STATISTICS 641 - Exam 3

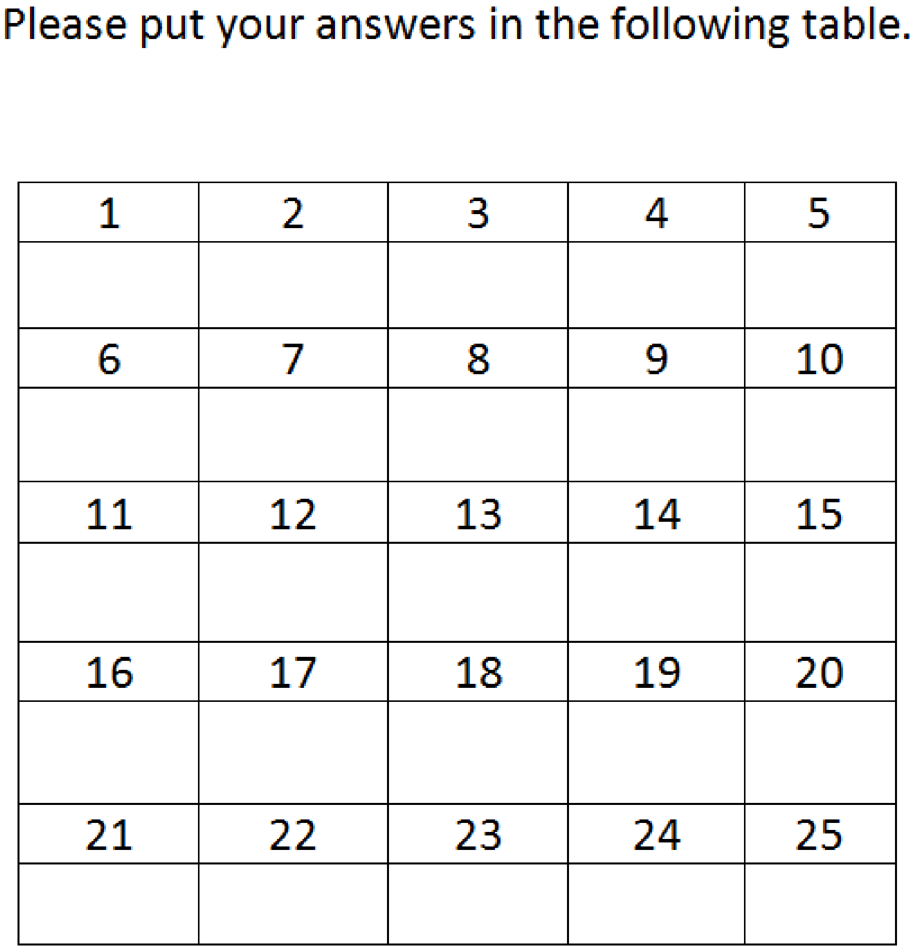
Total time is 120 minutes. The exam is available for 24 hours window, starting from noon (CST) May 07, 2015 to 11:59am (CST) May 08, 2015.

Instructions

1. The exam is for two hours.
2. Students are allowed to bring four pages of cheat sheets (front&back, front&back, front&back and front&back).
3. Students are allowed to use three pages (other than the cheat sheets) to do scratch work.
3. Students are allowed to use calculator and R.

Name ____

Email Address ____



Please put your answers in the following table.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

STATISTICS 641 - Exam #3

1. Suppose that we have four random samples from four different distributions. To test equality of the variances we should apply

   (a) Brown-Forsythe-Levene •

   (b) Kaplan-Meier test

   (c) F-test

   (d) t-test

   (e) χ^2 test

2. Suppose that we have data $X_1, \dots, X_T$ from the following model $X_t = \rho X_{t-1} + e_t$, e_t are iid Normal$(0, \sigma^2)$. The most reasonable estimator of the lag-one auto-correlation is

   (a) $1 - \sum_{t=2}^{T}(X_t - \overline{X})(X_{t-1} - \overline{X}) / \sum_{t=1}^{T}(X_t - \overline{X})^2$

   (b) $1 - \sum_{t=2}^{T} X_t X_{t-1} / \sum_{t=1}^{T}(X_t - \overline{X})^2$

   (c) $\sum_{t=2}^{T} X_t X_{t-1} / \sum_{t=1}^{T}(X_t - \overline{X})^2$

   (d) $\sum_{t=1}^{T}(X_t - \overline{X})^2 / \sum_{t=2}^{T}(X_t - \overline{X})(X_{t-1} - \overline{X})$

   (e) $\sum_{t=2}^{T}(X_t - \overline{X})(X_{t-1} - \overline{X}) / \sum_{t=1}^{T}(X_t - \overline{X})^2$ •

3. Suppose that we want to test equality of two population means while the two populations have the same shape. The appropriate method of testing the hypothesis is

   (a) pooled t-test

   (b) Wilcoxon Signed rank test •

   (c) Kolmogorov-Smirnov's test

   (d) Wilcoxon Rank sum test

   (e) Brown-Forsythe-Levene

   (f) Breslow test

Suppose that we have two random samples $X = (0.19, -1.15, -0.06, 1.42, 0.15, 1.16, -0.73, -0.78, -0.23, -0.76)$ and $Y = (0.81, 0.54, 0.44, 0.04, 0.05, 1.24)$ from two distributions, and $\overline{X} = -0.079$, $s_x^2 = 0.71734$, $n_x = 10$, $\overline{Y} = 0.52$, $s_y^2 = 0.21212$, and $n_y = 6$. Assume that the two distributions approximately normally distributed, and assume that the variances for the two groups are the same. Answer the next four questions.

4. The pooled variance estimate is

   (a) 0.537 •

   (b) 0.469

   (c) 0.501

   (d) 0.268

   (e) 0.656

   Hint: $s_p^2 = \{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2\}/(m + n - 2) = 0.5369$

5. Assume that the pooled variance is 0.84. The absolute value of the test statistic for testing equality of the two means is

   (a) 2.674

   (b) 2.451

   (c) 1.265 •

   (d) 1.172

   (e) 2.292

Hint: $T = (\overline{X} - \overline{Y})/s_p\sqrt{1/m + 1/n} = -1.265$.

6. In order construct a confidence interval for the variance ratio σ_x^2/σ_y^2 we need to use percentile from

   (a) χ^2 distribution with degrees of freedom 14

   (b) χ^2 distribution with degrees of freedom 15

   (c) F distribution with degrees of freedom $(9, 5)$ •

   (d) F distribution with degrees of freedom $(8, 4)$

   (e) t distribution with degrees of freedom 14

   (f) Z distribution

7. Suppose the sample sizes are equal $n_x = n_y = n$. Determine the value of n so that there is at least 90% power to detect the difference $|\mu_1 - \mu_2| = \sqrt{2}\sigma$ at the 5% level of significance.

   (a) 35

   (b) 16

   (c) 15

   (d) 8

   (e) 11 •

   Hint: HO13, $n = 2\{\sigma(Z_\alpha + Z_\beta)/\delta\}^2 = 2\{(1.96 + 1.28)/\sqrt{2}\}^2 = 10.49$

   Suppose that based on the sample data $X = (X_1, \ldots, X_{10})$, we want to test $H_0 : \tilde{\mu} = 0.25$ against $H_a : \tilde{\mu} > 0.25$, where $\tilde{\mu}$ is the median of the distribution. We decide to reject H_0 if $S+ > 8$. When the median is 0.25, $\text{pr}(X_i > 0.25) = 0.5$ and $\text{pr}(X > 0.45) = 0.28$, and when the median is 0.45, $\text{pr}(X_i > 0.25) = 0.68$ and $\text{pr}(X > 0.45) = 0.5$. Answer the next three questions.

8. Find the probability of type-I error.

   (a) 0.005

   (b) 0.12

   (c) 2.95×10^{-5}

   (d) 0.05

   (e) 0.01

   pr(Type-I error) $= \text{pr}(S+ > 8 | \tilde{\mu} = 0.25) =$ `1-pbinom(8, 10, 0.5)`$= 0.0107421$.

9. Refer to the previous question. Under H_0, and when $n = 10$, $\text{var}(S+)$ is

   (a) $0.5 \times 0.5/10$

   (b) $0.25 \times 0.75/9$

   (c) $0.25 \times 0.75/10$

   (d) $10/2$

   (e) $10/4$ •

   Hint: $\text{var}(S+) = np(1 - p) = n/4$.

10. What is probability of Type-II error when $\tilde{\mu} = 0.45$ and $n = 10$?

    (a) 0.064

    (b) 0.121

    (c) 0.879 •

    (d) 0.724

    (e) 0.145

Hint: pr(Type-II error) $=$ pr$(S+ \leq 8|\tilde{\mu} = 0.45) =$`pbinom(8, 10, 0.68)`0.879

A social scientist wants to determine if there is an association between political affiliation and annual income level. A random sample of 592 registered voters is selected and each of the selected individuals was asked their political affiliation and their annual income. These values were then summarized in the following table:

	Income level				Total
	$[0, 30k)$	$[30k, 50k)$	$[50k, 100k)$	$[100k, \infty)$	
Republican	20	84	17	94	215
Democrat	68	119	26	7	220
Independent	20	83	28	26	157
Total	108	286	71	127	592

Define $[100k, \infty)$ as the high income and $[0, 100k)$ as the not-high income group. Let p_{hi} and p_{nhi} be the proportion of republicans in the high and not-high income grops. Answer the next four questions.

11. What is the expected frequency of the cell "Republican and income level $[50k, 100k)$" under the assumption of independence of political affiliation and income level?

(a) 17

(b) 32

(c) 28

(d) 22

(e) 26 •

Hint: Under independence $592 \times (215/592)(71/592) = 25.78$

12. What is the degrees of freedom of the χ^2-test of independence?

(a) 4

(b) 5

(c) 10

(d) 6 •

(e) 11

Hint: $(r-1)(c-1) = (3-1)(4-1) = 2 \times 3 = 6$

13. The test statistic for testing $H_0 : p_{hi} = p_{nhi}$ is

(a) 0.654

(b) 1.211

(c) 2.212

(d) 5.543

(e) 9.96 •

Hint: $\widehat{p}_{hi} = 94/127$, $\widehat{p}_{nhi} = 121/465$. Now the test statistic is $T = (\widehat{p}_{hi} - \widehat{p}_{nhi})/\sqrt{\{p(1-p)(1/m + 1/n)\}}$, where $p = (94 + 121)/(127 + 465) = 0.3631$.

14. Find a 95% confidence interval for $p_{hi} - p_{nhi}$.

(a) $(0.552, 0.745)$

(b) $(0.323, 0.636)$

(c) $(0.362, 0.597)$

(d) $(0.393, 0.566)$ •

(e) $(0.407, 0.552)$

Hint:

$$\widehat{p}_{hi} - \widehat{p}_{nhi} \pm 1.96\sqrt{\frac{\widehat{p}_{hi}(1-\widehat{p}_{hi})}{m} + \frac{\widehat{p}_{nhi}(1-\widehat{p}_{nhi})}{n}},$$

and $\widehat{p}_{hi} = 94/127$, $\widehat{p}_{nhi} = 121/465$.

To know the efficacy of a new cost effective testing method to detect presence of E. coli, this testing method was applied on meat samples collected from different supermarkets. Also, these samples were tested for E. coli using an almost accurate method (we shall treat it as gold standard), and the results are given in the following table.

	Test result of the cost effective method	
E.coli status	+	−
Present	140	5
Absent	70	800

Assume that the chance that E. coli is present in a food sample is 0.08. Answer the next two questions.

15. Estimate the probability that a randomly chosen meat sample will be tested positive for E. coli bacteria using the given information?

    (a) 0.965
    (b) 0.151 •
    (c) 0.08
    (d) 0.206
    (e) 0.226

    Hint: pr(+) = pr(+|E. coli present)pr(E. coli present)+pr(+|E. coli absent)pr(E. coli absent) = (140/145)(0.08)+(70/870)(0.92) = 0.1512.

16. Estimate the probability that a meat sample actually contains E. coli bacteria given that the test result using the cost effect approach was negative?

    (a) 0.005
    (b) 0.126
    (c) 0.001
    (d) 0.003 •
    (e) 0.034

    Hint: pr(E. coli present|−) = pr(E. coli present ∩ −)/pr(−) = pr(−|E. coli present)pr(E. coli present)/pr(−) = (5/145)(0.08)/(1 − 0.1512) = 0.003.

An EPA researcher wants to design a study to estimate the mean lead level of fish in a lake near an industrial area. Based on the past sample data, the researcher estimate σ for the lead level in the fish population is approximately 0.016mg/g. He wants to use a 98% CI (two sided CI) having a margin of error no greater than 0.005mg. Answer the next four questions.

17. How many fish does he need to catch?

    (a) 56 •
    (b) 44
    (c) 222
    (d) 14
    (e) 166

    Hint: $n = (Z_\alpha \sigma / L)^2 = (2.326 \times 0.016/0.005)^2 = 55.4$

18. Based on the data from 60 randomly chosen fish display the formula for a 95% distribution-free confidence interval for the median lead level in the fish.

(a) $(Y_{(25)}, Y_{(36)})$

(b) $(Y_{(26)}, Y_{(44)})$

(c) $(Y_{(20)}, Y_{(40)})$

(d) $(Y_{(20)}, Y_{(41)})$

(e) $(Y_{(22)}, Y_{(39)})$ •

Hint: HO11, We need to find r such that $\text{pr}(Y_{(r)} \leq B \leq Y_{(n-r)}\} \geq 0.95$, where $B \sim \text{Binomial}(60, 0.5)$.

```
n=60
p=0.5
r=0
eps=0.99
while(eps>0.95){
r=r+1
eps= pbinom((n-r), n, p)-pbinom((r-1), n, p)
}
r=23
pbinom((n-r), n, p)-pbinom((r-1), n, p)
[1] 0.9481061
> r=22
> pbinom((n-r), n, p)-pbinom((r-1), n, p)
[1] 0.9726599
```

So, $r = 22$ and $n - r = 60 - 22 = 38$, and the CI is $(Y_{(22)}, Y_{(39)})$

19. Suppose the lead level in a fish approximately follows Gamma(6, 6) distribution so that its mean and variance are 1 and 1/6, respectively. Use the sampling distribution theory of the sample median ($\widehat{m}$), and construct a 95% CI for the median lead level (m) in the fish. Assume that for the sample with size $n = 60$, the sample median was 0.84mg/g. (Hint: $\text{var}(\widehat{m}) = 0.25/nf^2(m)$, where $f(m)$ denotes the density of random variable (lead level) evaluated at m)

    (a) $(0.823, 0.847)$

    (b) $(0.836, 0.847)$

    (c) $(0.719, 0.961)$ •

    (d) $(0.778, 0.901)$

    (e) $(0.755, 0.940)$

    Hint: HO10, Let m be the population median, and $\widehat{m}$ be the sample median. Then for large sample $\sqrt{n}(\widehat{m} - m) \sim \text{Normal}\{0, 0.25/f^2(m)\}$, where $f(m)$ is the Gamma density at m. Then 95% CI will be $[\widehat{m} \pm 1.96 \times 0.5/\sqrt{n}f(m)] \approx [\widehat{m} \pm 1.96 \times 0.5/\sqrt{n}f(\widehat{m})]$. So, the interval is $[0.84 \pm 1.96 \times 0.5/(\sqrt{n} \times 1.052637)] = (0.719, 0.960)$

20. Suppose that in the above question we want to test if the data actually follows a Gamma distribution. Which of the following will be the best way to test that assumption?

    (a) Kolmogorov-Smirnov test

    (b) McNemar's test

    (c) von Neumann Test

    (d) Anderson Darling •

    (e) Shapiro-Wilk's test

    (f) Breslow test

21. Which of the following statements is true when we have a random sample of n iid observations from a population F with finite mean and variance.

    (a) Sample mean is not necessarily unbiased for the population mean.

    (b) Sample mean is unbiased for the population mean only for a normal distribution.

(c) Sample median is always unbiased for the population median.

(d) Sample standard deviation is always unbiased for the population standard deviation.

(e) Sample variance is always unbiased for the population variance. •

22. Suppose that the incidence rate of HIV in the USA is 4% among adult men (15-49 years old). When 500 adult male are sampled at random, what is the probability that the sample proportion of HIV infected people will be less than 0.02? You should use normal approximation.

(a) 0.089

(b) 0.056

(c) 0.224

(d) 0.011 •

(e) 0.044

Hint: $\widehat{p} \sim \text{Normal}(0.04, 0.00876^2)$. So, $\text{pr}(\widehat{p} \leq 0.02) = \text{pr}\{Z \leq (0.02 - 0.04)/0.00876\} = 0.011$.

Suppose that p_G and $p_{\overline{G}}$ denote the probability of developing breast or ovarian cancer among the carrier of BRCA gene and non-carrier of BRCA gene, respectively. Consider the following table and answer the next 3 questions.

	Carrier status	
Disease	Yes	No
Cancer	40	450
No cancer	25	800

23. The odds of the disease among the BRCA carrier is (use numbers)

(a) 0.47

(b) 1.60 •

(c) 0.56

(d) 1.79

(e) 2.84

Hint: $p_G/(1 - p_G) = 40/25 = 1.6$

24. Suppose that we want to test if the disease and BRCA genes are associated. The null hypothesis is

(a) H_0: disease and BRCA gene are independent •

(b) H_0: disease and BRCA gene are dependent

25. To test the independence of the disease and BRCA genes we may use

(a) Agresti-Coull test

(b) Fisher's exact test •

(c) Sign test

(d) χ^2 test •

(e) McNemar's test