2017 First Year Exam –Theory
Statistics 200ABC
June 23, 2017
9:00 to 1:00

• There are 7 questions on the examination. Select any 5 of them to solve. If you attempt to solve more than 5 questions, you are only to turn in the 5 you want graded. If you turn in partial solutions to more than 5 questions, only 5 will be graded.

• Each of the 5 problems you attempt to solve will be worth equal credit, with each accounting for 20% of your final score on this examination.

• Your solutions to each problem should be written on separate sheets of paper. Label each sheet with your student identification number, the problem number, and the page number of that solution written in the upper right hand corner. For example, the labeling on a page may be:

ID# 912346378
Problem 2, page 3

• You have 4 hours to complete your solution. Please be prepared to turn in your exam at 1:00 pm.

Problem 1 Suppose a sequence of random variables is generated as follows. Let $X_1 \sim \text{uniform}(0,1)$, and $X_2|X_1 \sim \text{uniform}(X_1, 1)$. In general, let $X_{k+1}|(X_1, \ldots, X_k) \sim \text{uniform}(X_k, 1)$.

(a) Find the joint density of $(-\log(1-X_1), -\log(1-X_2))$ by using bivariate transformations. Be sure to specify the support of this density.

(b) Find the marginal density of $-\log(1-X_2)$.

(c) Find the conditional density of X_2 given (X_1, X_3). What about X_2 given $(X_1, X_3, X_4, \ldots, X_{20}$

(d) Find $EX_k, k = 1, 2, \ldots$ by using the law of iterated conditional expectations. (Hint: you may find it easier to work with $E(1-X_k)$.)

(e) Prove that $X_k \to 1$ in probability as $k \to \infty$.

Problem 2 Given constants $p \in (0,1)$ and $\mu_j \in (-\infty, \infty)$, $j = 0, 1$, let X_i, $i = 1, 2, \ldots$ be iid Bernoulli(p) and $Y_i|(X_i = j) \sim \mathrm{N}(\mu_j, 1)$ for $j = 0, 1$, with the pairs (X_i, Y_i) independent across i. Let $\bar{X} \equiv (1/n)\sum_{i=1}^{n} X_i$ and $\bar{Y} \equiv (1/n)\sum_{i=1}^{n} Y_i$.

(a) Find the marginal density of Y_i

(b) Find $Var(Y_i)$ and $Cov(X_i, Y_i)$.

(c) Find the conditional distribution of X_i given Y_i. Under what conditions do we have X_i independent of Y_i?

(d) Find the asymptotic distribution of $(\bar{X}, \bar{Y})$, suitably standardized.

(e) Use the Delta method to find the asymptotic distribution of $\bar{Y}/\bar{X}$.

Problem 3 Suppose $X \sim \text{Poisson}(\lambda_1)$.

(a) Show that $X(X-1)$ is the UMVUE for λ_1^2.

(b) Given a sample of n observations for X, show that $\bar{X}^2 - \bar{X}$ is a biased estimator of λ_1^2; then evaluate the amount of bias and use the result to derive an unbiased estimator for λ_1^2.

(c) Given a sample of n observations for X, find the MLE of $P(X \leq 1)$ and derive its asymptotic variance.

(d) Next, suppose that $Y \sim \text{Poisson}(\lambda_2)$ and $Z \sim \text{Poisson}(\lambda_1 + \lambda_2)$, such that X, Y, Z are independent given $\theta = (\lambda_1, \lambda_2)$. Find a sufficient statistic for θ. Is it complete? Justify your answer.

(e) Find the MLE of θ.

(f) Find the Cramer-Rao lower bound of an unbiased estimator of θ.

(g) Find the conditional distribution of X given $X + Y + Z = m$.

Problem 4 Let $X_1, \ldots, X_n$ be i.i.d. random variables with Bernoulli(p) distribution. We set $T = \sum_{i=1}^{n} X_i$.

(a) Show that $T(n-T)/(n(n-1))$ is an unbiased estimator of pq, where $q = 1 - p$.

(b) Find the MLE of pq.

(c) Show that the MLE of pq is biased, but it is unbiased asymptotically.

(d) Compare the variance of the unbiased estimator in part (a) to the variance of the MLE; which one is smaller?

(e) Assuming $p \sim \text{Beta}(a, b)$ and using the squared error loss function, write down the formal Bayes rule for estimating p.

(f) Using part (e), find the Bayes estimate given $n = 10$, $T = 4$, and $a = b = 1$.

Problem 5 Assume that $(X_1, X_2, \cdots, X_{2n}) \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown. For some reason, only the following statistics are available (i.e., the individual values are not available):

- $SS_y = \sum_{i=1}^{n} Y_i^2$ where $Y_i = X_{2i-1} - X_{2i}$.
- $SS_z = \sum_{i=1}^{n} Z_i^2$ where $Z_i = X_{2i-1} + X_{2i}$.
- $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

We are interested in inference on μ using the available information.

(a) Show that SS_z and $\bar{Z}$ are NOT independent. Furthermore, find the conditional distribution of $SS_z|\bar{Z} = 0$.

(b) Show that $SS_y/(2\sigma^2) \sim \chi^2_k$. Find and justify the value of k.

(c) Construct an F-statistic that follows $F_{n,n}$ when $\mu = 0$. The solution may depend on the above summary statistics, and appropriate constants. Please show your work.

(d) Construct an F-statistic that follows $F_{1,n}$ when $\mu = 0$. The solution may depend on the above summary statistics, and appropriate constants. Please show your work.

(e) Is it possible to construct an F-statistic whose null distribution is $F_{1,2n-1}$? Justify your answer.

Problem 6 In a two-way ANOVA one can study how two factors jointly affect the mean of an outcome variable. Consider a balanced situation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$$

for $i = 1, 2$, $j = 1, 2$, and $k = 1, 2$, where $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$.

(a) Is μ estimable? Explain.

(b) Show that $\delta_{11} - \delta_{12} - \delta_{21} + \delta_{22}$ is estimable. Let m denote the vector of sample means, i.e., $m = (\bar{Y}_{11.}, \bar{Y}_{12.}, \bar{Y}_{21.}, \bar{Y}_{22.})^T$, where $\bar{Y}_{ij.} = \frac{1}{2}\sum_{k=1}^{2} Y_{ijk}$. Show that the best linear unbiased estimator of $\delta_{11} - \delta_{12} - \delta_{21} + \delta_{22}$ is a linear function of m.

We further introduce the following constraints:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \delta_{ij} = \sum_j \delta_{ij} = 0$$

We can use inference on the $\delta_{ij}'s$ to investigate whether the two factors are additive. For example, we may be interested to test if the two factors are additive, i.e., there is no interaction between them, by considering

$$H_0 : \delta_{ij} = 0 \text{ for all } i = 1, 2; j = 1, 2$$

We can define an F-statistic based on the following sums of squares

$$\begin{aligned} SSE &= \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 \\ SSAB &= \sum_i \sum_j \sum_k (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \end{aligned}$$

(c) Express $SSAB$ in terms of m and some constants. Then show that $SSAB/\sigma^2$ follows a chi-squared distribution when H_0 is true.

(d) Construct an F-test for testing H_0. Please show all your work.

Finally, consider an augmented model that contains an additional variable Z, where $Z = (z_{111}, \cdots, z_{222})^T$ is a continuous variable:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \gamma z_{ijk} + \epsilon_{ijk}.$$

(e) Find the least squares estimator of γ.

Problem 7 Suppose $Y_{0i} \sim (\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$ and $Y_{1i} \sim (\mu_1, \sigma_1^2)$ for $i = 1, \ldots, n_1$, with $Cov(Y_i, Y_j) = 0$ for $i \neq j$ and $0 < \sigma_k^2 < \infty$ for $k = 0, 1$. Let $\overline{Y}_0 = \sum_{i=1}^{n_0} Y_{0i}/n_0$ and $\overline{Y}_1 = \sum_{i=1}^{n_1} Y_{1i}/n_1$ denote the sample means for each group, and let $s_0^2 = \sum_{i=1}^{n_0}(Y_{0i} - \overline{Y}_0)^2/(n_0 - 1)$ and $s_1^2 = \sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2/(n_1 - 1)$ denote the sample variances for each group, with pooled variance estimate $s_p^2 = ((n_0 - 1)s_0^2 + (n_1 - 1)s_1^2)/(n_0 + n_1 - 2)$. Suppose further that as $n_0 \to \infty$ and $n_1 \to \infty$, $n_0/(n_0 + n_1) \to \lambda$ for some known constant $0 < \lambda < 1$. We are interested in making inference about $\delta = \mu_1 - \mu_0$.

Under the assumption of equal variances (so $\sigma_0^2 = \sigma_1^2$), the usual statistical inference for this problem is based on the two sample t test for independent samples assuming equal variances using test statistic $T_e = (\overline{Y}_1 - \overline{Y}_0)/(s_p\sqrt{1/n_0 + 1/n_1})$ and assuming that T_e is distributed according to the t distribution with $n_0 + n_1 - 2$ degrees of freedom under the null hypothesis $H_0 : \delta = 0$. A confidence interval is constructed by inverting that test statistic.

Under the assumption of unequal variances (so $\sigma_0^2 \neq \sigma_1^2$), the usual statistical inference for this problem is based on the two sample t test for independent samples assuming unequal variances using test statistic $T_u = (\overline{Y}_1 - \overline{Y}_0)/\sqrt{s_0^2/n_0 + s_1^2/n_1}$ and assuming that T_u is distributed according to the t distribution with k degrees of freedom under the null hypothesis $H_0 : \delta = 0$, where k might be determined by the Satterthwaite or Aspin-Welch approximations. A confidence interval is constructed by inverting that test statistic.

(a) Show that under the assumption of equal variances, the statistical inference based on T_e is asymptotically correct in that the size of the hypothesis test is asymptotically at the correct level and the confidence interval has the correct coverage probability asymptotically.

(b) Show that under the assumption of equal variances, the statistical inference based on T_u is asymptotically correct.

(c) Show that under the assumption of unequal variances, the statistical inference based on T_u is asymptotically correct.

(d) Show that under the assumption of unequal variances, the statistical inference based on T_e is not necessarily asymptotically correct. Under what conditions will inference based on T_e be asymptotically valid in this setting? Under what conditions will it be conservative? anti-conservative?

(e) What do the above results suggest about the validity of regression based on linear regression models in the presence of heteroscedasticity?

Table 1: Common distributions and densities.

Distribution	Notation	Density
Bernoulli	$\text{Bern}(\theta)$	$f(y\|\theta) = \theta^y(1-\theta)^{1-y}$
Binomial	$\text{Bin}(n,\theta)$	$f(y\|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$
Multinomial	$\text{Multi}(n;\theta_1,\theta_2,\ldots,\theta_K)$	$f(y\|\theta) = \frac{n!}{y_1!y_2!\ldots y_K!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_K^{y_K}$
Beta	$\text{Beta}(a,b)$	$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)$
Uniform	$U(a,b)$	$p(\theta) = \frac{I_{(a,b)}(\theta)}{b-a}$
Poisson	$\text{Pois}(\theta)$	$f(y\|\theta) = \theta^y e^{-\theta}/y!$
Exponential	$\text{Exp}(\theta)$	$f(y\|\theta) = \theta e^{-\theta y}I_{(0,\infty)}(y)$
Gamma	$\text{Gamma}(a,b)$	$p(\theta) = [b^a/\Gamma(a)]\theta^{a-1}e^{-b\theta}I_{(0,\infty)}(\theta)$
Chi-squared	$\chi^2(n)$	Same as $\text{Gamma}(n/2, 1/2)$
Weibull	$\text{Weib}(\alpha,\theta)$	$f(y\|\theta) = \theta\alpha y^{\alpha-1}\exp(-\theta y^\alpha)\, I_{(0,\infty)}(\theta)$
Normal	$N(\theta, 1/\tau)$	$f(y\|\theta,\tau) = (\sqrt{\tau/2\pi})\exp[-\tau(y-\theta)^2/2]$
Student's t	$t(n,\theta,\sigma)$	$f(y\|\theta) = [1+(y-\theta)^2/n\sigma^2]^{(n+1)/2}$
		$\times\, \Gamma[(n+1)/2]/\Gamma(n/2)\sigma\sqrt{n\pi}$
Cauchy	$\text{Cauchy}(\theta)$	same as $t(1,\theta,1)$
Dirichlet	$\text{Dirichlet}(a_1,a_2,a_3)$	$p(\theta) = \Gamma(a_1+a_2+a_3)/\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)$
		$\times\, \theta_1^{a_1-1}\theta_2^{a_2-1}(1-\theta_1-\theta_2)^{a_3-1}$
		$\times\, I_{(0,1)}(\theta_1)I_{(0,1)}(\theta_2)I_{(0,1)}(1-\theta_1-\theta_2)$

Table 2: Means, Modes, and Variances.

Distribution	Mean	Mode	Variance
Bern(θ)	θ	0 if $\theta < .5$ 1 if $\theta > .5$	$\theta(1-\theta)$
Bin(n, θ)	$n\theta$	integer closest to $n\theta$	$n\theta(1-\theta)$
Beta(a, b)	$a/(a+b)$	$(a-1)/(a+b-2)$ if $a > 1, b \geq 1$	$ab/(a+b)^2(a+b+1)$
$U(a, b)$	$.5(a+b)$	everything a to b	$(b-a)^2/12$
Pois(θ)	θ	integer closest to θ	θ
Exp(θ)	$1/\theta$	0	$1/\theta^2$
Gamma(a, b)	a/b	$(a-1)/b$ if $a > 1$	a/b^2
$\chi^2(n)$	n	$n-2$ if $n > 2$	$2n$
Weib(α, θ)	$\Gamma[(\alpha+1)/\alpha]/\theta$	$[(\alpha-1)/\alpha]^{1/\alpha}/\theta$	$\Gamma[(\alpha+2)/\alpha] - \mu^2$
$N(\theta, 1/\tau)$	θ	θ	$1/\tau$
$t(n, \theta, \sigma)$	θ if $n \geq 2$	θ	$\sigma^2 n/(n-2)$ if $n \geq 3$
Cauchy(θ)	Undefined	θ	Undefined