3}^1 du_4\right\} du_1du_2du_3 \\
&= 4\theta^2 \int\int\int_{0\le u_1\le u_2\le u_3\le 1} (1-u_1)^{\theta-1}(1-u_2)^{\theta-1}(1-u_3)du_1du_2du_3 \\
&= 4\theta^2 \int\int_{0\le u_1\le u_2\le 1} (1-u_1)^{\theta-1}(1-u_2)^{\theta-1}\left\{\int_{u_2}^1 (1-u_3)du_3\right\} du_1du_2 \\
&= \frac{4\theta^2}{2}\int\int_{0\le u_1\le u_2\le 1} (1-u_1)^{\theta-1}(1-u_2)^{\theta+1}du_1du_2 \\
&= \frac{4\theta^2}{2\cdot(\theta+2)}\int_{0\le u_1\le 1} (1-u_1)^{2\theta+1}du_1 \\
&= \frac{4\theta^2}{2\cdot(\theta+2)\cdot(2\theta+2)} \\
&= \frac{2\theta^2}{(\theta+2)\cdot(2\theta+2)}.
\end{aligned}
$$

Similarly,

$$P_\theta(\underline{Q} = (1,3)) = \frac{\theta^2}{6}E[(1-U_{(1)})^{\theta-1}(1-U_{(3)})^{\theta-1}] = \frac{4\theta^2}{(\theta+1)\cdot(\theta+2)(2\theta+2)},$$

$$P_\theta(\underline{Q} = (1,4)) = \frac{\theta^2}{6}E[(1-U_{(1)})^{\theta-1}(1-U_{(4)})^{\theta-1}] = \frac{4\theta^2}{\theta(\theta+1)(\theta+2)(2\theta+2)},$$

$$P_\theta(\underline{Q} = (2,3)) = \frac{\theta^2}{6}E[(1-U_{(2)})^{\theta-1}(1-U_{(3)})^{\theta-1}] = \frac{4\theta^2}{(\theta+1)(2\theta+1)(2\theta+2)},$$

$$P_\theta(\underline{Q} = (2,4)) = \frac{\theta^2}{6}E[(1-U_{(2)})^{\theta-1}(1-U_{(4)})^{\theta-1}] = \frac{4\theta^2}{\theta(\theta+1)(2\theta+1)(2\theta+2)},$$

and, finally

$$P_\theta(\underline{Q} = (3,4)) = \frac{\theta^2}{6} E[(1-U_{(3)})^{\theta-1}(1-U_{(4)})^{\theta-1}] = \frac{4\theta^2}{\theta \cdot 2\theta(2\theta+1)(2\theta+2)}.$$

In particular, for $\theta = 2$ we get the following table:

$\underline{q}$	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)	total
$P_\theta(\underline{Q} = \underline{q})$	1/3	2/9	1/9	8/45	4/45	3/45	1

Thus for testing H versus the particular alternative in K with $\theta = 2$, by the Neyman-Pearson lemma, we would reject H if $\underline{Q} = (1,2)$.

D. For testing H versus K the locally most powerful test is based on the "log-rank statistic". If we take the "approximate scores" version of this statistic with $a_N(i) = -\log(1 - i/(N+1))$, then the test statistic is $S_n(\underline{Q}) = \sum_{j=1}^n a_N(Q_j)$. With $m = n = 2$, the possible values of the statistic are:

$S_2((1,2)) = -\log(1-1/5) - \log(1-2/5) = -\log[(4/5)(3/5)]$,
$S_2((1,3)) = -\log(1-1/5) - \log(1-3/5) = -\log[(4/5)(2/5)]$,
$S_2((1,4)) = -\log(1-1/5) - \log(1-4/5) = -\log[(4/5)(1/5)]$,
$S_2((2,3)) = -\log(1-2/5) - \log(1-3/5) = -\log[(3/5)(2/5)]$,
$S_2((2,4)) = -\log(1-2/5) - \log(1-4/5) = -\log[(3/5)(1/5)]$,
$S_2((3,4)) = -\log(1-3/5) - \log(1-4/5) = -\log[(2/5)(1/5)]$,

and we want to reject for small values of S_n. In this case the smallest value of the statistic corresponds to $\underline{Q} = (1,2)$, so at size $\alpha = 1/6$ we would reject if $\underline{Q} = (1,2)$.

5. (30 points). Suppose that X_{ij} for $j = 1, \ldots, n_i$, $i = 1, \ldots, I$ are independent, normally distributed random variables with common variance σ^2, and suppose that $EX_{ij} = \theta_i$. Thus $X_{i1}, \ldots, X_{i,n_i}$ is a sample of size n_i from the $N(\theta_i, \sigma^2)$ distribution. Consider testing $H : \theta_1 = \ldots = \theta_I$ versus $K : \theta_i \neq \theta_j$ for some $i \neq j$.
A. Explain briefly the canonical form of this testing problem, and the groups which leave the testing problem invariant.
B. Find the UMP-invariant test of H versus K, and specify its distribution under the null hypothesis H.
C. What is the distribution of the test statistic you found in B under the general hypothesis K?

Solution: A. In the canonical form of the testing problem, $Z_i \sim N(\eta_i, \sigma^2)$ are independent for $i = 1, \ldots, n$. Under the general hypothesis Θ, $\eta_i = 0$ for $i = k+1, \ldots, n$. Under the null hypothesis Θ_0, $\eta_i = 0$ for $i = 1, \ldots, r$ and for $i = k+1, \ldots, n$. Then the testing problem is invariant under the following groups:
G_1, consisting of arbitrary shifts of $Z_{r+1}, \ldots, Z_k$;
G_2, consisting of orthogonal transformations of the first r coordinates of $\underline{Z} \equiv (Z_1, \ldots, Z_n)$;
G_3, consisting of orthogonal transofrmations of the last $n-k$ coordinates of $\underline{Z}$;

G_4, consisting of scale changes of $\underline{Z}$.
B. The LS estimator $\hat{\hat{\xi}}$ of $\xi = \{\xi_{ij}\}$ under the null hypothesis is given by

$$\hat{\hat{\xi}}_{ij} = \overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}$$

where $n \equiv \sum_{i=1}^{I} n_i$. Under the general hypothesis, the LS estimator $\hat{\xi}$ is given by

$$\hat{\xi}_{ij} = \overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \text{for} \qquad j = 1, \ldots, n_i, i = 1, \ldots, I\,.$$

Thus we have

$$\|\hat{\xi} - \hat{\hat{\xi}}\|^2 = \sum_{i=1}^{I} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

and

$$\|X - \hat{\xi}\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2\,.$$

We have $I - 1$ degrees of freedom for the numerator, and $n - I$ degrees of freedom for the denominator, so that the UMP G-invariant test of H versus K is given by "reject H if $F > F_{I-1,n-I,\alpha}$" where

$$F \equiv \frac{\|\hat{\xi} - \hat{\hat{\xi}}\|^2/(I-1)}{\|X - \hat{\xi}\|^2/(n-I)}\,.$$

C. Under the general hypothesis $F \sim F_{I-1,n-I}(\delta^2)$ where the non-centrality parameter δ^2 is given by is

$$\delta^2 = \frac{1}{\sigma^2}\sum_{i=1}^{I} n_i(\theta_i - \overline{\theta}_{\cdot})^2$$

where $\overline{\theta}_{\cdot} \equiv \sum_{i=1}^{I} n_i\theta_i/n$.

6. (40 points). Suppose that $X_1, \ldots, X_m$ are i.i.d. exponential(λ), and that $Y_1, \ldots, Y_n$ are i.i.d. exponential(μ); thus the density of X_1 is $\lambda\exp(-\lambda x)1_{[0,\infty)}(x)$. Consider testing $H : \lambda \le \mu$ versus $K : \lambda > \mu$.
A. Show that this testing problem is invariant with respect to the group of scale changes, G given by $g_c(\underline{x}, \underline{y}) = (c\underline{x}, c\underline{y})$ where $c > 0$.
B. Find the UMP G-invariant test of H versus K. [Hint: you may use the fact that the family of distributions $\{\delta^{-1}F_{r,s} : \delta > 0\}$ has monotone decreasing likelihood ratio.]
C. Specify exactly how you would carry out the test derived in B.
D. Find a UMP - unbiased test of H versus K. Is this the same test as in part B or a different test?

Solution: A. If $X \sim$ exponential(λ), then

$$\begin{aligned} P_\lambda(cX > t) &= P_\lambda(X > t/c) = \exp(-\lambda t/c) \\ &= \exp(-(\lambda/c)t) = P_{\lambda/c}(X > t)\,, \end{aligned}$$

and similarly for $Y \sim$ exponential(μ). Hence the induced group on the parameter space is $\overline{g}(\lambda,\mu) = (\lambda/c, \mu/c)$. Note that for any $\overline{g} \in \overline{G}$ we have $\overline{g}\Theta_0 = \{(\lambda/c,\mu/c) : \lambda \le \mu\} = \{(\lambda,\mu) : \lambda \le \mu\} = \Theta_0$, and $\overline{g}\Theta = \{(\lambda/c,\mu/c) : (\lambda,\mu) \in R^+ \times R^+\} = \Theta$. Hence the testing problem is invariant under the group G.
B. By sufficiency we may reduce to consideration of $(S,T) = (\sum_{i=1}^m X_i, \sum_{j=1}^n Y_j)$. The induced group G^* on the space of the sufficent statistic is given by $G^* = \{g^*(s,t) = (cs, ct) : c > 0\}$, and the maximal invariant for the group G^* is $V = S/T$; the corresponding $\overline{G}-$maximal invariant is $\delta = \lambda/\mu$. Now $2\lambda X_i \sim \chi_2^2$, and similarly $2\mu Y_j \sim \chi_2^2$. Hence $2\lambda S \sim \chi_{2m}^2$ and $2\mu T \sim \chi_{2n}^2$. Hence

$$\frac{n}{m} V = \frac{\mu}{\lambda} \cdot \frac{2\lambda S/2m}{2\mu T/2n} = \delta^{-1} F_{2m,2n}$$

where $F_{2m,2n}$ has an $F-$distribution with degrees of freedom $2m, 2n$. Since the family $\delta^{-1}F_{r,s}$ has monotone decreasing monotone likelihood ratio, we conclude that the UMP G-invariant test of H versus K is given by "reject H if $nV/m < F_{2m,2n,\alpha}$ where $P(F_{2m,2n} \le F_{2m,2n,\alpha}) = \alpha$. (Alternatively, "reject H if $m/(nV) = (n^{-1}T/m^{-1}S) > F_{2n,2m,1-\alpha}$" where $P(F_{2n,2m} \ge F_{2n,2m,1-\alpha}) = \alpha$.
C. See part B.
D. The joint density of the X_i's and Y_j's is

$$\begin{aligned} \lambda^m \mu^n \exp(-\lambda \sum_1^m x_i - \mu \sum_1^n y_j) &= c_{m,n}(\lambda,\mu) \exp\left(-(\mu-\lambda)\sum_1^n y_j - \lambda(\sum_1^m x_i + \sum_1^n y_j\right) \\ &= c_{m,n}(\lambda,\mu) \exp\left(\theta U + \xi \tilde{T}\right) \end{aligned}$$

where $\theta \equiv \lambda - \mu$, $\xi \equiv -\lambda$, $U \equiv \sum_{j=1}^n Y_j$, and $\tilde{T} \equiv \sum_{i=1}^m X_i + \sum_{j=1}^n Y_j$. From our development of UMP-unbiased tests for exponential families, we know that the UMP-unbiased test ϕ of H versus K rejects H if $U > c_\alpha(\tilde{T})$ where $c_\alpha(\tilde{T})$ is the upper α percentage point of the conditional distribution of U given $\tilde{T}$. But now note that

$$\frac{U}{T} = \frac{\sum_1^n Y_j}{\sum_1^m X_i + \sum_1^n Y_j} = \frac{T/S}{1 + T/S}$$

is a monotone function of T/S, and under the null hypothesis T/S is ancillary, hence independent of $\tilde{T} \equiv T + S$ by Basu's theorem. Thus we see that the test which rejects if $(T/n)/(S/m) > F_{2n,2m,1-\alpha}$ is the UMP-unbiased test of H versus K, and this is exactly the UMP G-invariant test which we derived in B and C.