415 Final

1. Do not open until told to do so. When you start check that you have 4 sheets, 8 pages.
2. Write T or F inside the box for multiple choice questions.
3. Write response for other questions on separate paper. Leave 2 inch on top for staple.
4. Write your name below in print.
5. Staple problem sheet and response together and submit.
6. Pick up your HWs, quizes, and midterms. I will discard them otherwise.
7. Leave last row for me to sit.
8. If you have doubts about the exam, you have to let me know today during the exam.
9. If the quantiles given in the problem are not correct, write what you need in symbols and explain how you would use it.
10. I will not upload solutions. Once I finish grading, I will put the papers in my mailbox at 325 Thomas. Take your paper there. Beware, mail room will close during Winter, so take them fast. Post grading error concerns in the canvas forum, so everybody else can see. Do this fast as well. I will admit or not admit my errors. For ones I do admit my error, you need to send photos for proof. If you don't, score stays as is. If I erroenously give you points, you can still let me know. I will not take off points, I will just give points to other students as well. Grading is not comparative so, you are not at risk of getting lower grade due to others getting higher grade.
11. This paper will be uploaded to my website after the exam has finished by end of day.

MATH or STAT (circle one)

Name: ____________________________________

1. (18 pts) Answer T (true) or F (false) to the following statements. No need to show intermediate derivations.

1. Lets consider null $\theta = \theta_0$, alternative $\theta = \theta_1$. To find the best critical region, look for the region with large likelihood ratio $\frac{L(\theta_1)}{L(\theta_0)}$. $\square$
2. A test that uses the best critical region as rejection region might not be the most powerful test. Note among all tests with level α, the most powerful test achieves the highest power. $\square$
3. You can always find a uniformly most powerful test, even if the test is two sided. $\square$
4. $\frac{(Y_1-np_1)^2}{np_1(1-p_1)} = \frac{(Y_1-np_1)^2}{np_1} + \frac{(Y_2-np_2)^2}{np_2}$. Y_1 is $Bin(n, p_1)$. $Y_2 = n - Y_1$. $p_2 = 1 - p_1$. $\square$
5. You conduct test with a chi square test statistic. You use $\alpha = 0.05$. The statistic had value smaller than the degrees of freedom of chi square. You can conclude that you fail to reject the null hypothesis. $\square$
6. Lets say we perform h experiments each has k possible outcomes. Denoting probability of obtaining outcome i for experiment j as p_{ij}. We want to test the null that $p_{ij} = p_{ij'}$ for all $i,\ j,\ j'$. Under the null, $\sum_{j=1}^{h}\sum_{i=1}^{k} \frac{\left(y_{ij} - n_j \frac{\sum_{j=1}^{h} y_{ij}}{\sum_{j=1}^{h} n_j}\right)^2}{n_j \frac{\sum_{j=1}^{h} y_{ij}}{\sum_{j=1}^{h} n_j}}$ follows a chi square with $(h-1)(k-1)$ degrees of freedom. $\square$
7. Normality of data is not an assumption of ANOVA. $\square$
8. Common variance of data is not an assumption of ANOVA. $\square$
9. The means have to satisfy a certain condition to perform any ANOVA test. $\square$
10. You are only given the mean of each class in ANOVA. But you do not know how many samples there are in each class. You can still calculate the overall mean. $\square$
11. If you know two of SS_{tot}, SS_{err}, SS_{trt}, you can find out what the remaining one is. $\square$
12. If $SS_{trt} = 0$, $SS_{tot} = SS_{err}$. $\square$
13. SS_{tot}/σ^2, SS_{err}/σ^2, SS_{trt}/σ^2 all follows a chi square distribution. SS_{tot}/σ^2 always has the smallest degrees of freedom. $\square$
14. SS_{err}/σ^2 can have larger degrees of freedom than SS_{trt}/σ^2. $\square$
15. Sum of squares can be negative. $\square$
16. You reject the null hypothesis when the SS_{trt} is large compared to SS_{err}. $\square$

17. $\frac{\chi^2_{d_1}}{\chi^2_{d_2}}$ follows a F_{d_1,d_2} distribution. $\square$

18. Think of the ANOVA setting. $\frac{SS_{err}}{\sigma^2}$ will follow a chi squared only under the null. $\square$

2. (32 pts) For each of the following problems, answer T or F to each option. No need to show intermediate derivations.

1. We test $H_0: \ \theta = \theta_0$ against $H_1: \ \theta = \theta_1$. We consider tests with type I error probability α. We are to choose a rejection region C for this test. Denote the sample data as X.

   (a) 'tests with type I error probability α' means $P_{\theta_0}(X \in C) = \alpha$. $\square$

   (b) You want $P_{\theta_1}(X \in C) \geq P_{\theta_1}(X \in D)$ for all D s.t. $P_{\theta_0}(X \in D) = \alpha$. $\square$

   (c) If C satisfies (a) and (b), it is called the 'best critical region of size α'. $\square$

   (d) If we consider tests with type I error probability smaller than or equal to α, and looked for a test that has largest power among them, we will always result in a power larger than that for test with largest power among tests with type I error probability equal to α. $\square$

2. **Neyman-Pearson Lemma**. Let $X_1, \cdots, X_n$ be an i.i.d. random sample of size n from pdf or pmf $f(x, \theta)$. Let $H_0: \ \theta = \theta_0$ and $H_1: \ \theta = \theta_1$. Let $L(\theta) = f(x_1, \theta) \cdots f(x_n, \theta)$. If there exists a constant $k > 0$ and a subset C of the sample space s.t. $P_{\theta_0}[(X_1, \cdots, X_n) \in C] = \alpha$, $\frac{L(\theta_0)}{L(\theta_1)} \leq k$ when $(X_1, \cdots, X_n) \in C$, and $\frac{L(\theta_0)}{L(\theta_1)} \geq k$ when $(X_1, \cdots, X_n) \notin C$, then C is the best critical region of size α.

   (a) It is counter intuitive that we reject when likelihood of θ_0 over likelihood of θ_1 is smaller than k. $\square$

   (b) We will reject the null if $\frac{L(\theta_1)}{L(\theta_0)} \geq k$. $\square$

   (c) Ignore the condition that $P_{\theta_0}[(X_1, \cdots, X_n) \in C] = \alpha$. And take some arbitrary k' and reject null for for $\frac{L(\theta_0)}{L(\theta_1)} \leq k'$. The critical region for this test is best critical region for some size α'. $\square$

   (d) k can be negative. $\square$

3. Let us consider a single sample X from geometric distribution with density

$$(1-p)^{x-1}p$$

   for $x = 1, 2, 3, \cdots$ where $0 < p < 1$.

   (a) If we were testing $p = 0.5$ as null and $p > 0.5$ as alternative and a test was most powerful test against each simple alternative in H_1, the critical region C in this test is uniformly most powerful critical region of its size. $\square$

   (b) Lower value of X is evidence for smaller p. $\square$

(c) Tests of the form of $p = 0.5$ against $p = p_1$ where $p_1 > 0.5$ can have a critical region different from that for $p = 0.5$ against $p > 0.5$. ☐

(d) Even if you changed the test to $p = 0.1$ against $p > 0.1$, the uniformly most powerful test would not change. ☐

4. One set is a toss of two four sided die (demarcation from 1 to 4). You perform 16 sets. The sum of the dies was 2 for 1 of the time, 3 for 2 of the time, 4 for 3 of the time, 5 for 4 of the time, 6 fot 3 of the time, 7 for 2 of the time, and 8 for 1 of the time. We would like to perform a chi square goodness of fit test to investigate whether the two dies are fair.

   (a) The value of our test statistic is greater than 2. ☐

   (b) Under the null, the test statistic should follow a chi-square distribution with 7 degrees of freedom. ☐

   (c) If instead we got a result where all 16 attempts were 2, we would result in a higher test statistic and therefore would be more likely to reject the null. ☐

   (d) What if the null hypothesis was that the two dies are magically tied together, that is, the pair always show the same face, but are fair. The probability of seeing the data we saw (described in problem above), is 0. ☐

5. We would like to test whether the data we see is from a certain distribution. We partition the range of the distribution and calculate proportion falling in each. We also calculate proportion of sampled data falling in each.

   (a) You may not use a chi square goodness of fit approach. ☐

   (b) If you take only one partition (i.e. you do not partition) you will always accept the null (so you shouldn't do it this way). ☐

   (c) If you want to test that the data is from a certain distribution, with any parameters, you should first estimate the parameters from the data. You have to account for the estimated parameters in the degrees of freedom of the chi square. ☐

   (d) If you only take two partitions, divided at any true quantile, you will accept the null as long as the data has the same quantile as the distribution tested against. ☐

6. In both of two machines, there are k possible outcomes. We want to test whether the two machines outputs are given with same distribution (null).

   (a) If p_{ij} denotes probability of obtaining outcome i for experiment j, then the null written as an equation is $p_{i1} \neq p_{i2}$ for all i from 1 to k. ☐

   (b) We can think of a statistic $\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\left(y_{ij} - n_j \frac{y_{i1}+y_{i2}}{n_1+n_2}\right)^2}{n_j \frac{y_{i1}+y_{i2}}{n_1+n_2}}$. This will follow a chi squared with $2k - 2 - (k-1)$ degrees of freedom under the null. ☐

(c) Under the alternative, the statistic written in option (b) will tend to be small. $\square$

(d) Say you think the true proportions for both machines are some numbers q_{ij}. Lets say the null is that $p_{ij} = q_{ij}$ for all i and j. Then $\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{(y_{ij}-n_j q_{ij})^2}{n_j q_{ij}}$ follows a chi squared with $k-1$ degrees of freedom. $\square$

7. We want to test whether two sets of data has the same distribution (null).

(a) If we denote cdf of process generating first set of data as $F(x)$ and cdf of process generating second set of data as $G(x)$. The null is $F(x) = G(x)$ for all x. $\square$

(b) We partition the range into 3 parts. We calculate proportion that fell in each partition for each dataset. First data set had all samples come from the middle part. Second data set had all samples come from the left and right part. You will most likely accept the null. $\square$

(c) This test involves using a chi squared test statistic just like in the contingency table. The more partitions you take, the larger the degrees of freedom of this chi squared statistic. $\square$

(d) A similar approach can be taken for testing the null of h data generating processes having same distribution. If the number of partitions you take is the same, the chi squared statistic for this test will be smaller than that for the test of two data generating processes (h>2). $\square$

8. Let us think of the two way ANOVA setting. The first treatment has a possibilities. The second treatment has b possibilities. Each pair of treatment has one observation. We model the response with treatment i for the first treatment and treatment j for the second treatment by $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ for $i = 1, \cdots, a$, $j = 1, \cdots, b$. Where $\mu_{ij} = \mu + \alpha_i + \beta_j$, $\sum_{i=1}^{a}\alpha_i = 0$, and $\sum_{j=1}^{b}\beta_j = 0$.

(a) If the model is true $\bar{X}_{i\cdot}$ will have expected value $\mu + \alpha_i$. $\square$

(b) If the model is true $\bar{X}_{\cdot j}$ will have expected value β_j. $\square$

(c) $\bar{X}_{\cdot\cdot} = \mu$. $\square$

(d) Lets say our null, $\alpha_1 = \alpha_2 = \cdots = \alpha_a$ and $\beta_1 = \beta_2 = \cdots = \beta_b$, is false. Then $E[\bar{X}_{i\cdot}] = E[\bar{X}_{\cdot j}] = \mu$. $\square$

3. (4 pts) Exponential distribution with mean θ has density

$$\frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$$

on $0 < x < \infty$. θ is greater than 0. Its moment generating function is

$$\frac{1}{1-\theta t}$$

where $t < \frac{1}{\theta}$. On the other hand a gamma distribution has density

$$\frac{1}{\Gamma(\alpha)\theta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\theta}}$$

where $\alpha > 0$ and $\theta > 0$. We also have $0 < x < \infty$. It's moment generating function is

$$\frac{1}{(1-\theta t)^{\alpha}}$$

for $t < \frac{1}{\theta}$. Show that the sum of exponentials is a gamma.

4. (6 pts) Let $X_1, \cdots, X_n$ i.i.d. from exponential distribution with mean θ. That is, density is

$$\frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$$

on $0 \leq x < \infty$. θ is greater than 0. $H_0:\ \theta = 3$, $H_1:\ \theta = 7$. Find best critical region for $\alpha = 0.05$ when $n = 16$. Let U follow a Gamma distribution with $\theta = 3$ and $\alpha = 16$. Let it be given that $P(U > 69.291) = 0.05$.

5. (3 pts) Now find the best critical region when we change H_1 to $\theta = 11$.

6. (5 pts) Let X be a single sample from a Poisson distribution with density

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

for $x = 0, 1, 2, \cdots$ where $\lambda > 0$. We want to test $\lambda = 1$ against $\lambda > 1$. Let it be given that $P(X = 0) = P(X = 1) = 0.368$, $P(X = 2) = 0.184$, $P(X = 3) = 0.61$ Find the uniformly most powerful test at level $\alpha = 0.05$ (rejection probability has to be equal to or smaller than α).

7. (6 pts) We investigate a plot of land by dividing it into 81 equal area lots. We first record is vegetation. This can be 'tree', 'grass', or 'bare'. We also record primary terrain. This can be 'hill', 'valley', 'flat'. The tally is summarized in the below table.

	hill	valley	flat
tree	5	2	16
grass	8	10	11
bare	9	14	6

Is vegetation independent of terrain at $\alpha = 0.05$? You may use that $\chi^2_{4,0.95} = 9.488$

8. (5 pts) Let X_1, X_2, X_3, X_4 be independent $N(\mu_i, \sigma^2)$, $i = 1, \cdots, 4$. We test null of $\mu_1 = \mu_2 = \mu_3 = \mu_4$ against alternative of 'not null'. Below is the data we saw.

X_1	4	2	3
X_2	6	7	5
X_3	1	8	3
X_4	4	5	8

Would you reject the test at $\alpha = 0.01$? You may use that $F_{3,8,0.99} = 7.591$.

9. (6 pts) In the two way ANOVA, we have the following.

$$SS_{tot} = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{..})^2$$

$$SS_{err} = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

$$SS_a = \sum_{i=1}^{a} b(\bar{X}_{i.} - \bar{X}_{..})^2$$

$$SS_b = \sum_{j=1}^{b} a(\bar{X}_{.j} - \bar{X}_{..})^2$$

Show that $SS_{tot} = SS_{err} + SS_a + SS_b$.

10. (4 pts) We have two treatments and c observations per cell. The kth observation with first treatment i and second treatment j has distribution $X_{ijk} \sim N(\mu_{ij}, \sigma^2)$. We have $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, $\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, $\sum_{i=1}^{a} \gamma_{ij} = 0$ for all j, and $\sum_{j=1}^{b} \gamma_{ij} = 0$ for all i. The observed data are given below.

	b_1	b_2
a_1	9, 23, 11	13, 4, 10
a_2	1, 19, 7	17, 21, 18
a_3	20, 22, 24	8, 2, 14
a_4	5, 6, 3	16, 12, 15

Test $H_0: \ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ against H_1 : not null. You may use $F_{3,12,0.99} = 5.953$. Use $\alpha = 0.01$.

11. (4 pts) Test $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_a = 0$ against H_1 : not null. You may use $F_{1,12,0.99} = 9.330$. Use $\alpha = 0.01$.

12. (4 pts) Test $H_0: \ \gamma_{ij} = 0$ for all i and j, against H_1 : not null. Use $\alpha = 0.01$.

13. (3 pts) Test $H_0: \ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ and $\beta_1 = \beta_2 = \cdots = \beta_a = 0$, against H_1 : not null. That is, you are not putting any assumption on the γ_{ij} in the null. Notice that in the alternative, as always, there is no assumption on the γ_{ij}. Do you think this kind of test can be performed? Why?