Statistics 582, Problem Set 9 Solutions

Wellner; 3/11/2009

1. Let X and Y be independent random variables with geometric distributions

$$p_{X,Y}(x,y|\theta_1,\theta_2) = (1-\theta_1)(1-\theta_2)\theta_1^x\theta_2^y, \qquad x,y \in \{0,1,\ldots\}.$$

where $0 < \theta_j < 1$, $j = 1,2$. Find a UMP unbiased test of size $\alpha = .20$ for testing
(a) $H_0 : \theta_1 \le \theta_2$ versus $H_1 : \theta_1 > \theta_2$.
(b) $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \ne \theta_2$.
(c) For what functions $\varphi(\theta_1,\theta_2)$ do our methods guarantee existence of a UMP unbiased test of $H_0 : \varphi(\theta_1,\theta_2) = 0$ versus $H_1 : \varphi(\theta_1,\theta_2) \ne 0$?

Solution: When X and Y are independent geometric random variables, the joint density (with respect to the product of counting measure on the non-negative integers) can be rewritten as

$$\begin{aligned} P_{\theta_1,\theta_2}(X = x,\ Y = y) &= (1-\theta_1)(1-\theta_2)\exp(x\log\theta_1 + y\log\theta_2) \\ &= (1-\theta_1)(1-\theta-2)\exp(x[\log\theta_1 - \log\theta_2] + (x+y)\log\theta_2) \\ &= (1-\theta_1)(1-\theta_2)\exp(\theta U + T\xi) \end{aligned}$$

where $U(x,y) = x$, $\theta = \log(\theta_1/\theta_2)$, $T(x,y) = x+y$, and $\xi = \log\theta_2$.
(a) Testing $H : \theta_1 \le \theta_2$ versus $K : \theta_1 > \theta_2$ in this family is equivalent to testing

$$H : \theta \le 0,\ \xi = \text{ anything} \qquad \text{versus} \qquad K : \theta > 0,\ \xi = \text{ anything}.$$

On the boundary $\Theta_B = \{(\theta_1,\theta_2) : \theta_1 = \theta_2\}$, $T = X+Y$ is sufficient and complete. Hence by Theorem 2.4.2 the UMPU test of H versus K is given by

$$\phi(X,Y) = \begin{cases} 1 & \text{if } X > c(T) \\ \gamma(T) & \text{if } X = c(T) \\ 0 & \text{if } X < c(T) \end{cases}$$

where c and γ are determined by

$$E_{\theta_0=0}\{\phi(X,Y)|T = t\} = \alpha = .2. \tag{1}$$

Now by direct calculation

$$\begin{aligned} P(X=x|T=t) &= \frac{P_{\theta_1,\theta_2}(X=x,\ Y=t-x)}{P_{\theta_1,\theta_2}(T=t)} \\ &= \frac{\theta_1^x\theta_2^{t-x}}{\sum_{x'=0}^{t}\theta_1^{x'}\theta_2^{t-x'}} \\ &= \frac{(\theta_1/\theta_2)^x}{\sum_{x'=0}^{t}(\theta_1/\theta_2)^{x'}} \\ &= \frac{1}{t+1}, \qquad x\in\{0,\ldots,t\} \qquad \text{when} \quad \theta_1=\theta_2. \end{aligned}$$

Hence (1) becomes:

$$\frac{t-c(t)}{t+1}+\frac{\gamma(t)}{t+1}=.2.$$

Thus we choose

$$c(t)=\inf\{k:\ t-k\le .2(t+1)\}, \qquad \gamma(t)=.2(t+1)-(t-c(t)).$$

(b) Testing $H\theta_1=\theta_2$ versus $K:\ \theta_1\neq\theta_2$ in this family is equivalent to testing

$$H:\ \theta=0,\ \xi=\text{anything}, \qquad \text{versus} \qquad K:\ \theta\neq 0,\ \xi=\text{anything}.$$

Hence by Theorem 2.4.2 the UMPU test of H versus K is given by

$$\phi(X,Y)=\begin{cases} 1 & \text{if } X>c_2(T) \text{ or } X<c_1(T) \\ \gamma_i(T) & \text{if } X=c_i(T),\ i=1,2 \\ 0 & \text{if } c_1(T)<X<c_2(T). \end{cases}$$

where c_i and γ_i are determined by

$$E_{\theta_0=0}\{\phi(X,Y)|T=t\}=\alpha=.2 \tag{2}$$

and

$$E_{\theta_0=0}\{X\phi(X,Y)|T=t\}=\alpha E_{\theta_0=0}(X|T=t). \tag{3}$$

Since $(X|T=t)\sim$ Discrete Uniform on $\{0,\ldots,t\}$, the two equalities in (2) and (3) become

$$c_1(t)+t-c_2(t)+\gamma_1(t)+\gamma_2(t)=.2(t+1),$$

and, abbreviating $c_i(t)=c_i$, $\gamma_i(t)=\gamma_i$,

$$t(t+1)+2c_1\gamma_1+2c_2\gamma_2=\alpha t(t+1).$$

(c) For functions of the form $g(\theta_1, \theta_2) = a \log \theta_1 + b \log \theta_2 - c$ for fixed numbers $a,\ b,\ c$ with at least one of a, b different from 0 we can derive UMPU tests of $H : g(\theta_1, \theta_2) = 0$ versus $K : g(\theta_1, \theta_2) \neq 0$. This can be seen as follows: $g(\theta_1, \theta_2) = 0$ is equivalent to $\log(\theta_1 \theta_2^{b/a}) = c/a$ if $a \neq 0$. Then note that the exponential term of the joint density of X, Y can be written as

$$\begin{aligned} \exp(x \log \theta_1 + y \log \theta_2) &= \exp(x \log(\theta_1 \theta_2^{b/a}) + (y - (b/a)x) \log \theta_2) \\ &= \exp(\theta U + \xi T) \end{aligned}$$

with $U(x, y) = x$, $\theta = \log(\theta_1 \theta_2^{b/a})$, $T(x, y) = x + y$, and $\xi = \log \theta_2$.

2. (From Wasserman, *All of Statistics*, page 171). In 1861, 10 essays appeared in the *New Orleans Daily Crescent*. They were signed "Quintus Curtius Snodgrass" and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words founds in an author's work. From eight Twain essays we have

   .225, .262, .217, .240, .230, .229, .235, .217

   From 10 Snodgrass essays we have:

   .209, .205, .196, .210, .202, .207, .224, .223, .220, .201

   (a) Perform a Wald test for equality of the means. Give a $p-$value and a 95% confidence interval for the difference of means What conclusion do you reach?
   (b) Now use a permutation test (which avoids the use of large - sample methods). What is your conclusion?

   Solution: (a) Labelling the Twain proportions as X's and the Snodgrass proportions as Y's, we find that $\overline{X}_m = .231875$, $\overline{Y}_n = .2097$, $S_X = .01456$, and $S_Y = .00966$. Assuming that $X_i \sim N(\mu, \sigma^2)$ and $Y_j \sim N(\nu, \tau^2)$ with $\sigma \neq \tau$, the Wald statistic becomes

$$\begin{aligned} W_{m,n} &= \left\{ \frac{\sqrt{\frac{mn}{N}}(\overline{X}_m - \overline{Y}_n)}{\sqrt{(n/N)S_X^2 + (m/N)S_Y^2}} \right\}^2 \\ &= \left\{ \frac{\sqrt{\frac{8 \cdot 10}{18}}(.231875 - .2097)}{\sqrt{(10/18)(.000212125) + (8/18)(.0000933444)}} \right\}^2 \\ &= 3.70355^2 = 13.7163 \end{aligned}$$

   and the (approximate) p-value is $P(\chi_1^2 > 13.7163) = .000213$. If we use Welch's approximate $t-$test (see e.g. Lehmann and Casella, TSH, page 447), then the

degrees of freedom f becomes, with $R \equiv mS_X^2/(nS_Y^2)$

$$\frac{1}{f} = \left(\frac{R}{1+R}\right)^2 \frac{1}{m-1} + \frac{1}{(1+R)^2}\frac{1}{n-1} = 1/13.6148.$$

Thus the approximate p-value using Welch's approximation is $P(|t_{13.61}| \geq 13.7163) = .00265$. A 95% confidence interval for $\mu - \nu$ based on normal theory is given by

$$\begin{aligned} &\overline{X}_m - \overline{Y}_n \pm z_{.025}\sqrt{S_X^2/m + S_Y^2/n} \\ &= .022175 \pm 1.95996\sqrt{.000212125/8 + .0000933444/10} \\ &= (0.0104397, 0.0339103) \end{aligned}$$

The conclusion based on either of these tests is to reject the null hypothesis: from this evidence we would conclude that the Snodgrass and Twain essays were written by different authors.

(b) If we do an exact permutation t-test using the statistic introduced in class (involving the assumption of equal variances in the alternative), there are $\binom{18}{8} = 43758$ combinations to consider, and the observed value of the statistic (in the form $(\overline{X}_m - \overline{z})/\sigma_N$) is 2.78917. By my calculations the exact one-sided p-value is 0.000525618, and the exact two-sided p-value is 0.000777. In contrast, by drawing $10^5 = 100,000$ random permutations, the estimated p-values were 0.00058 and 0.00088 respectively.

I have not yet programmed the exact permutation test based on the Wald - type statistic used in part (a), but without squaring, which allows for the possibility of different variances. There are still $\binom{18}{8} = 43758$ combinations to consider, and the observed value of the test statistic is 3.70355. I have however, programmed the approximate permutation test based on sampling from the permutation distribution. By drawing $10^5 = 100,000$ random permutations, the estimated p-values I calculated are 0.00461 and 0.01064 respectively. It seems that the permutation distribution of the unequal variances version of the unsquared form of the Wald statistic is more nearly normal than that of the classical $t-$statistic. Note that while the two-sided permutation test still rejects at level $\alpha = .05$, this two-sided p-valued (0.01064)is not nearly as small as the estimated p-value of the permutation $t-$ test noted above (0.00088). We would continue to reject the null hypothesis at the level 0.05, but not at 0.01.

3. For observations $\underline{X} = (X_1, \ldots, X_n)$, let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the *order statistics* of the X_i's ($X_{(i)} \equiv \mathbb{F}_n^{-1}(i/n)$, $i = 1, \ldots, n$) and let $\underline{R} = (R_1, \ldots, R_n)$ denote the *ranks*; defined by $X_i = X_{(R_i)}$, $i = 1, \ldots, n$ (if $X_i = X_j$ for some $i < j$, define the ranks by $R_i < R_j$ and $X_i = X_{(R_i)}$).
(a) Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{ac}$ (the absolutely continuous df's F on R) with density f. Show that the order statistics $\underline{X}_{(\cdot)} \equiv (X_{(1)}, \ldots, X_{(n)})$ are independent of the ranks $\underline{R}$ and that the order statistics have joint density $\bar{p}$ given by

$$\bar{p}(\underline{x}_{(\cdot)}) = n! \prod_{i=1}^{n} f(x_{(i)}) \,, \qquad -\infty < x_{(1)} < \ldots < x_{(n)} < \infty$$

while

$$P(\underline{R} = \underline{r}) = \frac{1}{n!} \,, \qquad \underline{r} \in \Pi \equiv \{ \text{ all permutations of } \{1, \ldots, n\}\} \,.$$

(b) Show that (a) continues to hold for any joint distribution p of the $\underline{X}$ which is symmetric with respect to permutation of its coordinates: $p(\pi \underline{x}) = p(\underline{x})$ for all $\underline{x}$ and $\pi \in \Pi$ where $\pi x \equiv (x_{\pi(1)}, \ldots, x_{\pi(n)})$.
(c) If the joint distribution p of $\underline{X}$ is general (not permutation symmetric), show that the joint density $\bar{p}$ of the order statistics is given by

$$\bar{p}(\underline{x}_{(\cdot)}) = \sum_{\pi \in \Pi} p(\pi \underline{x}_{(\cdot)}) \,,$$

and

$$P(\underline{R} = \underline{r} | \underline{X}_{(\cdot)} = \underline{x}_{(\cdot)}) = \frac{p(r \underline{x}_{(\cdot)})}{\bar{p}(\underline{x}_{(\cdot)})} \,.$$

Solution: I will prove (c) first; then (a) and (b) follow as corollaries:
(c) Suppose that $\underline{X}$ has joint density p. Then for any set Borel set $A \subset \{\underline{x} \in$

$\mathbb{R}^n : \ x_1 < x_2 < \cdots < x_n\}$

$$\begin{aligned} P(\underline{X}_{(\cdot)} \in A) &= \int_{[\underline{x}_{(\cdot)} \in A]} p(x_1, \ldots, x_n)\ dx_1 \ldots dx_n \\ &= \sum_{r \in \Pi} \int_{[R(\underline{x})=r,\ \underline{x}_{(\cdot)} \in A]} p(x_1, \ldots, x_n)\ dx_1 \ldots dx_n \\ &= \sum_{r \in \Pi} \int_A p(x_{(r_1)}, \ldots, x_{(r_n)})\ dx_{(1)} \ldots dx_{(n)} \\ &= \int_A \overline{p}(x_{(1)}, \ldots, x_{(n)})\ dx_{(1)} \ldots dx_{(n)} \end{aligned}$$

where we have used the fact that the correspondence between $(x_1, ..., x_n)$ and $(x_{(1)}, \ldots, x_{(n)})$ is one-to-one and linear with Jacobian $= 1$ on each subset $[R = r]$, $r \in \Pi$. This proves that

$$\overline{p}(\underline{x}_{(\cdot)}) = \sum_{\pi \in \Pi} p(\pi \underline{x}_{(\cdot)}) \ .$$

Similarly,

$$\begin{aligned} P(R = r,\ \underline{X}_{(\cdot)} \in A) &= \int_{[R=r,\ \underline{x}_{(\cdot)} \in A]} p(x_1, \ldots, x_n)\ dx_1 \cdots dx_n \\ &= \int_A p(x_{(r_1)}, \ldots, x_{(r_n)})\ dx_{(1)} \ldots dx_{(n)} \\ &= \int_A \frac{p(x_{(r_1)}, \ldots, x_{(r_n)})}{\overline{p}(x_{(1)}, \ldots, x_{(n)})} \overline{p}(x_{(1)}, \ldots, x_{(n)})\ dx_{(1)} \ldots dx_{(n)} \end{aligned}$$

since $\overline{p}(x_{(1)}, \ldots, x_{(n)}) = 0$ implies $p(x_{r(1)}, \ldots, x_{r(n)}) = 0$ for each $r \in \Pi$. This implies that

$$P(\underline{R} = \underline{r} | \underline{X}_{(\cdot)} = \underline{x}_{(\cdot)}) = \frac{p(r\underline{x}_{(\cdot)})}{\overline{p}(\underline{x}_{(\cdot)})} \ .$$

(b) When $p(\underline{x}) = p(\pi \underline{x})$ for all $\pi \in \Pi$, then

$$\overline{p}(\underline{x}_{(\cdot)}) = n! p(\underline{x}_{(\cdot)}),$$

and

$$P(\underline{R} = \underline{r} | \underline{X}_{(\cdot)} = \underline{x}_{(\cdot)}) = \frac{p(r\underline{x}_{(\cdot)})}{\overline{p}(\underline{x}_{(\cdot)})} = \frac{p(r\underline{x}_{(\cdot)})}{n! p(\underline{x}_{(\cdot)})} = \frac{1}{n!}.$$

Hence R is independent of $\underline{X}_{(\cdot)}$, and $P(R = r) = 1/n!$ for each $r \in \Pi$.
(a) This follows easily from (a) since, in this case,

$$p(\underline{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{\pi(i)}) = p(\pi\underline{x}).$$

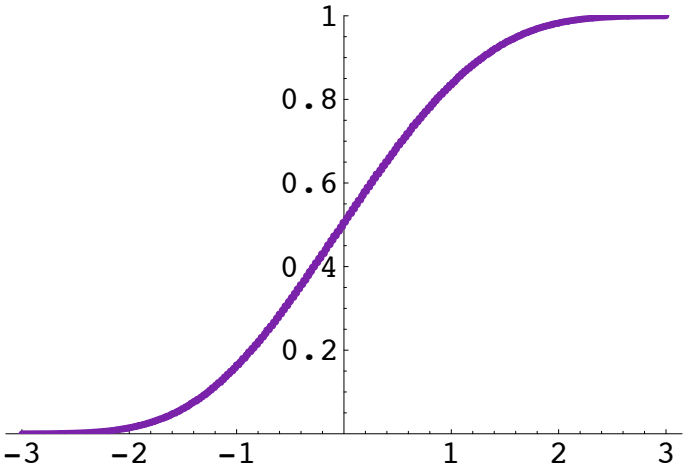

Figure 1: Exact permutation distribution, two-sample $t-$statistic $(\overline{X} - \overline{z})/\sigma_N$

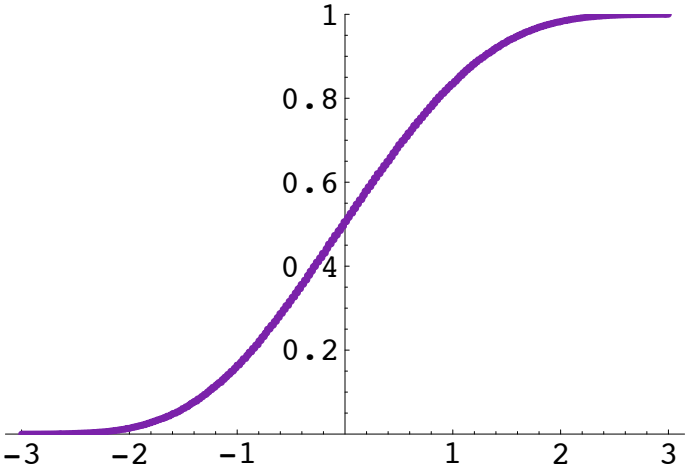

Figure 2: Approximate permutation distribution, two-sample $t-$statistic $(\overline{X} - \overline{z})/\sigma_N$, based on 10^5 random permutations

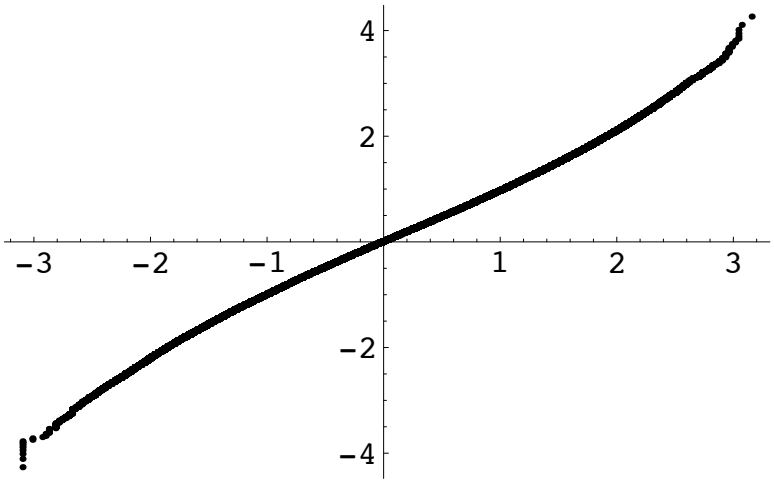

Figure 3: QQ-plot, Approximate permutation distribution, two-sample $t-$statistic $(\overline{X} - \overline{z})/\sigma_N$, based on 10^5 random permutations

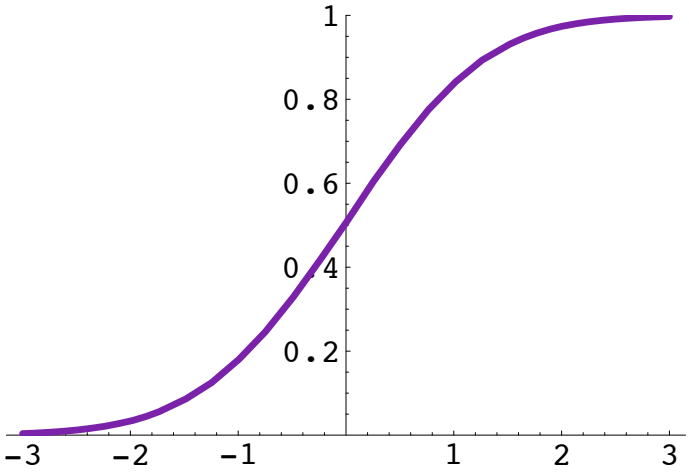

Figure 4: Approximate permutation distribution, one-sided Wald statistic $\overline{X} - \overline{Y})/\sqrt{S_X^2/m + S_Y^2/n}$ based on 10^5 random permutations

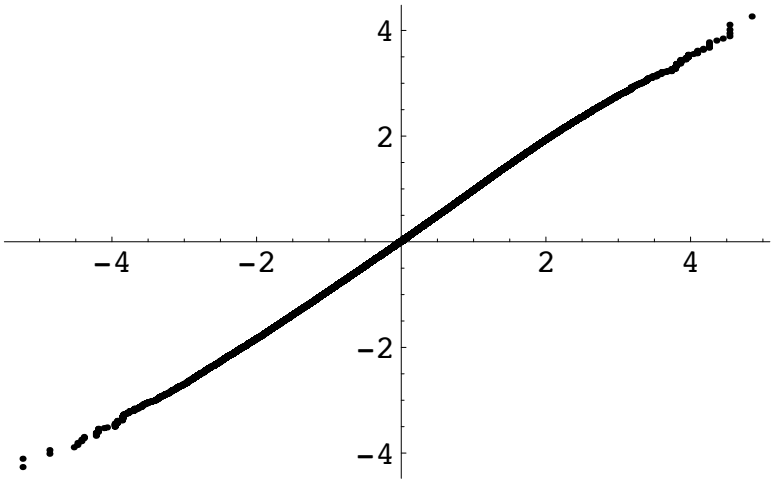

Figure 5: QQ-plot, Approximate permutation distribution, one-sided Wald statistic $\overline{X} - \overline{Y})/\sqrt{S_X^2/m + S_Y^2/n}$ based on 10^5 random permutations