Statistics 24400 - Winter 2016

Final Examination

March 11,14,15, 2015

Name (print): ______________________________

On my honor, I will not discuss this exam with **ANY PERSON** before 16:30 March 15, 2015.

Signature ______________________________

1. Please **print** your name in the space provided. If you are taking this exam on the 11th or 14th, you must sign the "temporary nondisclosure" line and conduct yourself accordingly in order to get credit for this final exam.

2. Do not sit directly next to another student.

3. Do not turn the page until told to do so.

4. This is a closed book examination. You are allowed a single page of notes, written on both sides. Please write your name on your notes and turn it in with the exam. You are permitted to have a calculator. Devices capable of communication (laptops, tablets, phones) must be powered down. Tables of the cumulative Normal and χ^2 distributions are at the end of the exam.

5. Please provide the answers in the space and blank pages provided. If you do not have enough space, please use the back of a nearby page, clearly indicating the identity of the continued problem.

6. Be sure to show your calculations. In order to receive full credit for a problem, you must show your work and explain your reasoning. Good work can receive substantial partial credit even if the final answer is incorrect.

7. Read through the exam before answering any questions. Our scale of credit for questions may not correlate with the level of difficulty you experience—use your time wisely!

Question	*Points*	*Score*
Question 1	20	
Question 2	20	
Question 3	20	
Question 4	20	
Question 5	20	
TOTAL	100	

1. **True or False?** (20 pts) As with the quiz, there is a "guessing penalty", but this time it is four questions (8 pts). You'll begin to accrue points with your **fifth** correct answer.

   (a) (2 pts) T F
   If random variables X and Y are uncorrelated (i.e. $\text{Cor}(X, Y) = 0$), then X and Y must be independent. **F**. Lack of correlation does not imply independence. Example: let X be a random var $X \sim N(0, 1)$, and let W be a random var that takes on the values +1 and -1 with probability 0.5. Then $Y = WX$ is not independent of X because $|Y| = |X|$, but X and Y are uncorrelated (you can check this).

   (b) (2 pts) T F
   If $P(A) < P(B)$ and $P(C) > 0$, then $P(A|C) \leq P(B|C)$. **F**. Examples in Lecture 6 and elsewhere,

   (c) (2 pts) T F
   The calculation of p-value does not depend on the alternative hypothesis once we know the null hypothesis. **T**. It only depends on α, not β.

   (d) (2 pts) T F
   The maximum likelihood estimator (MLE) is always unbiased. **F**. See Lecture 8.

   (e) (2 pts) T F
   For any hypothesis testing procedure, it is always possible to increase the power $\pi = 1 - \beta$ while keeping the type 1 error α the same. **F**. Look at p. 2, Lecture 14.

   (f) (2 pts) T F
   If X is a continuous random variable with cdf $F(X)$, and Y is a random variable such that $Y = F(X)$, then Y is distributed uniformly on $[0, 1]$. **T**. Look at Lecture 3, pp. 14-15.

   (g) (2 pts) T F
   $\Gamma(\frac{3}{2}) < \Gamma(\frac{1}{2})$ **T**. $\Gamma(n + 1) = n\Gamma(n)$. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

   (h) (2 pts) T F
   Fisher's method of combination on a set of tests of the same hypothesis with p-values $p_1, p_2, \ldots p_k$ means that the p-value of the combined tests is given by $P = p_1 p_2 \ldots p_k$. **F**. See Lecture 17,

   (i) (2 pts) T F
   P defined as above. Then $P \sim \chi^2_{2k}$. **T** See Lecture 17.

   (j) (2 pts) T F
   Suppose that $X_i \sim \mathcal{N}(\mu, \sigma^2)$, with μ and σ unknown. Then the distribution of $X_1 - \mu$ does not depend on any unknown parameter. **F**. It still depends on σ^2.

   (k) (2 pts) T F
   As above, suppose that $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ unknown, Then the distribution of $\frac{X_1 - X_2}{X_3 - X_4}$ does not depend on any unknown parameter. **T**. All $X_i \sim N(\mu, \sigma^2)$, so we can write each of them as $\sigma Z_i - \mu$, where $Z_i \sim N(0, 1)$ for $i = 1, 2, 3, 4$. Then the μ's are subtracted from each other and the σ's cancel.

(l) (2 pts) T F
For a Poisson process, **conditional on the number of events** $N(0,1] = n$, the number of events

$$N(0,1/3] \sim \text{Bin}(n, 1/3) \quad \mathbf{T}$$

Because the event in a Poisson process are independent, conditioning the whole interval does not condition a portion of it.

(m) (2 pts) T F
In order to do hypothesis testing on the median of rank ordered data, you'll need tables of the cumulative Binomial distribution. **T**. See Lecture 18 and Rice 10.4.2.

(n) (2 pts) T F
The proof of the Neyman-Pearson Lemma involves a Taylor expansion. **F**. See Lecture 14, pp. 3-5. It *does* use an indicator function that behaves as a Bernoulli random variable.

2. **Alice and Bob go to the Reg.** (20 pts). Alice and Bob are planning to meet at the Reg to work on STAT 244 homework together. Alice will arrive there at a random (uniform) time between 1:00 pm and 2:00 pm, while Bob will arrive at a random (uniform) time between 1:00 pm and 3:00 pm. Assume Alice and Bob arrive independently of one another.

   (a) What is the joint density of the arrival times of Alice and Bob?

   (b) What is the expected amount of time one person waits for the other?

   (c) What is the probability that Bob waits for Alice?

   Solution

   (a) Let A and B be the arrival times of Alice and Bob uniformly distributed on $[0, 1]$ and $[0, 2]$, respectively. Since A and B are independent, their joint density is $f_{A,B}(a, b) = f_A(a)f_B(b) = \frac{1}{2}$ if $(a, b) \in [0, 1] \times [0, 2]$ and 0 otherwise.

   (b) Consider $\min(A, B)$ (the arrival time of whoever arrives first) and $\max(A, B)$ (the arrival time of whoever arrives second). Then the waiting time is $\max(A, B) - \min(A, B)$. Using the joint density from part (a), the expected waiting time is

$$\int_0^1 \int_0^2 (\max(A, B) - \min(A, B)) f_{A,B}(a, b) \; \mathrm{d}a\mathrm{d}b =$$

$$\int_0^1 \int_0^a \frac{1}{2}(a - b) \; \mathrm{d}b\mathrm{d}a + \int_0^1 \int_a^2 \frac{1}{2}(b - a) \; \mathrm{d}b\mathrm{d}a = \frac{1}{12} + \frac{7}{12} = \frac{2}{3},$$

   or 40 minutes.

   (c) Integrating the joint density over the region where $b < a$, the probability that Bob waits for Alice is

$$\mathbb{P}(B < A) = \int_0^1 \int_0^a \frac{1}{2} \; \mathrm{d}b\mathrm{d}a = \frac{1}{4}.$$

3. **Test for the Binomial Distribution** (20 pts)
Nylon bars were tested for brittleness. Each of 280 bars was molded under similar conditions and was tested in five places. Assuming that each bar has uniform composition, the number of breaks on a given bar should be Binomially distributed with five trials and an unknown probability p of failure. If the bars are all of the same uniform strength, p will be the same for all of them; if they are of different strengths, p will vary from bar to bar. Thus, the null hypothesis is that the ps are all equal. The following table summarizes the outcome of the experiment:

Breaks/Bar	Frequency
0	157
1	69
2	35
3	17
4	1
5	1

(a) Under the given assumption, the data in the table consist of 280 observations of independent binomial random variables. Find the MLE of p.

(b) Pooling the last three cells, calculate Pearson's chi-square test statistic. How many degrees of freedom are there?

(c) The density for $X \sim \chi^2_k$ is $f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$. Use this fact to obtain an *exact* expression for the p-value.
Hint: if you don't see something that's easy to integrate, check your degrees of freedom again.

(d) With a significance level 0.05, what is your conclusion from the test?

4. **Test for the Binomial Distribution—Solution** (20 pts)

(a) Denote X_i as number of breaks on bar i, $i = 1, \dots, 280$ and p_i as failure probability for each bar. Then $X_i \sim \text{Binomial}(5, p_i)$.
Under the null hypothesis H_0: all p_i are equal, the likelihood function is

$f(x_1, \dots, x_{280}|p) = P(X_1 = 0)^{157} P(X_1 = 1)^{69} P(X_1 = 2)^{35} P(X_1 = 3)^{17} P(X_1 = 4)^1 P(X_1 = 5)^1$

so the log-likelihood is

$$\begin{aligned} l(p) &= (5 \times 157 + 4 \times 69 + 3 \times 35 + 2 \times 17 + 1) \log(1-p) \\ &+ (69 + 2 \times 35 + 3 \times 17 + 4 + 5) \log p + \text{Constant} \\ &= 1201 \log(1-p) + 199 \log p + \text{Constant} \end{aligned}$$

So MLE is $\hat{p} = \frac{199}{1400}$

(b) Under binomial distribution and null hypothesis, the expected counts in each category can be calculated as

$$\begin{aligned} E_i &= 280 \times P(X_1 = i) = 280 \times \binom{5}{i} \hat{p}^i (1-\hat{p})^{5-i} \qquad i = 0, 1, 2 \\ E_{3+} &= 280 \times P(X_1 \geq 3) = 280 - \sum_{i=0}^{2} E_i \end{aligned}$$

So we obtain the following table:

	0	1	2	≥ 3
O_i	157	69	35	19
E_i	130	108	36	6

Pearson's chi-square statistic is $X^2 = \sum_{i=0}^{3} \frac{(O_i - E_i)^2}{E_i} \approx 47.9$ which approximately distributed as chi-square with degree of freedom 2.

(c) Chi-square distribution with 2 degree of freedom is Gamma distribution with parameters $(\alpha = 1, \lambda = 1/2)$ which is Exponential(1/2). So

$$\begin{aligned} \text{p-value} &= P(\chi_2^2 > 47.9) \\ &= P(\text{Exp}(1/2) > 47.9) \\ &= e^{-\frac{47.9}{2}} \\ &\approx 4 \times 10^{-11} \end{aligned}$$

(d) From distribution table, we know p-value must be smaller than 0.005, since $X^2 \approx 47.9 > \chi^2_{.995}$. Even though you haven't worked out (c), you can do this part by comparing test statistic with critical value.
In either way, we reject the null hypothesis.

5. **The Pareto Distribution** (20 pts)
The Pareto Distribution has the density

$$f(x \mid x_0, \theta) = \theta x_0^\theta x^{-\theta-1}, \quad x \geq x_0, \ \theta > 1$$

Assume $x_0 > 0$ is given and that $X_1, X_2, ...X_n$ are i.i.d. samples from this distribution. Then

(a) Find the MLE of θ.

(b) Find the asymptotic variance of the MLE.

(c) Find a sufficient statistic for θ.

Solution.

(a)

$$\begin{aligned} f(x/\theta) &= \theta x_0^\theta x^{-\theta-1} \\ L(\theta) &= \theta^n x_0^{n\theta} \left(\prod_{i=1}^n x_i \right)^{-\theta-1} \\ l(\theta) &= n \log \theta + n\theta \log x_0 - (\theta+1)(\sum_{i=1}^n \log x_i) \\ l'(\theta) &= \frac{n}{\theta} + n \log x_0 - (\sum_{i=1}^n \log x_i) \end{aligned}$$

Equating this to zero, we get the MLE to be

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \log(x_i/x_0)}$$

(b)

$$\frac{\mathrm{d}^2 \log f(x_1|\theta)}{\mathrm{d}\theta^2} = \frac{\mathrm{d}}{\mathrm{d}\theta}(\frac{1}{\theta} + \log x_0 - \log x_1) = -\frac{1}{\theta^2}$$

So the fisher information, $I(\theta) = \frac{1}{\theta^2}$ and the asymptotic variance of $\hat{\theta}$ is $\frac{1}{nI(\theta)} = \frac{\theta^2}{n}$.

(c)

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^n x_0^{n\theta} \left(\prod_{i=1}^n x_i \right)^{-\theta-1}$$

So by factorization theorem, the sufficient statistic is $\prod_{i=1}^n X_i$.

6. **Some p-values in the literature.** (20 pts) When certain conditions are met, the p-value from a hypothesis test has Uniform$(0, 1)$ distribution under the null hypothesis.

The following is a table of p-values from 10,000 instances of a hypothesis test.

p-values	Observed Counts	Expected Counts
$0 \le p \le .001$	3,859	
$.001 < p \le .005$	972	
$.005 < p \le .01$	1,116	
$.01 < p \le .02$	712	
$.02 < p \le .03$	539	
$.03 < p \le .04$	423	
$.04 < p \le .05$	445	
$.05 < p \le .10$	370	
$.10 < p \le 1$	1,564	
Total	10,000	

We would like to test the hypothesis that these data indeed follow a Uniform$(0, 1)$ distribution.

(a) Fill out the last column of the table under the null hypothesis that the data follows a Uniform$(0, 1)$ distribution.

(b) Write out (but do not compute) the expression for the test statistic χ^2. State the degrees of freedom for the test statistic χ^2, and explain your answer.

(c) Is the null hypothesis supported? Please reason in a precise and quantitative manner using the Table in this exam, computing as necessary.

(a) If $V \sim$ Uniform$(0, 1)$, then $\Pr\{V \in (a, b]\} = b - a$ for any $(a, b] \subseteq [0, 1]$.

p-values	Observed Counts	Expected Counts
$0 \le p \le .001$	3,859	10
$.001 < p \le .005$	972	40
$.005 < p \le .01$	1,116	50
$.01 < p \le .02$	712	100
$.02 < p \le .03$	539	100
$.03 < p \le .04$	423	100
$.04 < p \le .05$	445	100
$.05 < p \le .10$	370	500
$.10 < p \le 1$	1,564	9,000
Total	10,000	10,000

(b)

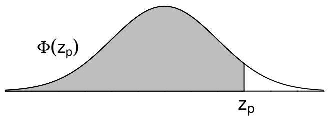
$$X^2 = \sum_{\text{cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

(It is acceptable to write this out in terms of actual values.) The degrees of freedom is 8, as there are 9 cells, and no parameter has been estimated from the data.

(c) The values for χ^2 with 8 degrees of freedom is not listed in the table. But observe that $\chi^2_8(.99) < \chi^2_{10}(.99)$. Each summand of X^2 is strictly positive, so we have

$$X^2 > \frac{(3859-10)^2}{10} = \frac{3849^2}{10} > \frac{3000^2}{10} = 900000 > \chi^2_{10}(.99) > \chi^2_8(.99).$$

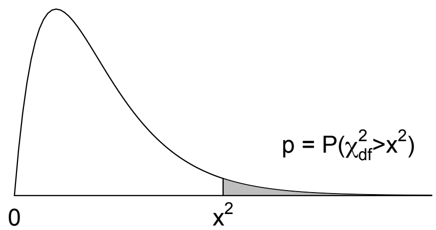
Thus, we reject the hypothesis that the data follows Uniform$(0,1)$ distribution.



Cumulative Normal Distribution

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Also for $z_p = 4.0$, 5.0, and 6.0, the values of p are 0.99997, 0.9999997, and 0.999999999.



χ^2 Distribution

df	Tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2