Instructions:
1. There are six questions and 12 pages. Answer each question to the best of your ability.
2. Be as specific as possible and write as clearly as possible.
3. This is an in-class examination; do not discuss any part of this exam with anyone while you are taking the exam. **NO BOOKS, NO NOTES, NO INTERNET DEVICES, NO OUTSIDE ASSISTANCE.**
4. You may leave the examination room to use the restroom or to step out into the hallway for a short breather. **HOWEVER, YOU MUST LEAVE YOUR CELL PHONE AND ALL EXAM MATERIALS IN THE EXAMINATION ROOM.** If there is an emergency please discuss this with the exam proctor.
5. Vanderbilt’s academic honor code applies.

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Score</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>255</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Answer the following True or False questions on a separate sheet of paper and include any special explanations or reasoning that you think is important.

a. **True or False**: A Bayes factor is a ratio of averaged likelihoods.

b. **True or False**: The most powerful rejection region is always in the tails of the test statistic's sampling distribution under the null hypothesis.

c. **True or False**: When two studies yield the exact same \( p \)-value, both studies have generated equivalent amounts of statistical evidence about the parameter of interest.

For the remaining questions refer to the table below of observed results from \( N \) hypothesis. The table displays acceptance status versus the truth.

<table>
<thead>
<tr>
<th>( H_0 )</th>
<th>( H_0 ) Accepted</th>
<th>( H_0 ) Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) True</td>
<td>( a )</td>
<td>( b )</td>
<td>( a + b )</td>
</tr>
<tr>
<td>( H_0 ) False</td>
<td>( c )</td>
<td>( d )</td>
<td>( c + d )</td>
</tr>
<tr>
<td>Total</td>
<td>( a + c )</td>
<td>( b + d )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

d. What is the observed the Type I Error rate?

e. Use the table to define the probability of rejecting at least one true null hypothesis.

f. What is the observed the False Discovery Proportion?

g. **True or False**: The Type I Error rate and the False Discovery Proportion cannot be controlled simultaneously (assume the proportion of true null hypotheses and the number of hypothesis tests, \( N \), are fixed).
2. Consider the following R code and graphs:

```r
a = rnorm(10^6, mean=0, sd=1)
b = sort(a, decreasing = FALSE)  # ascending order
c = round( b[975000], digits=2 )
d = pnorm(median(b),mean=0,sd=1)
e = a^2
f = sum(e < 3.84 )
g = round( f/10^6, 2 )
h = mean(e)
i = rnorm(10^6, mean=0, sd=1)
j = a/i
k = sort(j, decreasing = FALSE)  # ascending order
m = mean(j)
```

Graph 1

Graph 2

Graph 3

Graph 4
Question 2 continued:

a. Make an educated guess for the value of $c$. Explain your guess or explain why no reasonable guess can be made.

b. Make an educated guess for the value of $d$. Explain your guess or explain why no reasonable guess can be made.

c. Make an educated guess for the value of $g$. Explain your guess or explain why no reasonable guess can be made.

d. Make an educated guess for the value of $h$. Explain your guess or explain why no reasonable guess can be made.

e. Make an educated guess for the value of $m$. Explain your guess or explain why no reasonable guess can be made.

f. Which graph best depicts the histogram of $j$? Explain your reasoning.

g. Assume that each graph represents a different empirical distribution. Explain how to generate Graph 4 from any of the other graphs. Be sure to define any notation that you introduce.
3. The following questions are based on the article “Cannabis Use Is Quantitatively Associated with Nucleus Accumbens and Amygdala Abnormalities in Young Adult Recreational Users.” (Gilman, et al. Journal of Neuroscience. 2014.)

The paper received a lot of attention from that nation news media and blogosphere. Much of the paper’s analysis, results, and conclusions were debated intensely. The Washington Post wrote:

Even casually smoking marijuana can change your brain, study says  
By Terrence McCoy, April 16 at 3:11 am

The days when people thought only heavy Cheech-and-Chong pot smokers suffered cognitive consequences may be over. A study in The Journal of Neuroscience says even casual marijuana smokers showed significant abnormalities in two vital brain regions important in motivation and emotion.

"Some of these people only used marijuana to get high once or twice a week," said co-author Hans Breiter, quoted in Northwestern University's Science Newsline. Breiter hailed the study as the first to analyze the effects of light marijuana use. "People think a little recreational use shouldn’t cause a problem, if someone is doing OK with work or school," he said. "Our data directly says this is not the case."

"This study raises a strong challenge to the idea that casual marijuana use isn’t associated with bad consequences," he added.

The study analyzed 20 pot smokers and 20 non-pot smokers between 18 and 25. Scientists asked them to estimate how much marijuana they smoked and how often they lit up over a three-month test period. ...

One frequently discussed figure is the scatterplot below. It shows the volume of the brain's left accumbens region versus the number of joints consumed per smoking occasion.

![Scatterplot showing the volume of the brain's left accumbens region versus the number of joints consumed per occasion.](image)
Question 3 continued:

a. Discuss the author’s claim that "our data directly says [a little recreational use alters the brain]" based only on your impressions from this figure. Be sure to point out important features of the figure that either lend support or cast doubt on the claim.

The data for this problem has been gleaned from the figure above (the original data has not been released). A quick analysis of these data can be found on the next page.

b. Which differences between group means, if any, appear to support the author's claim? Which, if any, do not? Be specific.

c. Explain whether a t-test based on the equal-variance assumption would be appropriate for testing group differences. Are there any features of this particular dataset that would make you wary of the two-sample equal-variance assumption in this case? Explain any rules of thumb that you used to make this judgment.

d. If the sample sizes in each group were equal, would that change your answer to part (c)? Explain your answer.

e. In mathematical notation, describe the null and alternative hypotheses for a *two-sided* two-sample equal-variance t-test.

f. As specifically as possible, describe the rejection region for a *two-sided* two-sample equal-variance t-test.

g. What is the formula for the test statistic for a two-sample equal-variance t-test? Be specific.

h. Write a conclusion for a *two-sided* two-sample equal-variance t-test comparing the groups zero occasions and at least one occasion.

i. Also included in the analysis was a non-parametric test of group differences (zero versus at least one). Describe this test and interpret its results.

j. Explain how this non-parametric test is different from the t-test. Discuss their relative advantages and disadvantages. Some critiques of the paper have argued that the maximum volume in the ‘two joints per occasion’ group is an outlier. Explain how this is addressed by the analysis.
Output for Question #3

```
. bysort jointsper: summarize volume

-------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
</table>
| jointsper = 0
| volume   | 20  | 560.55 | 79.74728  | 390       | 700   |

-------------------------------------------------------------------------------
| jointsper = 1
| volume   | 10  | 572   | 98.15521  | 380       | 690   |

-------------------------------------------------------------------------------
| jointsper = 2
| volume   | 6   | 685.8333 | 125.1566 | 570       | 900   |

-------------------------------------------------------------------------------
| jointsper = 3
| volume   | 4   | 635   | 78.20912  | 535       | 725   |

. summarize volume

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>volume</td>
<td>40</td>
<td>589.65</td>
<td>99.74314</td>
<td>380</td>
<td>900</td>
</tr>
</tbody>
</table>

. summarize volume if j!=0

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>volume</td>
<td>20</td>
<td>618.75</td>
<td>110.8089</td>
<td>380</td>
<td>900</td>
</tr>
</tbody>
</table>
```
Output for Question #3

`. ttest volume, by(joints)`

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>560.55</td>
<td>17.83203</td>
<td>79.74728</td>
<td>523.2271 597.8729</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>618.75</td>
<td>24.77763</td>
<td>110.8089</td>
<td>566.8898 670.6102</td>
</tr>
<tr>
<td>combined</td>
<td>40</td>
<td>589.65</td>
<td>15.77078</td>
<td>99.74314</td>
<td>557.7506 621.5494</td>
</tr>
</tbody>
</table>

`diff | mean(0) - mean(1) t = -1.9065 Ho: diff = 0 degrees of freedom = 38`

`. ttest volume, by(joints) uneq`

Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>560.55</td>
<td>17.83203</td>
<td>79.74728</td>
<td>523.2271 597.8729</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>618.75</td>
<td>24.77763</td>
<td>110.8089</td>
<td>566.8898 670.6102</td>
</tr>
<tr>
<td>combined</td>
<td>40</td>
<td>589.65</td>
<td>15.77078</td>
<td>99.74314</td>
<td>557.7506 621.5494</td>
</tr>
</tbody>
</table>

`diff | mean(0) - mean(1) t = -1.9065 Ha: diff < 0 Satterthwaite's degrees of freedom = 34.5187`

`. ranksum volume, by(joints)`

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

<table>
<thead>
<tr>
<th>joints</th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>348</td>
<td>410</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>472</td>
<td>410</td>
</tr>
<tr>
<td>combined</td>
<td>40</td>
<td>820</td>
<td>820</td>
</tr>
</tbody>
</table>

unadjusted variance 1366.67
adjustment for ties -0.77
adjusted variance 1365.90

Ho: volume(joints==0) = volume(joints==1)
   z = -1.678
   Prob > |z| = 0.0934
The paper looked at several outcomes measures, including grey matter density, and several exposures, including smoking occasions per day. Suppose we categorized the outcome grey matter density as "thick" or "thin". Here is a contingency table displaying hypothetical results of grey matter density by smoking occasions per day, along with two analyses of the table (next page).

<table>
<thead>
<tr>
<th>Grey Matter Density</th>
<th>Smoking Occasions Per Day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Thin</td>
<td>10</td>
</tr>
<tr>
<td>Thick</td>
<td>10</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
</tr>
</tbody>
</table>

a. Write down, in mathematical notation, the null hypothesis(s) that are being tested.

b. If you could only report one of the two tests used, which one would you choose and why?

c. Does the analysis you chose in part (b) support the claim that the "data directly says [a little recreational use alters the brain]"? Explain.

d. Collapse the table into cannabis users vs. non-users to examine the association between cannabis use (y/n) on grey matter density (thick/thin). What is the collapsed table?

e. Write down, in mathematical notation, the null hypothesis(s) that are being tested and compare this with your answer in part (a). Note the key differences, if any.

f. For your collapsed table from part (d), name and write out three different point estimates that describe the observed association. You don’t need to solve for the final numeric answers.

g. Explain how you might calculate a 95% confidence interval for one of the measures you presented in the preceding question. Be sure to define any new notation that you introduce.
Output for Question #4

\[ m = \text{matrix}( \begin{c} 10, 10, 8, 7, 0, 3, 0 \end{c}, \text{nrow}=2 ) \]
> chisq.test(m)
   Pearson's Chi-squared test
   data:  m
   X-squared = 8.5714, df = 3, p-value = 0.03557
   Warning message:
   In chisq.test(m) : Chi-squared approximation may be incorrect

> fisher.test(m)
   Fisher's Exact Test for Count Data
   data:  m
   p-value = 0.04323
   alternative hypothesis: two.sided

.tabi 10 8 7 3 \ 10 2 0 0,chi col row e

+------------------+
<table>
<thead>
<tr>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
</tr>
<tr>
<td>row percentage</td>
</tr>
<tr>
<td>column percentage</td>
</tr>
</tbody>
</table>
+-------------------+

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>row</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>35.71</td>
<td>28.57</td>
<td>25.00</td>
<td>10.71</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>80.00</td>
<td>100.00</td>
<td>100.00</td>
<td>70.00</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>83.33</td>
<td>16.67</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>20.00</td>
<td>0.00</td>
<td>0.00</td>
<td>30.00</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>25.00</td>
<td>17.50</td>
<td>7.50</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Pearson chi2(3) = 8.5714  Pr = 0.036
Fisher's exact = 0.043
5. The conditional expectation of continuous measure $Y$ (cm) given another continuous measure $X$ (cm) is modeled as a linear relation. The 25 observed values of each variable are standardized (to mean 0 and stdv. 1) and denoted as $ZX$ and $ZY$ respectively. The regression results are given below.

a. Provide a 95% confidence interval for the coefficient of $X$ and interpret both the interval and the coefficient.

b. Provide a 95% confidence interval for the coefficient of $ZX$ and interpret both the interval and the coefficient.

c. Estimated the correlation between $Y$ and $X$. Provide a 95% CI and interpret.

d. What is the conditional mean of $Y$ given $X = 1$? Interpret this number.

e. Is there evidence of a linear relationship between $X$ and $Y$? Can you conclude the relationship is linear? Explain.

Output for Problem 5

```
. regress y x

Source |        SS       df       MS              Number of obs =      25
-------------+-------------------------------------------+-----------------------------+--------------------
Model |  6.4777658     1  6.4777658           Prob > F      =  0.0134
     Residual | 20.7835533    23  .903632754           R-squared     =  0.2376
-------------+-------------------------------------------+-----------------------------+--------------------
     Total |  27.2613191    24 1.1358883           Root MSE      =   .9506
-------------+-------------------------------------------+-----------------------------+--------------------

. regress zy zx

Source |        SS       df       MS              Number of obs =      25
-------------+-------------------------------------------+-----------------------------+--------------------
Model |  5.7028193     1  5.7028193           Prob > F      =  0.0134
     Residual | 18.297181    23  .795529608           R-squared     =  0.2376
-------------+-------------------------------------------+-----------------------------+--------------------
     Total |  24.0000003    24 1.00000001          Root MSE      =  .89192
-------------+-------------------------------------------+-----------------------------+--------------------
```

```
y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------+-----------------------------+--------------------
x | -.4693033   .1752819    -2.68   0.013   -.8319015   -.1067052
    _cons |  .1282957   .1901395     0.67   0.507   -.2650378    .5216293

zy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------+-----------------------------+--------------------
   zx | -.4874602   .1820634    -2.68   0.013   -.864087   -.1108335
    _cons | -1.98e-09   .1783849    -0.00   1.000   -.3690174    .3690173
```
6. Researchers are interested in modeling county-level adjusted mortality rates (AMR: deaths per 100,000, adjusted for age distribution) by county characteristics. The characteristics of interest are educational attainment (EDU: percent population with at least some college education), household income (INC: percent population whose income falls at or above twice the federal poverty level), and urbanicity (URB: percent population living in an urbanized area). The proposed model, below, is fit on the next page for a sample of U.S. counties based on 2010 data.

\[ AMR = \beta_0 + \beta_1 \text{EDU} + \beta_2 \text{INC} + \beta_3 \text{URB} + \epsilon \]

a. State three key assumptions for this linear regression model.

b. Using the table of correlations below (first table), discuss the relationship between AMR and each of EDU, INC, and URB. What impact does this have in interpreting the model or finding the 'best-fit' model.

c. Interpret the partial and semi-partial correlations for URB shown in the second table. Note that the partial correlation is positive, but the unadjusted correlation is negative. What is a possible explanation for this change in sign?

d. Interpret \( \beta_2 \).

e. Interpret R-squared.

f. What is the Root MSE and what does it estimate? Explain how this parameter is used in estimation and prediction.

g. Suppose that the variance of AMR increases with income. Explain how to diagnose this problem and how to resolve it.
Output for Problem 6

```
. corr amr edu inc urb
(obs=491)
      amr    edu    inc    urb
-------------
   amr |   1.000
   edu |  -0.559   1.000
   inc |  -0.710   0.637   1.000
   urb |  -0.278   0.551   0.416   1.000
```

```
. pcorr amr edu inc urb
(obs=491)

Partial and semipartial correlations of amr with

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>edu</td>
<td>-0.2277</td>
<td>-0.1603</td>
<td>0.0519</td>
<td>0.0257</td>
<td>0.0000</td>
</tr>
<tr>
<td>inc</td>
<td>-0.5616</td>
<td>-0.4653</td>
<td>0.3154</td>
<td>0.2165</td>
<td>0.0000</td>
</tr>
<tr>
<td>urb</td>
<td>0.1202</td>
<td>0.0830</td>
<td>0.0144</td>
<td>0.0069</td>
<td>0.0078</td>
</tr>
</tbody>
</table>
```

```
. regress amr edu inc urb

Source |       SS       df       MS
--------|--------------|------------|-------------|
Model   |  2494373.14   3  831457.712
Residual |  2211018.41   487  4540.07887
Total   |  4705391.55   490  9602.83989

Number of obs =     491
F(  3,   487) =  183.14
Prob > F      =  0.0000
R-squared     =  0.5301
Adj R-squared =  0.5272
Root MSE      =   67.38

| amr | Coef.     | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-----|-----------|-----------|-------|-------|----------------------|
| edu | -2.1451   | .415652   | -5.16 | 0.000 | -2.961825            | -1.328436 |
| inc | -5.98589  | .3995832  | -14.98| 0.000 | -6.77101             | -5.200771 |
| urb | .3143184  | .1176768  | 2.67  | 0.008 | .0831014             | .5455353 |
| _cons | 854.6526 | 20.46401  | 41.76 | 0.000 | 814.444              | 894.8613 |
```