Statistics 416: Homework # 2 Due: 2/16/17

1. Suppose a researcher was interested in studying the effects of different temperature conditions (normal, cold, heat) on the expression of one interesting gene in two different breeds of chicken. Six houses that could have four chickens in each house were available. For each house, two chickens were randomly picked from each breed and assigned to the house, so that we have 2 chickens from each breed for each house. All chickens in one house received the same temperature condition. The researchers randomly assigned two houses to N (normal temperature), two to C (cold) and the remaining two to H (heat). After 3 days, the chicken spleen was obtained and one RNA sample was prepared for each chicken. The expression of the interesting gene was measured using Northern Blot for each RNA sample.

   a) Name the treatment factors considered in this experiment.
   b) Name the levels of each treatment factor.
   c) What are the experimental units in this experiment? How many experimental units are used in this experiment? (If there are two or more types of experimental units, specify the number of experimental units for each type.)
   d) What are the observational units in this experiment?
   e) Does this experiment involve blocking? If so, name the blocks.
   f) Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? *Justify your answer*.

2. For the experiment as described in question 1, suppose the researcher was interested in studying all expressed genes using an RNA-seq experiment, and the primary interesting question was to find genes differentially expressed between the two breeds under each temperature condition, and the secondary goal is to compare the temperature conditions for each breed. Suppose each RNA sample and library was prepared independently of the others, and the processing order was random. In another word, the design for phase 2 is a completely randomized design. There are 2 flowcells with 6 lanes available each. And we will apply multiplexing technique using two different barcodes for each lane. How would you assign samples to the lanes of the two flowcells in this RNA-seq experiment?

   Please use the symbolic notation discussed in class to describe your design. That is, you should draw one circle for each smallest (or smaller) experimental unit. Each circle should be labeled to describe the treatment applied to the experimental unit. Each lane should be represented by a line connecting the experimental units sequenced in the same lane. You may either use color to code the barcoding of samples. Or you may use the direction of the arrow to indicate barcode assignment (tail=barcode1 and head=barcode2).
   Justify your design with respect to the investigation of the research questions (Would data from your designed experiment be able to provide answers to the interesting

research questions? Why is your design a good one?). If you apply useful confounding in your design, describe them.

3. For the experiment as described in question 2, suppose the researcher was still interested in studying all expressed genes using an RNA-seq experiment, and the primary interesting question was to find genes differentially expressed between the two breeds under each temperature condition, and the secondary goal is to compare the temperature conditions for each breed. Different from the independent sample preparation as described in question 2, suppose each RNA sample and library was prepared in sessions of four samples/libraries each (as in example 3 and 4 of our lecture about 3-phase design). There are 2 flowcells with 6 lanes each available. And we will still apply multiplexing technique using two different barcodes for each lane. How would you design this RNA-seq experiment for phase 2 and phase 3?

   Specify the name of your design for each of phase 2 and 3, if you can identify your design as one of the common designs we discussed in class. Otherwise, you may just describe how you design the experiment. Justify your design with respect to the investigation of the research questions (Would data from your designed experiment be able to provide answers to the interesting research questions? Why is your design a good one?). If you apply useful confounding in your design, describe them.

4. This question is intended to help you review or gain a better understanding of variance components and calculations used to determine standard errors.

(a) The expected value of a random variable Y is denoted by $E(Y)$. It is the mean or average value taken by Y. If $E(Y) = \mu$, then $E(a + bY) = a + b\mu$ for any real constant values a and b. This says that if Y has value μ on average, then $a+bY$ has value $a+b\mu$ on average. If $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, then $E(Y_1 + Y_2) = \mu_1 + \mu_2$. You can put these rules together if necessary to complete the following problems. Suppose Y_i is a random variable with mean μ_i and μ_i is known for all i. Calculate:

i. $E(2Y_1 + 3)$.
ii. $E(Y_1 + Y_2 + Y_3)$.

(b) Consider the random variable $\{Y - E(Y)\}^2$, i.e., the squared deviation of the random variable Y from its mean value. The mean of this random variable is defined as the variance of Y; i.e., the variance of a random variable Y is $Var(Y)=E[\{Y - E(Y)\}^2]$. Thus the variance of Y is the expected squared deviation of a random variable Y from its mean. Random variables that tend to yield values close to their means will have small variances. Random variables that tend to yield values far from their means will have large variances. A constant real number c can be viewed as a random variable that always takes the value c. Calculate

i. $E(c)$.
ii. $Var(c)$.

(c) If the random variables Y_1 and Y_2 are independent and c_1 and c_2 are any real constants, then $Var(c_1Y_1 + c_2Y_2) = c_1^2Var(Y_1) + c_2^2Var(Y_2)$..

If $Var(Y_i) = \sigma^2$ for i = 1, 2, that is, the two random variables Y_1 and Y_2 have the same variance; and σ^2 is a known constant, calculate

i. $Var(Y_1 - Y_2)$

ii. $Var\{(Y_1 + Y_2)/2\}$

(d) If the random variables $Y_1, Y_2, \ldots, Y_n$ are independent and $c_1, c_2, \ldots, c_n$ are any real constants, then

$Var(\sum_{i=1}^{n} c_iY_i) = Var(c_1Y_1 + c_2Y_2 + ... + c_nY_n) = c_1^2Var(Y_1) + c_2^2Var(Y_2) + ... + c_n^2Var(Y_n)$. If Var($Y_i$) = σ^2 for i = 1, 2, …, n, that is, the random variables Y_i have the same variance; and σ^2 is a known constant, calculate:

(i) $Var(\sum_{i=1}^{n} Y_i)$

(ii) $Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = Var(\overline{Y})$

(e) For the design examples in lecture notes given by "05 Experimental Design_Replication.pdf", derive the variance of $\overline{Y}_{1..} - \overline{Y}_{2..}$ for each of designs 1 and 2 given on slide 40. That is, derive the following by applying the results you obtained from (c) and (d).

(i) $Var(\overline{Y}_{1..} - \overline{Y}_{2..}) = (\sigma_b^2 + \sigma_e^2)/4$ for Design 1

(ii) $Var(\overline{Y}_{1..} - \overline{Y}_{2..}) = (2\sigma_b^2 + \sigma_e^2)/4$ for Design 2