Final Exam, Biostatistics 341
8 December, 2014

Please show all your work and perform all calculations to whatever degree of exactness you are able. This test is closed book and no calculators are allowed.

1. (40) Miscellaneous Questions.
(a) $P(A|B) = 1/2, P(A|B^C) = 1/4$, and $P(B) = 1/3$. What is $P(B|A)$?
(b) Let $X \sim Exp(1)$. What is the pdf of $Y = e^X$?
(c) What is the moment generating function for the gamma distribution?
(d) $Z \sim N(0,1), Y \sim \chi^2_p$, and $Z \perp Y$. State the distribution of $U = Z/\sqrt{Y/p}$.
(e) What is the probability of getting two 1's, two 2's, and one 6 if five fair and independent 6-sided dice are rolled?
(f) A fair, 6-sided die is rolled until a six is obtained. What is the probability of rolling the die exactly 5 times?

2. (40) Let X be a random variable with continuous pdf $f(\cdot)$. A biased sample of $f(\cdot)$ could be created by selecting X with probability of selection given by $w(x)$. The pdf of the resulting biased sample is

$$f^{bias}(x) = \frac{w(x)f(x)}{\int w(x)f(x)dx}.$$

Suppose $f(x) = 1$ for $0 < x < 1$ and $w(x) = x^2$.
(a) What is $f^{bias}(x)$?
(b) Describe a procedure for generating data from $f^{bias}(x)$ using the probability integral transformation and uniform(0,1) random variables.
(c) Describe a procedure for generating data from $f^{bias}(x)$ using the accept-reject algorithm and uniform(0,1) random variables.
(d) What is the probability of accepting in your accept-reject algorith? (Stated slightly differently, if you generated 1000 uniform(0,1) random variables, what is the expected number that you would accept?)

3. (40) Let $X|Y \sim Binomial(n, Y)$ and $Y \sim Beta(\alpha, \beta)$.
(a) What is $E(Y)$?
(b) What is $Var(Y)$?
(c) What is $E(X)$?
(d) What is $Var(X)$?
(e) What is the joint distribution of X and Y?
(f) What is the marginal distribution of X?

4. (40) Let X_i be *iid* random variables with continuous cdf $F(\cdot)$ and pdf $f(\cdot)$ for $i = 1, \cdots, n$.

(a) What is the probability density function of the sample median if $n = 3$? For the remainder of this problem, assume that $X_i \sim^{iid} Exp(2)$.

(b) What is the pdf of $X_{(50)}$ if $n = 99$?

(c) What is the population (theoretical) median of X_i?

(d) For $0 < p < 1$,

$$\sqrt{n}(X_{(np)} - F^{-1}(p)) \rightarrow^d N(0, \frac{p(1-p)}{f(F^{-1}(p))^2}).$$

Using this fact, derive the large sample distribution of the sample median.

(e) What is the large sample distribution of the natural logarithm of the sample median?

(f) What is the large sample distribution of the sample mean?

5. (40) Let $\mu_x = E(X), \mu_y = E(Y), \sigma_x^2 = Var(X), \sigma_y^2 = Var(Y)$, and $\rho_{xy} = Corr(X, Y)$. In our second exam, we showed that if the expectation of Y conditional on X was linear, then

$$\begin{aligned} E(Y|X = x) &= a + bx \\ &= \mu_y - \rho_{xy}\frac{\sigma_y}{\sigma_x}\mu_x + \rho_{xy}\frac{\sigma_y}{\sigma_x}x. \end{aligned}$$

Supppose that X is measured with error so instead of observing X, we observe $W = X + U$, where $E(U) = 0, Var(U) = \sigma_u^2$, and U is independent of X and Y. (Give all answers in terms of parameters that have been defined.)

(a) What is the expectation and variance of W?

(b) What is the correlation between W and Y?

(c) $E(Y|W = w) = c + dw$. What is d? (Hint: Use an expression analogous to that given above for $E(Y|X = x)$.)

(d) Suppose (Y_i, X_i, W_i) are *iid* random vectors as defined above for $i = 1, \cdots, n$. Furthermore, suppose that $V_i \sim^{iid} Bin(1, p)$ where V_i is independent of all other variables and $V_i = 1$ implies that both X_i and W_i are observed for subject i, whereas $V_i = 0$ implies that only W_i is observed. In other words, complete data are available for a subsample of n_v subjects, where $n_v = \sum_{i=1}^n V_i$. Give expressions for unbiased estimators for σ_x^2 and σ_u^2?

(e) If you got the answer to part (d) correct, these estimators are consistent. In other words, they converge in probability to what they are trying to estimate. What does it mean for a statistic to converge in probability to σ_u^2? (i.e., Write the definition of convergence in probability for this statistic.)

(f) Let $t(\underline{Y}, \underline{W})$ be a consistent estimator of d (the slope derived in part (c)). We are interested in assessing the relationship between Y and X, not Y and W. With that in mind, propose a consistent estimator for b (the slope of $E(Y|X = x)$ and x) that uses $t(\underline{Y}, \underline{W})$ and the estimators for σ_x^2 and σ_u^2 derived in part (d).

1

a $P(A|B) = \frac{1}{2}$, $P(A|B^c) = \frac{1}{4}$, $P(B) = \frac{1}{3}$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

$$= \frac{(\frac{1}{2})(\frac{1}{3})}{(\frac{1}{2})(\frac{1}{3}) + (\frac{1}{4})(\frac{2}{3})} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}(2)} = \frac{\frac{1}{2}}{1} = \boxed{\frac{1}{2}}$$

b $X \sim Exp(1)$

$f(x) = e^{-x}$ for $x > 0$

$y = e^x \Rightarrow g^{-1}(y) = x = \log(y)$ $\quad \frac{dx}{dy} = \frac{1}{y} = \frac{d}{dy} g^{-1}(y)$

$$f(y) = f(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$$

$$= e^{-\log(y)} \frac{1}{y} = \frac{1}{e^{\log(y)}} \frac{1}{y} = \boxed{\frac{1}{y^2} \text{ for } y > 1}$$

$\int_1^\infty y^{-2} = -1(y^{-1})\Big|_1^\infty = -1(0-1) = 1$

c $$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

$$E(e^{tx}) = \int e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx$$

$$= \frac{(\frac{1}{\beta} - t)^{-\alpha}}{\beta^\alpha} \int \frac{(\frac{1}{\beta} - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x(\frac{1}{\beta} - t)}$$

$$= \frac{\left(\frac{1-t\beta}{\beta}\right)^{-\alpha}}{\beta^\alpha} = \frac{\left(\frac{\beta}{1-t\beta}\right)^\alpha}{\beta^\alpha} = \boxed{\left(\frac{1}{1-t\beta}\right)^\alpha}$$

(d) $Z \sim N(0,1)$, $Y \sim \chi^2_p$, $Z \perp Y$

$U = Z / \sqrt{Y/p} \Rightarrow$ $\boxed{U \sim t\text{-distribution with } p \text{ degrees of freedom}}$

(e) $P(X=x) = \frac{1}{6}$ for $x = 1, \ldots, 6$

$P(2\ 1\text{'s}; 2\ 2\text{'s}; 0\ 3\text{'s}, 4\text{'s}, 5\text{'s}; 1\ 6) =$

$$\frac{5!}{2!\,2!\,0!\,0!\,0!\,1!}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^1$$

$$= \boxed{\frac{5!}{2!\,2!}\left(\frac{1}{6}\right)^5}$$

Multinomial Distribution

(f) Geometric distribution:

$P(X=x) = p(1-p)^{x-1}$ for $x = 1, \ldots$

$\Rightarrow P(X=5) = \boxed{\frac{1}{6}\left(\frac{5}{6}\right)^4}$

2 $f^{bias}(x) = \dfrac{w(x)\, f(x)}{\int w(x) f(x)\, dx}$

$f(x) = 1$ for $0 < x < 1$, $w(x) = x^2$

a $\Rightarrow \int w(x) f(x)\, dx = \int_0^1 x^2\, dx = \left.\dfrac{x^3}{3}\right]_0^1 = \dfrac{1}{3}$

$\Rightarrow \boxed{f^{bias}(x) = 3x^2 \quad \text{for } 0 < x < 1}$

b $F^{bias}(x) = \int_0^x 3t^2\, dt = \left.\dfrac{3t^3}{3}\right]_0^x = x^3$ for $0 < x < 1$

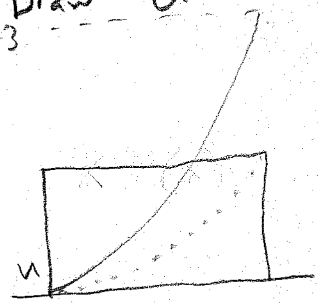
$\Rightarrow \quad t = x^3 \Rightarrow t^{1/3} = x$

$F^{-1}(t) = t^{1/3}$

Draw $U \sim Unif(0,1)$ and Let $X = U^{1/3}$

$M = 3$

c

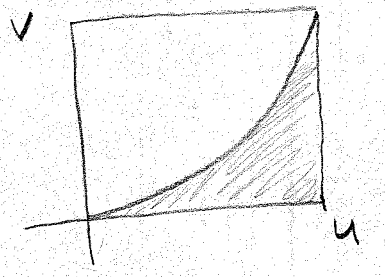


$\Rightarrow$ Draw $U \sim Unif(0,1)$.

Draw $V \sim Unif(0,1)$. $\quad U \perp V$

If $V < \frac{1}{3} \dfrac{3u^2}{1} = u^2$ then accept.

$\boxed{d}$ $P(V < U^2)$



$$= \int_0^1 \int_0^{u^2} dv\, du$$

$$= \int_0^1 u^2\, du = \left.\frac{u^3}{3}\right]_0^1 = \frac{1}{3}$$

$\boxed{3}$ $X \mid Y \sim Bin(n, Y)$, $Y \sim Beta(\alpha, \beta)$

$\boxed{a}$ $E(Y) = \displaystyle\int_0^1 y \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\, dy$

$$= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1+\beta)} \int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)} y^{\alpha+1-1} (1-y)^{\beta-1}\, dy$$

$$= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+1+\beta)} = \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\alpha+\beta)\Gamma(\alpha+\beta)} = \boxed{\frac{\alpha}{\alpha+\beta}}$$

$\boxed{b}$ $Var(Y) = E(Y^2) - (E(Y))^2$

$E(Y^2) = \displaystyle\int_0^1 y^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\, dy$

$$= \frac{\Gamma(\alpha+2)\Gamma(\alpha+\beta)}{\Gamma(\alpha+2+\beta)\Gamma(\alpha)} \int \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\Gamma(\beta)} y^{\alpha+2-1}(1-y)^{\beta-1}$$

$$= \frac{(\alpha+1)\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

$$\Rightarrow Var(Y) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2$$

$$= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \frac{\cancel{\alpha^3} + \cancel{\alpha^2\beta} + \cancel{\alpha^2} + \alpha\beta - \cancel{\alpha^3} - \cancel{\alpha^2\beta} - \cancel{\alpha^2}}{(\alpha+\beta)^2(\alpha+\beta+1)} = \boxed{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

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

$$= E(nY)$$

$$= nE(Y)$$

$$= \boxed{n\frac{\alpha}{\alpha+\beta}}$$

d $Var(X) = Var(E(X|Y)) + E(Var(X|Y))$

$$= Var(nY) + E(nY(1-Y))$$

$$= n^2 Var(Y) + nE(Y) - nE(Y^2)$$

derived in part (b)

$$= n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + n\frac{\alpha}{\alpha+\beta} - n\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

$$= \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{n\alpha}{\alpha+\beta}\left[1 - \frac{\alpha+1}{\alpha+\beta+1}\right]$$

$$= \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{n\alpha}{\alpha+\beta}\left[\frac{\cancel{\alpha}+\beta+\cancel{1}-\cancel{\alpha}-\cancel{1}}{\alpha+\beta+1}\right]$$

$$= \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{n\alpha\beta(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= \boxed{\frac{n\alpha\beta(n+\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

$\boxed{e}$ $f(x,y) = f(x|y)\, f(y)$

$$= \binom{n}{x} y^{x} (1-y)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}$$

for $x = 0, \ldots, n$

$0 < y < 1$

$\boxed{f}$ $f(x) = \displaystyle\int_0^1 f(x,y)\, dy$

$$= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(x+\alpha+n-x+\beta)} \underbrace{\int_0^1 \frac{\Gamma(x+\alpha+n-x+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}}_{1}$$

$$= \boxed{\binom{n}{x} \frac{\Gamma(\alpha+\beta)\, \Gamma(x+\alpha)\, \Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\, \Gamma(\alpha+\beta+n)} \quad \text{for } x = 0, 1, \ldots, n}$$

4

a $n=3$. Sample median $= X_{(2)}$

$$f_{X_{(2)}}(x) = \frac{3!}{1!\,1!\,1!} F(x)^1 f(x)^1 (1-F(x))^1$$

$$= \boxed{6\,F(x)\,f(x)\,(1-F(x))}$$

b $n=99$, $X_i \overset{iid}{\sim} \text{Exp}(2)$ $\Rightarrow$ $f(x) = \frac{1}{2}e^{-x/2}$ for $x>0$

$F(x) = 1-e^{-x/2}$

$$\Rightarrow \boxed{f_{X_{50}}(x) = \frac{99!}{49!\,49!}\left(1-e^{-x/2}\right)^{49} \frac{1}{2}e^{-x/2}\left(e^{-x/2}\right)^{49} \quad \text{for } x>0}$$

c $F(x) = 1-e^{-x/2}$ $\qquad F(x) = \frac{1}{2}$ if $x=$ median

Theoretical median is $F_x^{-1}(\frac{1}{2})$.

$$1-e^{-x/2} = \frac{1}{2} \Rightarrow -e^{-x/2} = -\frac{1}{2} \Rightarrow e^{-x/2} = \frac{1}{2}$$

$$\Rightarrow -\frac{x}{2} = \log\left(\frac{1}{2}\right) \Rightarrow -x = 2\log\left(\frac{1}{2}\right) \Rightarrow x = -2\log\left(\frac{1}{2}\right)$$

$$\Rightarrow x = -2(\log(1)-\log(2)) \Rightarrow x = \boxed{2\log(2)}$$

d $0 < p < 1$

$$\sqrt{n}\left(X_{(np)} - F^{-1}(p)\right) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f(F^{-1}(p))^2}\right)$$

$p = \frac{1}{2}$, $F^{-1}(p) = 2\log(2)$ from (a)

$$f(F^{-1}(p)) = \frac{1}{2}e^{-2\log(2)/2}$$

$$= \frac{1}{2}e^{-\log(2)}$$

$$= \frac{1}{2}e^{\log(\frac{1}{2})}$$

$$= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$$

$$\Rightarrow \sqrt{n}\left(X_{(\frac{n}{2})} - 2\log(2)\right) \xrightarrow{d} N\left(0, \frac{\frac{1}{4}}{\left(\frac{1}{4}\right)^2}\right) = \boxed{N(0, 4)}$$

e $$\sqrt{n}\left(\log(X_{(np)}) - \log(2\log(2))\right) \xrightarrow{d} N\left(0, \left(\frac{1}{2\log(2)}\right)^2 4\right)$$

by Δ-method $$= N\left(0, \frac{1}{(\log(2))^2}\right)$$

because $\frac{d}{d\theta}\log(\theta) = \frac{1}{\theta}$ and $\theta = 2\log(2)$

F $\sqrt{n}\left(\bar{X} - E(X)\right) \xrightarrow{d} N(0, \sigma^2)$ by CLT

$\Rightarrow \sqrt{n}\left(\bar{X} - 2\right) \xrightarrow{d} N(0, 4)$ because $E(X) = 2$, $Var(X) = 2^2 = 4$

$\boxed{5}$

$\boxed{a}$ $E(W) = E(X+U) = E(X) + E(U) = \mu_x + 0 = \boxed{\mu_x}$

$$Var(W) = Var(X+U) = Var(X) + Var(U) + 2Cov(X,U)$$
$$= \sigma_x^2 + \sigma_u^2 + 0 = \boxed{\sigma_x^2 + \sigma_u^2}$$

$\boxed{b}$ $Cor(W,Y) = Cor(X+U,Y) = \dfrac{Cov(X+U,Y)}{\sqrt{Var(X+U)\,Var(Y)}}$

$$Cov(X+U,Y) = Cov(X,Y) + Cov(U,Y)$$
$$= Cov(X,Y)$$
$$= \rho_{xy}\sigma_x\sigma_y$$

$$\Rightarrow Cor(W,Y) = \frac{\rho_{xy}\sigma_x \cancel{\sigma_y}}{\sqrt{\sigma_x^2+\sigma_u^2}\ \cancel{\sigma_y}} = \boxed{\frac{\rho_{xy}\sigma_x}{\sqrt{\sigma_x^2+\sigma_u^2}}}$$

$\boxed{c}$ $E(Y|W=w) = c + dw$

$$= \mu_y + \rho_{wy}\frac{\sigma_y}{\sigma_w}\mu_w + \rho_{wy}\frac{\sigma_y}{\sigma_w} w$$

$$\Rightarrow c = \mu_y + \frac{\rho_{xy}\sigma_x}{\sqrt{\sigma_x^2+\sigma_u^2}}\frac{\sigma_y}{\sqrt{\sigma_x^2+\sigma_u^2}}\mu_x = \mu_y + \frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_u^2}\mu_x$$

$$d = \frac{\rho_{xy}\sigma_x}{\sqrt{\sigma_x^2+\sigma_u^2}}\frac{\sigma_y}{\sqrt{\sigma_x^2+\sigma_u^2}} = \boxed{\frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_u^2}}$$

$\boxed{d}$ (Y_i, X_i, W_i) iid $i = 1, \ldots, n$

$V_i \overset{iid}{\sim} Bin(1, p)$

If $V_i = 1$ then X_i is observed

S_x^2 & S_u^2 are unbiased for σ_x^2 & σ_u^2, respectively.

$$S_x^2 = \frac{\sum V_i \left(X_i - \frac{\sum V_i X_i}{n_v}\right)^2}{n_v - 1}$$

$$S_u^2 = \frac{\sum V_i \left(W_i - X_i - \frac{\sum V_i (W_i - X_i)}{n_v}\right)^2}{n_v - 1}$$

$\boxed{e}$ $\lim_{n \to \infty} P\left(\left|S_u^2 - \sigma_u^2\right| \geq \varepsilon\right) = 0$ for $\varepsilon > 0$

(note: As $n \to \infty$, $n_v \to \infty$ because $E(n_v) = np$.

$\boxed{f}$ $t(\underline{Y}, \underline{W}) \xrightarrow{P} \dfrac{\rho_{xy} \sigma_x \sigma_y}{\sigma_x^2 + \sigma_u^2}$

But we want to estimate $\rho_{xy} \dfrac{\sigma_y}{\sigma_x}$.

$$\frac{\rho_{xy} \sigma_x \sigma_y}{\sigma_x^2 + \sigma_u^2} \lambda = \rho_{xy} \frac{\sigma_y}{\sigma_x} \Rightarrow \lambda = \frac{\cancel{\rho_{xy}} \cancel{\sigma_y} (\sigma_x^2 + \sigma_u^2)}{\sigma_x \sigma_x \cancel{\sigma_y} \cancel{\rho_{xy}}} = \frac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2}$$

A consistent estimator for $\dfrac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2}$ is $\dfrac{S_x^2 + S_u^2}{S_x^2}$

$\Rightarrow$ A consistent estimator for b is $\boxed{t(\underline{Y}, \underline{W}) \dfrac{S_x^2 + S_u^2}{S_x^2}}$