Take-home exam section 'Introduction to R and foundational statistics'

General information:

-the hand-in deadline for this exam is at the start of next class.
-you may work on the exam in duos, let me know who you work with.
-you may talk to other class participants but keep this limited.
-no interactions with other people not in the class, asks questions on web forums, etc.
-write answers concisely. You may lose points if you include many potential answers in the hope to get something right.
-show how you do the calculations.
-questions which require using R have a '***R***' after the final punctuation.
-provide the R script for these answers in one R script, clearly indicating the question number at the start of each answer (e.g. # --- Questions 2c ----).

Good luck.

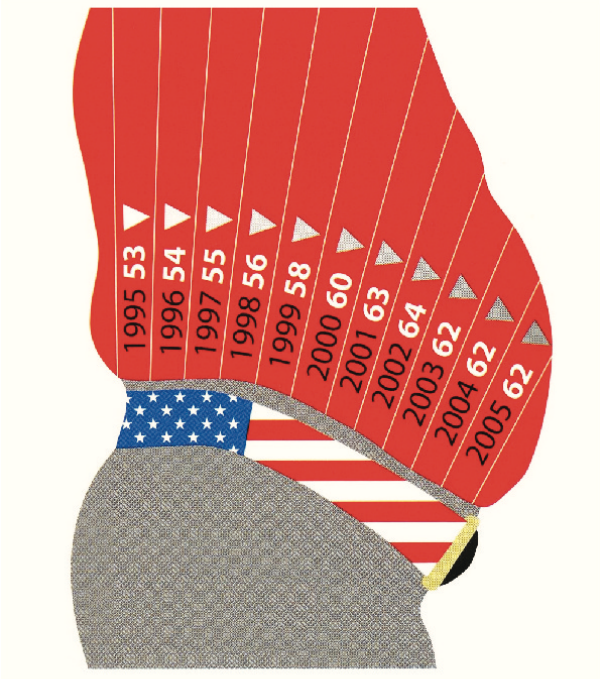
Questions

1. In each of the following examples, i) indicate which variable is the explanatory variable and which is the response variable and ii) indicate whether the study is an experimental or observational study. Provide a one sentence rational for your answer for i and ii.
a. The anticoagulant warfarin is often used as a pesticide against house mice, *Mus musculus*. Some populations of the house mouse have acquired a mutation in the *vkorc1* gene from hybridizing with the Algerian mouse, *M. spretus* (Song et al. 2011). In the Algerian mice, this gene confers resistance to warfarin. In a hypothetical follow-up study, researchers collected a sample of house mice to determine whether presence of the *vkorc1* mutation is associated with warfarin resistance in house mice as well. They fed warfarin to all the mice in a sample and compared survival between the individuals possessing the mutation and those not possessing the mutation.
b. Cooley et al. (2009) randomly assigned either of two treatments, naturopathic care (diet counseling, breathing techniques, vitamins and a herbal medicine) or standardized psychotherapy (psychotherapy with breathing techniques and a placebo added), to 81 individuals having moderate to severe anxiety. Anxiety scores decreased an average of 57% in the naturopathic group and 31% in the psychotherapy group.
c. Individuals highly sensitive to rewards tend to experience more food cravings and are more likely to be overweight or develop eating disorders than other people. Beaver et al. (2006) recruited 14 healthy volunteers and scored their reward sensitivity using a questionnaire (they were asked to answer "yes" or "no" to questions like: "I'm always willing to try something new if I think it will be fun"). The subjects were then presented with images of appetizing foods (e.g., chocolate cake, pizza) while activity of their fronto–striatal–amygdala–midbrain was measured

using functional MRI. Reward sensitivity of subjects was found to correlate with brain activity in response to the images.

d. Endostatin is a naturally occurring protein in humans and mice that inhibits the growth of blood vessels. O'Reilly et al. (1997) investigated its effects on growth of cancer tumors, whose growth and spread requires blood vessel proliferation. Mice having lung carcinoma tumors were randomly divided into groups that were treated with doses of either 0, 2.5, 10, and 20 mg/kg of endostatin injected once daily. They found that higher doses of endostatin led to inhibition of tumor growth.

2. Examine the following figure, which displays the percentage of adults over 18 with a "body mass index" greater than 25 in different years (modified from *The Economist* 2006, with permission). Body mass index is a measure of weight relative to height.



a. What is the main result displayed in this figure?

b. Which of the four principles for drawing good graphs is violated here? How are they violated?

c. Redraw the figure in R using the most appropriate method discussed. What type of graph did you use? ***R***

3. Measurements of lifetime reproductive success (LRS) of individual wild animals reveal the disparate contributions they make to the next generation. Jensen et al. (2004) estimated LRS of male and female house sparrows in an island population in Norway. They measured LRS of an

individual as the total number of "recruits" produced in its lifetime, where a recruit is an offspring that survives to breed one year after birth. Parentage of recruits was determined from blood samples using DNA techniques. Their results are tabulated as follows:

	Frequency	
Lifetime reproductive success	**Females**	**Males**
0	30	38
1	25	17
2	3	7
3	6	6
4	8	4
5	4	10
6	0	2
7	4	0
Total	80	84

a. Which sex has the higher mean lifetime reproductive success? Calculate the in R without using the mean() function. ***R***
b. Every recruit must have both a father and a mother, so it is not easy to see why male and female LRS should differ. Can you think of a biological explanation?
c. Which sex has the higher variance in reproductive success? Calculate the in R without using the var() function. ***R***
d. Calculate the median in R without using the median() function. ***R***

4. A massive survey of sexual attitudes and behavior in Britain between 1999 and 2001 contacted 16,998 households and interviewed 11,161 respondents aged 16–44 years (one per responding household). The frequency distributions of ages of men and women respondents were the same. The following results were reported on the number of heterosexual partners individuals had had over the previous five-year period (Johnson et al. 2001).

	Sample size, *n*	Mean	Standard deviation
Men	4620	3.8	6.7
Women	6228	2.4	4.6

a. What is the standard error of the mean in men? What is it in women? Assume that the sampling was random.
b. Which is a better descriptor of the variation among men in the number of sexual partners, the standard deviation or the standard error? Why?
c. Which is a better descriptor of uncertainty in the estimated mean number of partners in women, the standard deviation or the standard error? Why?
d. A mysterious result of the study is the discrepancy between the mean number of partners of heterosexual men and women. If each sex obtains its partners from the other sex, then the true

mean number of heterosexual partners should be identical. Considering aspects of the study design, suggest an explanation for the discrepancy.

5. Can parents distinguish their own children by smell alone? To investigate, Porter and Moore (1981) gave new T-shirts to children of 9 mothers. Each child wore his or her shirt to bed for three consecutive nights. During the day, from waking until bedtime, the shirts were kept in individually sealed plastic bags. No scented soaps or perfumes were used during the study. Each mother was then given the shirt of her child and that of another, randomly-chosen child and asked to identify her own by smell. Eight of 9 mothers identified their children correctly. Use this study to answer the following questions, using if necessary a two-sided test and a significance level of $\alpha = 0.05$.

a. To carry out a statistical test based on these data, what is the appropriate null hypothesis?

b. What is the alternative hypothesis?

c. What test statistic should you use?

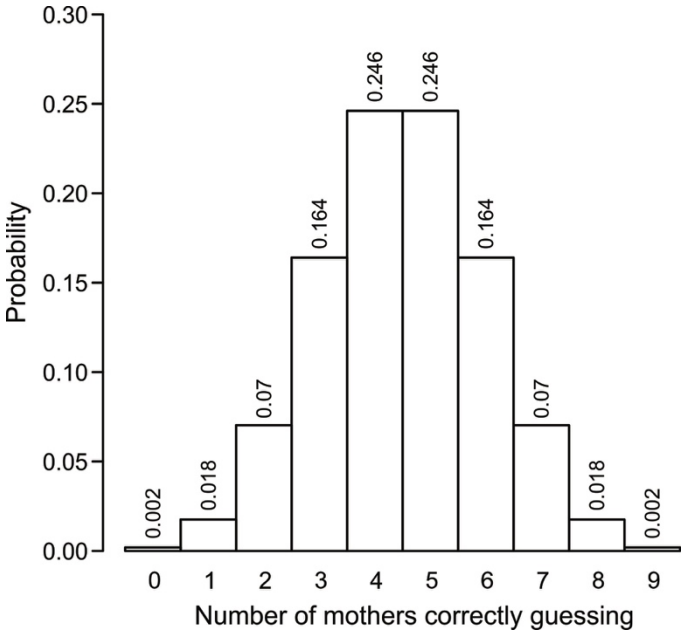
d. The accompanying figure shows the null distribution for the number of mothers out of 9 guessing correctly. The probability of each outcome is given above the bars. If the null hypothesis were true, what is the probability of exactly 8 correct identifications?

e. If the null hypothesis were true, what is the probability of obtaining 8 or more correct identifications?

f. What is the *P*-value for the test?

g. What is the appropriate conclusion?

h. As part of the analysis of these data, why would it be a good idea to calculate a 95% confidence interval for the true proportion of correct identifications?



6. In a test of Murphy's Law, pieces of toast were buttered on one side and then dropped. Murphy's Law predicts that they will land butter side down. Out of 9821 total slices of toast dropped, 6101 landed butter side down. (Believe it or not, these are real data!)
a. What is a 95% confidence interval for the probability of a piece of toast landing butter side down?
b. Using the results of part (a), is it plausible that there is a 50:50 chance of the toast landing butter side down or butter side up?

7. Aging workers of the Neotropical termite, *Neocapritermes taracua*, develop blue, crystal-containing glands ("backpacks") on their backs When they fight intruding termites and are hampered, these "blue" termites explode, the glands erupting a sticky liquid (Šobotník et al. 2012). The data below are from an experiment that measured the toxicity of the blue substance. A single drop of the liquid extracted from blue termites was placed on individuals of a second termite species, *Labiotermes labralis*, and the number that were immobilized (dead or paralyzed) within 60 minutes was recorded. The frequency of this outcome was compared with a control treatment in which liquid from glands of "white" termites lacking the blue crystals was dropped instead. Is the blue liquid toxic compared to liquid from white termites?
a. What is the H_0 and H_A?
b. Calculate the expected frequencies using R. ***R***
c. What test can you do to test the hypothesis, what are the assumptions and are they met?
d. Test if your H_0 can be rejected. ***R***

	Number of *L. labralis* individuals	
	Unharmed	Immobilized
Blue workers	3	37
White workers	31	9

8. The proportion of traffic fatalities for each U.S. state resulting from drivers with high alcohol blood levels in 1982 was approximately normally distributed, with mean 0.569 and standard deviation 0.068 (U.S. Department of Transportation Traffic Safety Facts 1999).
a. What proportion of states would you expect to have more than 65% of their traffic fatalities from drunk driving?
b. What proportion of deaths due to drunk driving would you expect to be at the 25th percentile of this distribution? In absence of physical tables, you can use R. ***R***

9. The normal distribution is very common in the natural world. It may not come as a surprise that many articles use statistics that are based on the normal distribution. The following formula describes the normal distribution.

$$f(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$

This gives the probability density *f(Y)* for a value *Y*. The value *Y* can be any real number ranging between negative and positive infinity; μ is the mean and σ the standard deviation.

~~a. What is the *standard* normal distribution?~~
~~b. Write your own function in R to draw a normal distribution with a mean and standard deviation provided by the user. ***R***~~
~~c. Make a figure with a normal distribution and indicate with a vertical line the critical test statistic values for α = 0.05 (assume a two-sided distribution, thus 2.5% surface to either side of the mean). ***R***~~