1. Suppose we conduct a microarray study to compare gene expression in tumor vs normal tissue. The dataset contains measurements for 1000 genes in 52 tumor and 50 tumor-adjacent normal samples from the same patients.

   (a) To identify differentially regulated genes, we wish to test each gene for differences in expression in tumor vs normal samples. Should we use a paired or an unpaired test? Justify your answer.

   Paired! We are looking at the same individual for two different samples, so doing a paired t-test will account for individual variability and give us a more significant p-value

   (b) Suppose we are able to use a t-test. We compute the t statistic and p-value for each gene. The top 10 most significant ones are shown below. Perform a Bonferroni adjustment for all the p values above, writing the adusted p-values as an additional column in the table. Circle the genes you would report as significant.

Bonferroni:
alpha / n OR p * n
n = 1000

Note: cannot change alpha AND p, one or other!!!

	t	p	p.adjust
Gene.17	-8.18	1.04e-05	0.0104
Gene.1	7.71	2.16e-05	0.0216
Gene.766	-6.46	9.09e-05	0.0909
Gene.2	6.20	1.25e-04	0.125
Gene.924	-4.72	9.29e-04	0.929
Gene.19	-3.75	4.12e-03	1
Gene.951	-3.58	5.44e-03	1
Gene.720	-3.51	6.06e-03	1
Gene.511	3.49	6.31e-03	1
Gene.349	-3.42	7.00e-03	1

*cannot have p-value > 1, so these max out at 1

   (c) Suppose we are interested in a particular pathway comprising 24 genes, which happen to be Gene.1 – Gene.24. What is the probability (i.e. p value) that this pathway is enriched for significant genes? (You may leave fractions, etc, unsimplified.) *Hint: How many of the genes in the pathway are significant? How many genes not in the pathway are significant? Use your answers from (b).*

This is an example for a hyper-geometric distribution (i.e. fisher's exact test)

$$\frac{\text{How many ways to choose 2 significant from 24 in pathway} \times \text{How many ways to choose 0 significant from the 1000-24 not in pathway}}{\text{How many ways to choose 2 significant out of 1000 total}^{2}} = \frac{({}_{24}C_2)({}_{1000-24}C_0)}{{}_{1000}C_2}$$

$= 0.0005525526$

2. In order to discover novel genetic variants that predispose women to postmenopausal breast cancer, we genotype a million SNP loci across the human genome in 1000 cases (women with breast cancer) and 1000 controls (women without breast cancer). At each SNP locus, the sample genotype may be homozygous for the major allele (coded "AA"), heterozygous (coded "Aa") or homozygous minor (coded "aa").

   List **all steps** you would take to analyze this data to identify significant SNPs. Be sure to clearly state your **hypotheses**, any **assumptions** you make, the **test(s)** you would use, any **diagnostics** you would perform, and any **distributions** you would use. Include any **relevant numbers** (such as degrees of freedom if using a test that requires them) and **formulas** if relevant. *Note*: you are not required to give R code.

Hypotheses:

H_0: no association between SNP and breast cancer

H_A: association between SNP and breast cancer

Assumptions:

Each expected value must be > 5 (n large enough)
Each SNP must be independent from each other

Test:

Chi-square test (could use Fisher if assumption of sample size is not met
> non-directional - either allele could be associated

Steps:

1. Check assumptions
2. Make contingency table
3. Calculate expected values:
4. Calc X^2 statistic
5. Compare to X^2 distribution with df = 1 (2x2 contingency table, df = 2 for 3x2 - this is not a goodness of fit test)
6. P < 0.05/1,000,000 = significant (multiple hypothesis correction - bonferroni. Either adjustment is okay).

$$e = \frac{(row_T)(col_T)}{grand_T}$$

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

Note: The "correct" way to do this is probably to combine AA, Aa, and aa into A's and a's for a 2x2 contingency table, but most of you did a 3x2 table, which I counted as fine. However, remember df is (r-1)(k-1) = 2 in this case

Note: suggesting a t-test or ANOVA is not appropriate because we are looking at COUNT data, not quantitative data. We could do a t-test if we were looking at the expression of a marker for cancer vs. non-cancer patients, but this data is just whether or not the individual has cancer and has the allele.

[15pt] 3. Consider the following outputs from data which contains observations on the percentage of people biking to work each day, the percentage of people who smoke, and the percentage of people with heart disease in 498 imaginary towns:

```
Call:
lm(formula = heart.disease ~ biking + smoking, data = heart_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -0.4463  0.0362  0.4422  1.9331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
biking      -0.200133   0.001366 -146.53   <2e-16 ***
smoking      0.178334   0.003539   50.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

This model fits the data best, because the interaction term below is not significant.

```
Call:
lm(formula = heart.disease ~ biking * smoking, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20619 -0.44862  0.02892  0.44099  1.94142

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.0527397  0.1248112 120.604   <2e-16 ***
biking         -0.2019916  0.0029472 -68.536   <2e-16 ***
smoking         0.1740065  0.0070359  24.731   <2e-16 ***
biking:smoking  0.0001177  0.0001653   0.712    0.477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6544 on 494 degrees of freedom
Multiple R-squared:  0.9796,Adjusted R-squared:  0.9795
F-statistic:  7922 on 3 and 494 DF,  p-value: < 2.2e-16
```

(Problem continues on the next page.)

(a) Write the regression equation (in format y = mx + b) for the model that best fits the data.

heart.disease = 14.98 - 0.2(biking) + 1.78(smoking)

(b) Write a few sentences detailing the results from your chosen model (a) as you would in a manuscript (i.e. how would you interperet this model?).

We performed a linear regression to see if biking and smoking were associated with levels of heart disease in a population. The analysis indicated that both biking (B = -0.2, p < 2.2e-16) and smoking (B = 0.178, p < 2.2e-16) are significant predictors of heart disease. As the percent of the population who smokes increases by 1%, we expect the heart disease rate in the population to also increase by 0.178%. Additionally, as the percent of the population who bikes increases by 1%, we expect the heart disease rate in the population to decrease by 0.2%. This model explains 97.96% of the total variation in heart disease in this study.

(c) My imaginary hometown has 5% of its population biking to work and 7% of its population smokes. What would you predict the rate of heart disease is in this town?

heart.disease = 14.98 - 0.2(5) + 0.178(7)

heart.disease = 15.226%

(d) Based on this summary, do you believe the model fits the data well? Provide at least two reasons to support your statement. Additionally, provide one more test/parameter not shown that could be useful to answer this question.

Yes, this model seems to fit the data pretty well.

(1) The R^2 is very high (> 0.97) which means that 97% of the total variation in heart disease can be explained by biking and smoking.
(2) The residual standard error is pretty low (0.654) which means that on average the data fall 0.65% from the regression line.
(3) The F test is highly significant (p < 2.2e-16) which shows that our model fits the data better than the intercept only model
(4) … (there could be others)

Additional tests/parameters to check fit:

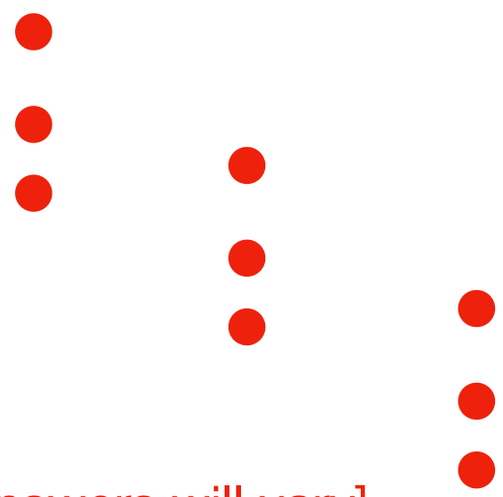
(1) look at residual plot for shape
(2) Anova to compare models to see if another one is significantly different
(3) … (others)

4. In the boxes below, sketch:

   (a) A plot illustrating a situation where $Cov(X, Y) < 0$ and where using a rank correlation coefficient would be advisable.

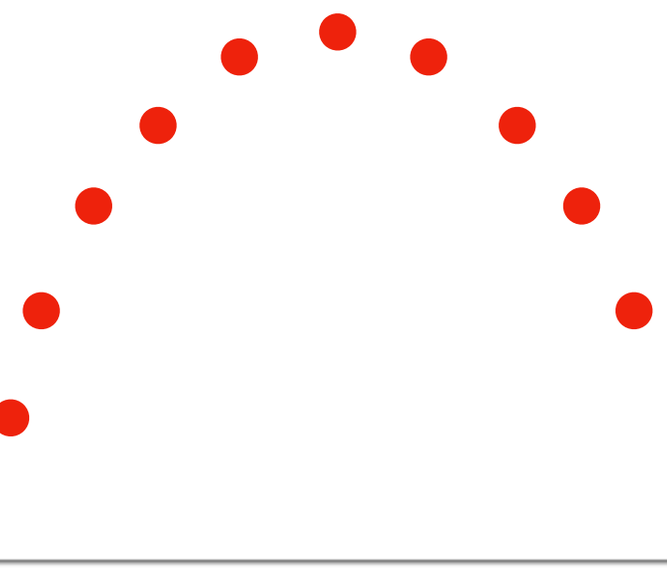
   (b) A plot illustrating a situation in which y depends on x but $r_{xy} = 0$.

a NOTE: rank correlation became extra credit, I realized I didn't cover it well



[Answers will vary]

b [Answers will vary]



Another common answer might be data with obvious outlier

5. If the odds of an event happening are 0.5, what is the probability of the event happening?

$$odds = \frac{P}{1-P} \qquad 0.5 = \frac{P}{1-P} \qquad P = 0.333$$

6. True or false (circle one):

T F : ANOVA is based on the idea that the variation of data can be decomposed into a systematic and a random part.

T F : In regression analysis, every time that an insignificant and unimportant variable is added to the regression model, the R^2 decreases

T F : Multicollinearity (when multiple independent variables are correlated) may cause the signs of some estimated regression parameters to be the opposite of what we would expect.

T F : The correlation coefficient indicates the change in y associated with a unit change in x.

T F : The value of the covariance between X and Y does not change if all values of X are multiplied by 100.

T F : Nonparametric and parametric tests will yield equivalent p-values

T F : The degrees of freedom in a χ^2 test depends on the sample size.

T F : FDR is a less strict multiple hypothesis comparison than Bonferroni.

7. Consider an experiment in which we measure the red-cell folate levels of 21 female and 45 male cardiac bypass operation patients randomized to three different ventilation protocols.

(a) Fill in the following ANOVA table:

MS = SS / df

F = MS(between)/ MS(within)

We did our best to give as much partial credit as possible based on equations

	df	SS	MS	F-ratio
between (sex)	1	8.41	8.41	0.29
between (treatment)	2	317.59	158.79	5.47
within group	62	1798	29.0	—
total	65	2124	—	—

df_total = n -1

n = 21 + 45 = 66

SST = SSW + SSB

(b) Based on the above, what is the standard deviation of red-cell folate levels for *all* patients in the study?

$$PooledSD = \sqrt{MS(within)} = \sqrt{29} = 5.38$$

(c) If we wanted to test the hypothesis that sex does not significantly contribute to the variance in red-cell folate levels, what would out test statistic be, and to what distribution would we compare it?

Test statistic: F = 0.29

Distribution: F distribution with df 1, 62

(d) Suppose you tested the "treatment" statistics from your ANOVA table and obtained $p = 0.0065$. (i) What were your hypotheses? (ii) How would you interpret this result? (iii) What could your next step be?

(i) Hypotheses

H_0: Each group comes from the same population (or means are equal, or groups don't vary, etc.)
H_A: Null is false: at least one or more groups are not from the same population

(i) Interpretation

A p-value of 0.0065 is less than alpha of 0.05, so we can reject the null hypothesis and conclude that at least one or more of the treatments has a different result.

(i) Next steps

Since we found a significant result, the next step would likely be to see which of the treatments differed. This could be done with pairwise t-tests with multiple hypothesis correction (like a Turkey HSD)