# Homework 1 2x2s

Even before you do the 2x2 problems below, try to install R, RStudio, 11 add-on packages, Rprofile, and cufunctions. Instructions are in
ctsa-stats-setup.pdf
As it says on slide 17, if you don't see the message shown there when you launch RStudio and type ?curead, email me the Errors document. No need to email me if the installation was successful.

Here are quite a few clinical research reports that require 2x2 tables to be constructed and analyzed. Do as many as you wish (or none!). Check your solution against mine, which is at
http://biomath.net/resources/ctsa-stats-hwsoln.xlsx

Email me just questions you have, along with the R Errors document if any problems there. We will start the class next week with a discussion of any questions people have.

1. NY Times 12/20/2000 Study Links Use of Caffeine to Higher Risk of Miscarriage. The study (NEJM 343:1839-45) involved 562 women who had miscarriages at 6 to 12 weeks of pregnancy, and 953 matched controls. The research found that 116 cases and 307 controls consumed less than 100 mg daily of caffeine.

2. NY Times 8/6/2002 Weighing Odds and Babies. Giving birth at home is riskier both for the child and the mother, even in uncomplicated pregnancies, according to a new study. The study (Obstetrics & Gynecology 100:253-9) looked at records of the 6,133 births that were completed or begun in a home in Washington state from 1989 to 1996. These were compared with 10,593 hospital deliveries, matched for other risk factors. When the delivery was planned and begun in the home, there were 20 neonatal deaths vs 18 in hospital). Low Apgar scores evaluating the newborn's health were more common in children born at home (24 vs. 21).

3. NY Times 7/3/2012 Drinking Coffee and Colon Cancer Risk. The NIH-AARP Diet and Health Study followed half a million older Americans over 15 years (AJCN 96:374-81), and found that those who drank four or more cups of coffee a day - regular or decaf - had a lower risk of colon cancer compared with coffee abstainers (1011/77,000 vs 647/50,000).

4. NY Times 6/11/2012 Diabetes Drugs Carry Vision Risks. A popular class of drugs used to treat Type 2 diabetes may increase the risk of vision problems, a new study (Archives of Internal Medicine 172:1005-11) suggests. About 1.3% of people taking a thiazolidinedione (TZD) developed diabetic macular edema (41/3,227), compared to a rate of 0.2% among those who were not on one of the medications (227/100,141).

5. NY Times 2/13/2013 Folic Acid Lowers Autism Risk. Women taking folic acid, a B vitamin, before pregnancy were less likely to give birth to children with the most severe form of autism, a new study has found. Researchers in Norway and at Columbia University followed more than 80,000 Norwegian mothers who gave birth between 2002 and 2008. Among the children, 270 with autism spectrum disorders, including 114 with autism disorder: 50/24,134 among women not taking folic acid, versus 64/61,042 among those who did.

6. NY Times 2/28/2017 Smokers: Eat Your Fruits and Vegetables. The study, in Thorax, looked at Swedish men who completed health and dietary questionnaires and were followed from 1998 to 2012. Nearly two-thirds had smoked at some point, and roughly one in four were current smokers. Ever-smokers who ate five or more servings of fruits and vegetables a day were much less likely to develop lung disease COPD (207 out of 5223) than those who ate two servings or less (577 out of 6115). There was no benefit for nonsmokers (32/3596 vs 60/5537). Analyze ever-smokers and nonsmokers separately.

7. Does second-hand smoke increase risk of chronic kidney disease (CKD)? (CJASN 0.2215/CJN.09540818) 1,948 (381 unexposed to 2nd-hand smoke) followed at 2 Korean sites; 319 (35 unexposed) developed CKD over 2y.

8. Does using prescription opioids increase risk of invasive pneumococcal disease (IPD)? (Annals of Internal Medicine, doi:10.7326/M17-1907): 1,233 with IPD matched to 24,399 controls by diagnosis date, age, residence; 311 and 3521, respectively, on opioids.

9. Is fast eating associated with obesity? (BMJ Open 2018;8:e019589) 59,717 Japanese Type 2 diabetics: categorized as fast (22070, 9884 obese), normal (33455, 9886 obese), or slow eaters.

10. Does Alzheimer's Disease (AD) increase risk of 30-day mortality after cardiovascular event (stroke, MI)? (J Alzheimer's Dis. 2015;48:241-9) all 73,005 Finns diagnosed with AD in 2005-2012 (7641 events, 2093 deaths) and 215,449 age-, gender- and region-matched non-AD (20977 events, 4450 deaths).

## The following problems are harder:

11. Was Columbia's stricter implementation of guidelines (mask, goggles, hand-sanitizing) among 29,000 healthcare personnel (HCP) after 2013 effective? (Infect Control Hosp Epidemiol 2017;38:1361–136) Number infected with the flu virus was 229 in 2012-13 and 89 in 2014-15. Assume flu case rate to be unchanged (actually 11.6% and 12.7% nationally).

[Additional qn: In 2012-13, the rate of vaccination among HCP was 86%, and the number of infected who had been vaccinated was 129. Calculate the efficacy of the vaccine.]

12. Does physiotherapy help recover from ankle sprain? (BMJ 355:i5650) Data from a randomized trial in "1-ankle" tab of http://biomath.net/resources/ctsa-stats-hw.xlsx. Do intent-to-treat (ITT) analysis, that is, include all subjects in analysis. Same as above, except do per-protocol (P-P) analysis, that is, include only those who stuck to the protocol (yes for PerProtocol). You may find the countifs function useful here.

13. Survey of Republicans on Global Warming http://www.democracycorps.com/attachments/article/1025/Dcor_RPP_Memo_2.29.2016_FOR%20RELEASE.pdf 800 Republicans were surveyed – 17% Tea Party, 30% Evangelical, 14% Observant Catholic, 31% Moderate (the groups are mutually exclusive, meaning no one belonged to more than one category; 8% were Establishment and excluded from analysis). The questions were prefaced with: "Please tell me if you think that it might be true OR it might be the liberal media trying to justify their agenda." One question: "Scientists say that 2015 was the hottest year in historical record by a wide margin, and 98% of climate scientists agree that human activity is a significant factor in climate change." Those choosing "True" were 23% of Tea Party, 26% of Evangelical, 52% of Observant Catholic, 65% of Moderate. Compare proportion choosing "True" between every 2 categories (you will need to construct a table of the numbers in different categories that chose "True" vs Not.)

What is the RR or OR of Moderates choosing "True" compared to (1) the rest of the cohort, and (2) just Tea Party. Comment on differences (or not) in calculated RR/OR between the two cases.

# 2 t-tests in Excel

The **2-cereal** tab has the data from a paper on cheerios (*J Am Diet Assoc.* 105:967-70, 2005). [The data are slightly different from what is in the paper; I made the two groups equal in size to make your work easier.] Calculate means, SDs, and p-values by paired and unpaired t-tests as you did in class with the NEJM data. [Hint: You should sort the data first so the two groups are separated and not interleaved. Click on a cell in the group column with B and Y, then click on Data and A-Z under Sort.] Remember that, while the cheerios data are similar to the NEJM data, there are differences: the number of subjects is different, as are the number of groups. Also, the NEJM spreadsheet has some calculated variables that the cheerios file doesn't have; you have to do this first.

# 3 Power Analysis

Table 3. Estimated Effect of Overweight in Adolescence on Subsequent Social and Economic Characteristics and Self-Esteem among Men.*

| VARIABLE | OBSERVED VALUE | | CRUDE ESTIMATE OF DIFFERENCE (95% CI) | | ADJUSTED ESTIMATE OF DIFFERENCE (95% CI) | | P VALUE |
|---|---|---|---|---|---|---|---|
| | OVERWEIGHT (N = 175) | NONOVERWEIGHT (N = 4726) | | | | | |
| Married (%) (n = 4642) | 40 | 48 | −8 | (−1 to −15) | −11 | (−3 to −18) | 0.005 |
| Household income ($) (n = 4022) | 26,008 | 31,462 | −5,454 | (−2,488 to −8,420) | −2,876 | (−10 to −5,742) | 0.05 |
| Income below poverty level (%)† (n = 4022) | 17 | 10 | 7 | (1 to 13) | 5 | (−1 to 9) | 0.07 |
| Education (yr) (n = 4610) | 12.4 | 13.1 | −0.8 (−0.5 to −1.1) | | −0.2 (−0.5 to 0.0) | | 0.08 |
| Completed college (%) (n = 4022) | 10 | 23 | −13 | (−8 to −18) | −5 | (0 to 10) | 0.05 |
| Self-esteem in 1987 (n = 4901) | 32.8 | 33.9 | −1.1 (−0.5 to −1.7) | | −0.2 (−0.8 to 0.30) | | 0.17 |

The table above is from a 1993 NEJM paper on marriage and income prospects of overweight teens. Suppose you want to design a study to look at the same question in a contemporary teenage male population. The plan is to recruit equal numbers of overweight and lean boys.

1. If marital status is the outcome of interest, how many boys would you need in each group to find the difference between overweight and lean to be significant at p=0.05 (with 80% power), assuming your outcomes will be the same as in the NEJM paper?

2. How many if household income were the outcome of interest? The paper does not give the SD for income. Assume it is $10,000.

3. How many if the proportion living below poverty level were the outcome of interest?

4. How many if you were interested in all three, still testing at p=0.05?

5. How many if the smallest marriage difference of practical interest (I certainly don't think any difference is of interest) is 5%, assuming the overweight boys' marital status will be the same as in the NEJM paper?

6. What if the smallest income difference of interest is $2,000?

7. Answer the six questions above if you plan to recruit 10 times as many lean boys as overweight boys.

8. If you only have resources to recruit 200 overweight and 200 lean boys, how small a difference in marital status and income do you have 80% power to find significant at p=0.05 (again assuming the overweight boys' marital status and income will be the same as in the NEJM paper)?

9. What if you recruit 100 overweight and 300 lean boys?

10. What if you recruit 100 overweight and 1000 lean boys?

11. Anything interesting between the answers to 8 and 9, noting that both designs involve 400 boys total?

12. Anything interesting between the answers to 8 and 10?

# 4 M8209 Biostatistics Midterm Examination 2018 (*Plus More)*

**\*\* Starred problems may be harder; you can get 100%
without them.
[Questions 6, 10 and 12 are quite hard.
Scatter plots for #3 and #7 go in 4-GLC tab.]**

The **4-GLC tab** has gun death rates/100,000 (**GDR**)
and state gun law ratings (**Score**) for 2010 and
2016, very kindly provided, solely for use in this
exam, by folks at lawcenter.giffords.org. "**10**" is
2010, "**16**" is 2016; **D_GDR** is change in GDR
from '10 to '16 (I've collapsed some grades, scaled
'10 scores).

**(14%) 1:** Fill in the column for D_GDR. Fill in
Table 1 (4-R Results Tab: mean and SD or median
and IQR as appropriate; answer Rank question).
In the following questions, assume data are normal
so *t*-tests, correlations and regressions are valid.
**(5%) 2:** Do pairwise correlations among Score10,
Rank10, GDR10, Score16, Rank16, GDR16, and
D_GDR. Fill in Table 2.
**(5%) 3:** Do scatter plots of **GDR16** vs Score16 and
GDR16 vs Rank16; include equation and $R^2$.
**\*\*(5%):** Improve plots visually (labels, axes, etc).
**(5%) 4:** Compare GDR16 among the 2016 gun law
grade categories. Fill in Table 3.
**(5%) 5:** If a state were to increase its score by 20
points, what effect would you expect in that state's
2016 gun death rate? What if it changed its grade
from F to C? From C to A-?
**\*\*(10%) 6:** Use the residual errors (or $R^2$) to decide
which of 2016 gun law score, gun law rank and gun
law grade is best correlated with gun death rate.
**(5%) 7:** Do scatter plots of **D_GDR** vs Score16 and
D_GDR vs Rank16; include equation and $R^2$.
**\*\*(5%):** Improve plots visually (labels, axes, etc).
**(5%) 8:** Compare D_GDR among the 2016 gun law
grade categories. Fill in Table 4.
**(5%) 9:** If a state were to increase its score by 20
points, what effect would you expect in the 2010 to
2016 change in that state's gun death rate? What if
it changed its grade from F to C? From C to A-?
**\*\*(10%) 10:** Use the residual errors (or $R^2$) to
decide which of 2016 gun law score, gun law rank
and gun law grade is best correlated with D_GDR.
**(20%) 11:** Regress GDR16 on GDR10 & Grade16.
Fill in Table 5; paste R graph of GDR16 vs GDR10.

**\*\*(10%) 12:** Do the regression p-values for the
grades seem inconsistent with the p-values in Q8
(Table 4)? If so, try to reconcile the two results.
**\*\*(25%) 13:** Regress, with interaction, GDR16 on
GDR10 and Grade16. Fill in Table 6 at GDR10=5
and 15; paste R graph of GDR16 vs GDR10.
**\*\*(15%) 14:** What overall conclusion(s) do you
draw? Which tests, significant or not, are relevant?
**\*\*(10%) 15:** What, if any, other questions or other
data can help in the analysis?

For each of 4 studies below, enter type of study, risk
factor, outcome variable, and RR or OR. Calculate
whatever are appropriate among sensitivity,
specificity, positive predictive value, negative
predictive value, false positive rate, and false
negative rate. **Enter NA if inappropriate.**
**Hint:** Use formulas in Excel to calculate these
quantities, so you can copy cells from one problem
to others and avoid recalculations.

**(7%) 16:** Qn: Does using prescription opioids
increase risk of invasive pneumococcal disease
(IPD)? (Annals of Internal Medicine,
doi:10.7326/M17-1907): 1,233 with IPD matched to
24,399 controls by diagnosis date, age, residence;
311 and 3521, respectively, on opioids.
**(7%) 17:** Qn: Is fast eating associated with obesity?
(BMJ Open 2018;8:e019589) 59,717 Japanese Type
2 diabetics: categorized as fast (22070, 9884 obese),
normal (33455, 9886 obese), or slow eaters.
**(7%) 18:** Qn: Does AD increase risk of 30-day
mortality after cardiovascular event (stroke, MI)? (J
Alzheimer's Dis. 2015;48:241-9) all 73,005 Finns
diagnosed with AD in 2005-2012 (7641 events,
2093 deaths) and 215,449 age-, gender- and region-
matched non-AD (20977 events, 4450 deaths).
**(10%) 19:** Qn: Was Columbia's stricter
implementation of guidelines (mask, goggles, hand-
sanitizing) among 29,000 healthcare personnel
(HCP) after 2013 effective? (Infect Control Hosp
Epidemiol 2017;38:1361–136) Number infected
with the flu virus was 229 in 2012-13 and 89 in
2014-15. Assume flu case rate to be unchanged
(actually 11.6% and 12.7% nationally).

**(10%) 20:** In the last study above, in 2012-13, the rate of vaccination among HCP was 86%, and the number of infected who had been vaccinated was 129. Calculate the efficacy of the vaccine.

21. Work in the 4-NEJM tab, which has three columns for LDL cholesterol, to be calculated from the Friedewald equation (this is how laboratories get LDL cholesterol values for TG<400 mg/dL):
ldlpre = tcpre - hdlpre - 0.2*tgpre
ldlstudy = tcstudy - hdlstudy - 0.2*tgstudy
ldlchange = ldlstudy-ldlpre
What is gained by my positioning of the LDL cholesterol columns where they are?

22. Use cu1way to do the anova-post-hoc t and the Dunn tests, to compare the three diets for ldlchange. Comment on how the p-values compare.

23. This was in the 2017 midterms: The "4-crimedata" tab has **real** data on crime rates (#/100K/day) in 2492 counties categorized into six levels of urbanness and as sanctuary or non-sanctuary, kindly provided by Prof. Tom K Wong of UC San Diego (thousands of rows, so copy data by selecting column headings). I learned about his work from this Center for American Progress report:
https://www.americanprogress.org/issues/immigration/reports/2017/01/26/297366/the-effects-of-sanctuary-policies-on-crime-and-the-economy/
[The report defines sanctuary counties as counties that do not assist federal immigration enforcement officials by holding people in custody beyond their release date.]
Use cu2way to assess the sanctuary effect (i.e., sanctuary vs nonsanctuary crime rates) within each of the six urbanness levels (L-central, L-fringe, etc).