1. See class notes.

2. See Chapter 7 and 8 notes.

3. A. Let $\mathbb{H}_{n,i}$ denote the empirical distribution of the data with the ith pair (X_i, Y_i) omitted. Let $T_{n,i} \equiv T(\mathbb{H}_{n,i})$, $T_{n,\cdot} \equiv n^{-1}\sum_{i=1}^n T_{n,i}$ and let $T^*_{n,i} \equiv nT_n - (n-1)T_{n,i}$. Then the Jackknife estimator of $nVar_H(T_n)$ is

$$\frac{1}{n-1}\sum_{i=1}^n \{(T^*_{n,i} - \overline{T}^*_n)^2 \ .$$

B. The ideal bootstrap estimator of $nVar_H(T_n)$ is $nVar_{\mathbb{H}_n}(T_n)$. To implement this, we would draw B bootstrap samples

$$(X^*_{j1}, Y^*_{j1}) \ , \cdots , (X^*_{jn}, Y^*_{jn}) \ , \quad j = 1, \cdots, B \ ,$$

let $\mathbb{H}^*_{j,n}(x,y) \equiv n^{-1}\sum_{i=1}^n 1_{[X^*_{ji} \le x \ , \ Y^*_{ji} \le y]}$ be the empirical distribution function of the jth bootstrap sample, and compute $T^*_{j,n} \equiv T(\mathbb{H}^*_{j,n})$, $j = 1, \cdots, B$. Then the bootstrap estimator of $nVar_H(T_n)$ is just

$$n\,\frac{1}{B}\sum_{j=1}^B \{T^*_{j,n} - \overline{T^*_n}\}^2 \ .$$

C. Because $T(H)$ is a smooth functional of the marginal first and second moments, and the natural substitution estimators of these parameters are jointly asysmptotically normal under the hypotheses $EX^4 < \infty$, $EY^4 < \infty$, it is clear that $\sqrt{n}(T(\mathbb{H}_n) - T(H)) \to_d N(0, V^2)$ for some V^2 by the delta method. Because $T(H)$ is a ratio of moments, and because the central limit theorem does a better job of approximating sums rather than moments, it seems likely that use of the logarithmic transformation $g(x) = \log x$ might be helpful: $\sqrt{n}(\log(T(\mathbb{H}_n)) - \log(T(H))$ will probably converge to normality faster than $\sqrt{n}(T(\mathbb{H}_n) - T(H)$.
D. Because $T(H)$ is a smooth functional of moments, and because the bootstrap works for moments, it follows by preservation of bootstrap convergence under differentiable mappings that the bootstrap will "work" in this situation. Similarly, the jackknife will also estimate $nVar_H(T_n)$ consistently in this problem.

4. (This problem is from the midterm exam!)
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 $c > 0$.
Solution: If $X \sim$ exponential(λ), then

$$P_\lambda(cX > t) \;=\; P_\lambda(X > t/c) \;=\; \exp(\,-\lambda t/c)$$

$$=\; \exp(\,-(\lambda/c)t) \;=\; P_{\lambda/c}(X > t)\,,$$

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. 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 likelihood ratio.]
[I should have said *monotone decreasing likelihood ratio*.]

Solution: By sufficiency we may reduce to consideration of $(S\,,T) \equiv (\sum_1^m X_i\,, \sum_1^n Y_j)$. The induced group G^* on the space of the sufficient statistic is given by $G^* = \{g^*(s,t) = (cs\,,ct) : c > 0\}$, and the maximal invariant for the group G^* is $V \equiv S/T$; 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 $2\lambda S \sim \chi^2_{2m}$ and $2\mu T \sim \chi^2_{2n}$. 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 freeddom $2m\,, 2n$. 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 $m/(nV) = (n^{-1}T/m^{-1}S) > F_{2m,2n,1-\alpha}$" where $P(F_{2m,2n} \ge F_{2m,2n,1-\alpha}) = \alpha$.

C. Specify as exactly as possible how you would carry out the test derived in B.

Solution: See part B above.

5. In the context of testing for a disease, let $X \sim F$ denote the outcome of the test for a diseased individual and let $Y \sim G$ denote the outcome of the test for a non-diseased individual. Assuming that $X > x$ (or $Y > x$) leads to classifying the individual as "diseased", the *Receiver Operating Characteristic* or ROC curve R is a plot of *sensitivity* $\equiv P(X > x) = \overline{F}(x)$ versus $1 -$ *specificity* $\equiv 1 - P(Y \le x) = P(Y > x) = \overline{G}(x)$. Thus the ROC curve $R = R_{F,G}$ can be written as

$$R(t) \;=\; \overline{F}(\overline{G}^{-1}(t)) \;=\; 1 - F(G^{-1}(1-t))\,, \quad 0 < t < 1\,.$$

A. A good test for a disease has ROC curve with values close to 1 for small t and is everywhere above the line $I(t) = t$. Show that $R(t) \ge t$ for $0 \le t \le 1$

with strict inequality for some t if and only if $G <_s F$.

Solution: $R(t) = 1 - F(G^{-1}(1-t)) \geq t$ for all $0 \leq t \leq 1$ if and only if $1 - t \geq F(G^{-1}(1-t))$ for all $0 \leq t \leq 1$, if and only if $u \geq F(G^{-1}(u))$ for all $0 \leq u \leq 1$, if and only if $F^{-1}(u) \geq G^{-1}(u)$ for all $0 \leq u \leq 1$, if and only if $F(x) \leq G(x)$ for all $-\infty < x < \infty$, and the latter means $G \leq_s F$. If strict inequality holds for some t in the first inequality ($R(t) > t$), then strict inequality holds in $F(x) < G(x)$ for some x, and hence $G <_s F$.

B. Consider $A \equiv \int_0^1 R(t)dt$ as a measure of the quality of the disease test (values close to 1 indicating an excellent test). Show that A can be expressed in terms of the Mann-Whitney-Wilcoxon functional $\int F\, dG$.

Solution:

$$\int_0^1 R(t)\, dt = 1 - \int_0^1 F \circ G^{-1}(1-t) dt$$

$$= 1 - \int_0^1 F \circ G^{-1}(u)\, du$$

$$= 1 - EF \circ G^{-1}(U) \quad \text{where } U \sim \text{Uniform}(0,1).$$

$$= 1 - EF(Y) \quad \text{since } Y \equiv G^{-1}(U) \text{ has distributon } G$$

$$= 1 - \int F\, dG = P(Y < X).$$

C. Suppose that $X_1, \cdots, X_m$ are i.i.d. F and $Y_1, \cdots, Y_n$ are i.i.d. G and consider estimation of the ROC curve $R \equiv R_{F,G}$ on the basis of the data.

(i) Propose a nonparametric estimator $I\!R_{m,n}(t)$ of $R(t) = R_{F,G}(t)$.

Solution: A natural estimator is $I\!R_{m,n}(t) \equiv R_{IF_m, IG_n}(t) = 1 - IF_m(IG_n^{-1}(1-t))$.

(ii) Give conditions on F and G which imply that your estimator in (i) is consistent for a fixed $t \in (0,1)$.

Solution: Now $IF_m(x) \to_{a.s.} F(x)$ uniformly in x by the Glivenko-Cantelli theorem, and similarly, $IG_n(y) \to_{a.s.} G(y)$ uniformly in y. Furthermore, $IG_n^{-1}(t) \to_{a.s.} G^{-1}(t)$ for any t for which G^{-1} is continuous at t. It follows that $I\!R_{m,n}(t) \to_{a.s.} R(t)$ for any t which is a point of continuity of G^{-1}.

(iii) Compute the Gateaux derivatives of the functionals $T_1(F) \equiv R_{F,G}(t)$ for fixed G and t and $T_2(G) \equiv R_{F,G}(t)$ for fixed F and t.

Solution: Let F_1 be a fixed distribution function $\neq F$, and set $F_\epsilon \equiv (1-\epsilon)F + \epsilon F_1$. Then

$$\frac{d}{d\epsilon}\, T_1(F_\epsilon)|_{\epsilon=0} \;=\; \frac{d}{d\epsilon}\, \{1 - F_\epsilon \circ G^{-1}(1-t)\}|_{\epsilon=0}$$

$$= \; -\,(F_1 \;-\; F)(G^{-1}(1-t)) \;\equiv\; \dot{T}_1(F\,; F_1 - F)\,.$$

Similarly, Let G_1 be a fixed distribution function $\neq G$, and set $G_\epsilon \;\equiv\; (1-\epsilon)G + \epsilon G_1$. Then

$$\frac{d}{d\epsilon}\, T_2(G_\epsilon)|_{\epsilon=0} \;=\; \frac{d}{d\epsilon}\, \{1 - F \circ G_\epsilon^{-1}(1-t)|_{\epsilon=0}$$

$$= \; -\, f(G^{-1}(1-t))\, \frac{d}{d\epsilon}\, G_\epsilon^{-1}(1-t)|_{\epsilon=0}$$

$$= \; \frac{f(G^{-1}(1-t))}{g(G^{-1}(1-t))}\, \{G_1(G^{-1}(1-t)) \;-\; (1-t)\}$$

because

$$0 \;=\; \frac{d}{d\epsilon}\, (1-t)|_{\epsilon=0} \;=\; \frac{d}{d\epsilon}\, G_\epsilon(G_\epsilon^{-1}(1-t))|_{\epsilon=0}$$

$$= \; \frac{d}{d\epsilon}\, \{G(G_\epsilon^{-1}(1-t)) \;+\; \epsilon(G_1 - G) \circ G_\epsilon^{-1}(1-t))\}|_{\epsilon=0}$$

$$= \; g(G^{-1}(1-t))\, \frac{d}{d\epsilon}\, G_\epsilon^{-1}(1-t)|_{\epsilon=0}$$

$$+ \; (G_1 - G) \circ G^{-1}(1-t))\,.$$

(iv) Give conditions on F and G which imply that your estimator in (ii) is asymptotically normal (for a fixed $t \in (0,1)$). Find the influence function of your estimator (with help from (iii)).

Solution: Suppose that F and G are differentiable at $G^{-1}(1-t)$ with derivatives f and g respectively and suppose that $g(G^{-1}(1-t)) > 0$. Then, for the case $m = n$

$$\sqrt{n}(I\!R_{n,n}(t) \;-\; R(t)) \;=\; -\; \sqrt{n}(I\!F_n(I\!G_n^{-1}(1-t)) \;-\; F(G^{-1}(1-t)))$$

$$= \; -\; \sqrt{n}(I\!F_n(I\!G_n^{-1}(1-t)) \;-\; F(I\!G_n^{-1}(1-t)))$$

$$-\; \sqrt{n}(F(I\!G_n^{-1}(1-t)) \;-\; F(G^{-1}(1-t)))$$

$$= \; -\; \sqrt{n}(I\!F_n(G^{-1}(1-t)) \;-\; F(G^{-1}(1-t))) \;+\; o_p(1)$$

$$- \frac{(F(I\!G_n^{-1}(1-t)) - F(G^{-1}(1-t)))}{I\!G_n^{-1}(1-t) - G^{-1}(1-t)} \sqrt{n}(I\!G_n^{-1}(1-t) - G^{-1}(1-t))$$

$$= - \sqrt{n}(I\!F_n(G^{-1}(1-t)) - F(G^{-1}(1-t))) + o_p(1)$$

$$- \frac{f}{g}(G^{-1}(1-t))\sqrt{n}(I\!G_n^{-1}(1-t) - G^{-1}(1-t)) + o_p(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ - (1_{(-\infty, G^{-1}(1-t)]}(X_i) - F(G^{-1}(1-t)))$$

$$+ \frac{f}{g}(G^{-1}(1-t))(1_{(-\infty, G^{-1}(1-t)]}(Y_i) - (1-t))\} + o_p(1) ,$$

so the influence function in this case is

$$\psi_{F,G}(x,y) = - (1_{(-\infty, G^{-1}(1-t)]}(x) - F(G^{-1}(1-t)))$$

$$+ \frac{f}{g}(G^{-1}(1-t))(1_{(-\infty, G^{-1}(1-t)]}(y) - (1-t))$$

When $m \neq n$ it is natural to normalize by $\sqrt{mn/N}$ (as we did in the case of the Mann-Whitney-Wilcoxon statistic), and then, assuming that $\lambda_N \equiv m/N \to \lambda$, the influence function for $I\!R_{m,n}(t)$ is given by the pair of functions

$$\begin{pmatrix} - \sqrt{1-\lambda}(1_{(-\infty, G^{-1}(1-t)]}(x) - F(G^{-1}(1-t))) \\ \sqrt{\lambda}\frac{f}{g}(G^{-1}(1-t))(1_{(-\infty, G^{-1}(1-t)]}(y) - (1-t)) \end{pmatrix} .$$

(v) What can you say about your estimator $I\!R_n$ as an estimator of the function R uniformly in $0 \leq t \leq 1$?

Solution: First, note that

$$\sup_{0 \leq t \leq 1} |I\!R_{n,n}(t) - R(t))| = \sup_{0 \leq t \leq 1} |I\!F_n(I\!G_n^{-1}(1-t)) - F(G^{-1}(1-t))|$$

$$\leq \sup_{0 \leq t \leq 1} |I\!F_m(I\!G_n^{-1}(1-t)) - F(I\!G_n^{-1}(1-t))|$$

$$+ \sup_{0 \leq t \leq 1} |F(I\!G_n^{-1}(1-t)) - F(G^{-1}(1-t))|$$

$$\leq \sup_{-\infty < x < \infty} |I\!F_m(x) - F(x)|$$

$$+ \sup_{0 \le t \le 1} |F(G^{-1}(\Gamma_n^{-1}(1-t))) - F(G^{-1}(1-t))|$$

where, in the last line, we have replaced IG_n^{-1} by something equal in distribution (jointly in n), namely $G^{-1}(\Gamma_n^{-1})$ with $\Gamma_n(t) = n^{-1}\sum_{i=1}^{n} 1_{[0,t]}(\xi_i)$ the empirical d.f. of n i.i.d. Uniform$(0,1)$ random variables. By the Glivenko-Cantelli theorem we have $||IF_m - F||_\infty \to_{a.s.} 0$, so the first term converges a.s. to 0. Also by the Glivenko-Cantelli theorem and symmetry about the identity, $||\Gamma_n^{-1} - I||_\infty \to_{a.s.} 0$. Thus we see that the second term will converge a.s. to 0 if the function $F(G^{-1}(1-t))$ is uniformly continuous on $[0,1]$; i.e. if $F \circ G^{-1}$ is continuous on the closed interval $[0,1]$.

(vi) Explain how and why (or why not) you could use the jackknife or bootstrap to estimate the variance of the estimator in (i). Be as detailed and explicit as possible.

Solution: Because the jackknife fails for the median and other quantiles, and because $R_{F,G}$ involves $G^{-1}(1-t)$, it seems likely that the jackknife will fail for $R_{F,G}$. On the other hand, the bootstrap "works" for quantiles and for linear statistics (and differentiable functionals more generally), and hence it will work for $R_{F,G}(t)$.