10th Problem Set

1. Here is a generalization of a problem we did in class on April 24th. Suppose you have one of three biased coins. You are uncertain which it is, but you know the frequencies with which they turn up "HEADS". Those frequencies are given by the second column below, and your state of uncertainty about which coin you have is characterized by the first column below.

$$\begin{aligned} \mathbf{Pr}(\text{1st coin}) &= 0.4 & \quad \mathbf{Pr}(\text{HEADS} \mid \text{1st coin}) &= 0.3 \\ \mathbf{Pr}(\text{2nd coin}) &= 0.25 & \quad \mathbf{Pr}(\text{HEADS} \mid \text{2nd coin}) &= 0.55 \\ \mathbf{Pr}(\text{3rd coin}) &= 0.35 & \quad \mathbf{Pr}(\text{HEADS} \mid \text{3rd coin}) &= 0.8 \end{aligned}$$

(a) Let X be the number of times "HEADS" turns up when the coin is tossed n times. Show that

$$\begin{bmatrix} \log\left\{\dfrac{\mathbf{Pr}(\text{1st coin} \mid X = x)}{\mathbf{Pr}(\text{2nd coin} \mid X = x)}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{2nd coin} \mid X = x)}{\mathbf{Pr}(\text{3rd coin} \mid X = x)}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{3rd coin} \mid X = x)}{\mathbf{Pr}(\text{1st coin} \mid X = x)}\right\} \end{bmatrix} = \begin{bmatrix} \log\left\{\dfrac{\mathbf{Pr}(\text{1st coin})}{\mathbf{Pr}(\text{2nd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{2nd coin})}{\mathbf{Pr}(\text{3rd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{3rd coin})}{\mathbf{Pr}(\text{1st coin})}\right\} \end{bmatrix} +$$

$$+ \quad x \begin{bmatrix} \log\left\{\dfrac{\mathbf{Pr}(\text{HEADS} \mid \text{1st coin})}{\mathbf{Pr}(\text{HEADS} \mid \text{2nd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{HEADS} \mid \text{2nd coin})}{\mathbf{Pr}(\text{HEADS} \mid \text{3rd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{HEADS} \mid \text{3rd coin})}{\mathbf{Pr}(\text{HEADS} \mid \text{1st coin})}\right\} \end{bmatrix} + (n - x) \begin{bmatrix} \log\left\{\dfrac{\mathbf{Pr}(\text{TAILS} \mid \text{1st coin})}{\mathbf{Pr}(\text{TAILS} \mid \text{2nd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{TAILS} \mid \text{2nd coin})}{\mathbf{Pr}(\text{TAILS} \mid \text{3rd coin})}\right\} \\[3em] \log\left\{\dfrac{\mathbf{Pr}(\text{TAILS} \mid \text{3rd coin})}{\mathbf{Pr}(\text{TAILS} \mid \text{1st coin})}\right\} \end{bmatrix}$$

Call this vector the **"generalized logit"** of the posterior probability distribution. Similarly, the first term in the sum of three terms to the right of "=" is the **"generalized logit"** of the prior probability distribution.

(b) Let $\mathbf{p} + x\mathbf{a} + (n - x)\mathbf{b}$ be the vector to the right of "=" in part (a). Show that the set $\{\, \mathbf{a}, \mathbf{b} \,\}$ is linearly independent, so that Figure 1 on page 2 makes sense. Next, show that if

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \log\{\mathbf{Pr}(\text{1st} \mid X = x)/\,\mathbf{Pr}(\text{2nd} \mid X = x)\} \\ \log\{\mathbf{Pr}(\text{2nd} \mid X = x)/\,\mathbf{Pr}(\text{3rd} \mid X = x)\} \\ \log\{\mathbf{Pr}(\text{3rd} \mid X = x)/\,\mathbf{Pr}(\text{1st} \mid X = x)\} \end{bmatrix}$$

then

$$\begin{bmatrix} \mathsf{\mathbf{Pr}}(\text{1st} \mid X = x) \\ \\ \mathsf{\mathbf{Pr}}(\text{2nd} \mid X = x) \\ \\ \mathsf{\mathbf{Pr}}(\text{3rd} \mid X = x) \end{bmatrix} = \frac{1}{1+e^{-u}+e^{w}} \begin{bmatrix} 1 \\ e^{-u} \\ e^{w} \end{bmatrix} = \frac{1}{e^{-w}+e^{v}+1} \begin{bmatrix} e^{-w} \\ e^{v} \\ 1 \end{bmatrix} = \frac{1}{e^{u}+1+e^{-v}} \begin{bmatrix} e^{u} \\ 1 \\ e^{-v} \end{bmatrix}.$$

In other words, you can find a probability distribution if you know its generalized logit.

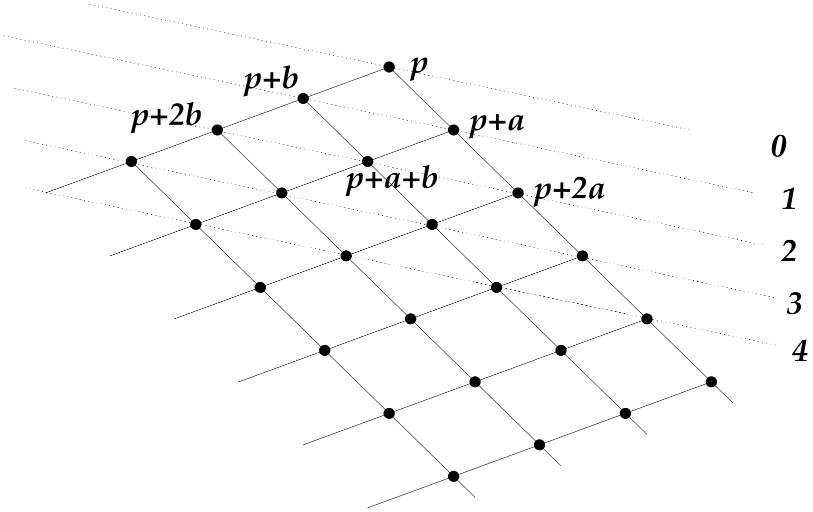

Figure 1:

(c) Each probability distribution ($\mathsf{\mathbf{Pr}}$(1st), $\mathsf{\mathbf{Pr}}$(2nd), $\mathsf{\mathbf{Pr}}$(3rd)) is a point in the triangle depicted in Figure 2 on page 6, with $(1, 0, 0)$ at one corner, $(0, 1, 0)$ at another, and $(0, 0, 1)$ at another. **BY THINKING ABOUT FIGURE 1 AND FIGURE 2, i.e., NOT BY SOME OTHER METHOD**, argue that if the number n of times the coin has been tossed is very big, then at least one of the three posterior probabilities is very close to 0. A HINT is in a footnote[1].

(d) Consider this statement:

> In Figure 1, the point labeled "**p**" is not on a straight line between the points labeled "$\mathbf{p} + \mathbf{a}$" and "$\mathbf{p} + \mathbf{b}$".

At what earlier point in this problem set did you address the content of this statement in somewhat different language? Now consider this statement:

> In Figure 2, the point labeled "**p**" is on a straight line between the points labeled "$\mathbf{p} + \mathbf{a}$" and "$\mathbf{p} + \mathbf{b}$".

Prove this second statement by interpreting those three points as probability distributions of particular events involved in this problem. A HINT is in a footnote[2].

[1] HINT: The numbers in Figure 1 count how many times the coin has been tossed. Where in Figure 2 would you see the images of the dotted lines shown in Figure 1?

[2] HINT: Being between them, means being a weighted average of them. The weights are probabilities.

2. Consider the model $Y_i \sim N_1(\beta_0 + \beta_1 x_i, \sigma^2)$ and $Y_1, \ldots, Y_{15}$ are independent. Or, more tersely,

$$Y \sim N_{15}(X\beta, \sigma^2 I_{15}),$$

where $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$. Suppose the data are as follows:

x	y
3	44
3	28
3	33
4	45
4	35
4	31
5	40
5	29
5	30
7	38
7	25
7	31
8	28
8	20
8	18

(a) Carefully draw the scatterplot.

(b) Test the null hypothesis that $\beta_1 = 0$ at the 4% level.

Now alter the model, but keep the data the same. Treat the x-values as identifying categories, rather than as quantities. Let μ_x be the expected value of individual observations in category x, for $x \in \{3, 4, 5, 7, 8\}$. The model will then state that $Y_{ij} \sim N(\mu_i, \sigma^2)$ for $i \in \{3, 4, 5, 7, 8\}$, $j \in \{1, 2, 3\}$, and all 15 of these random variables are independent.

(c) Test the null hypothesis $\mu_3 = \mu_4 = \mu_5 = \mu_7 = \mu_8$ against the alternative that says these are not all equal, at the 4% level.

(d) Now consider testing the null hypothesis that this straight-line model is right, against the alternative hypothesis that the a certain other model is right:

$$\begin{array}{lll} H_0 : & Y_{ij} \sim N(\beta_0 + \beta_1 x_i, \sigma^2) & \text{(the ``straight-line'' model)} \\ H_1 : & Y_{ij} \sim N(\mu_i, \sigma^2) & \text{(the ``categories'' model)} \end{array}$$

Recall the notation of #6 on the 5th problem set:

$$\begin{aligned} \varepsilon &= Y - X\beta = \text{vector of ``errors''} \\ \widehat{\varepsilon} &= Y - X\widehat{\beta} = (I - H)Y = \text{vector of ``residuals''} \end{aligned}$$

CONTINUED⟶

Find a matrix M such that the categories model can be expressed thus: $Y \sim N_{15}(M\mu,\, \sigma^2 I_{15})$. Let $K = M(M'M)^{-1}M'$. Write

$$Y = \underbrace{HY} + \underbrace{(K-H)Y} + \underbrace{(I-K)Y}$$

Show that the components of these three vectors can be written thus:

$$\underbrace{(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)} + \underbrace{\overline{Y}_{i\bullet} - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)} + \underbrace{Y_{ij} - \overline{Y}_{i\bullet}}$$

Show that

$$\begin{aligned} \|(K-H)Y\|^2/\sigma^2 &\sim \chi^2_3 \quad \text{if } H_0 \text{ is true,} \\ \|(I-K)Y\|^2/\sigma^2 &\sim \chi^2_{10} \quad \text{regardless of which of the two models is true,} \end{aligned}$$

and that these are independent. A Hint is in a footnote[3].

(e) Carry out the test contemplated in part (d), using the data given about part (a), at the 4% level.

3. In *Mathematical Statistics and Data Analysis, Second Edition*, by John A. Rice, we read about

"... 85 Hodgkin's patients who had a sibling of the same sex who was free of the disease and whose age was within 5 years of the patient's. These investigators presented the following table:

	Tonsillectomy	No tonsillectomy
Hodgkin's	41	44
Control	33	52

"They calculated a chi-square statistic of 1.53, which is not significant.. . .
. . . [they] had made an error in their analysis by ignoring the pairings.. . . [their] samples were not independent, because the siblings were paired . . . . . . set up a table that exhibits the pairings:

		Sibling	
		No Tonsillectomy	Tonsillectomy
Patient	No Tonsillectomy	37	7
	Tonsillectomy	15	26

". . . . . . The appropriate null hypothesis states that the probabilities of tonsillectomy and no tonsillectomy are the same for patients and siblings. . . ."

(a) If you know the numbers in the second table, how would you find the ones in the first?

(b) If you know the numbers in the first table, why is it not possible to find the ones in the second? In particular, given the numbers in the first table, what is the smallest number that could have appeared where "37" appears in the second table, and what is the largest?

[3] Hint: To show independence of $\|(K-H)Y\|^2$ and $\|(I-K)Y\|^2$, it suffices to show independence of $(K-H)Y$ and $(I-K)Y$

(c) Suppose the probabilities of being in the four cells in the first table are given by

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix}$$

(so that $p + q = 1/2$ and $r + s = 1/2$), and their counterparts for the second table are given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

(so that $a + b + c + d = 1$).

i. Express the null hypothesis of independence in the first table in the form of an equation in p, q, r, and s. A HINT is in a footnote[4].

ii. The second-to-last line of the quote from Rice's book refers to "the appropriate null hypothesis." Express that null hypothesis in the form of an equation in a, b, c, and d. A HINT is in a footnote[5].

iii. Show that the null hypothesis in part (i) above is true if and only if the null hypothesis in part (ii) above is true.

[4] HINT: After simplifying, it's a first-degree equation involving only two of the four variables.

[5] HINT: Also a first-degree equation involving only two of the four variables. No simplifying should be needed if you think about the fact that the probabilities of no tonsillectomy should be the same for both patients and siblings, and if the two probabilities of no tosillectomy are the same, then it's redundant to say the two probabilities of tonsillectomy are the same.

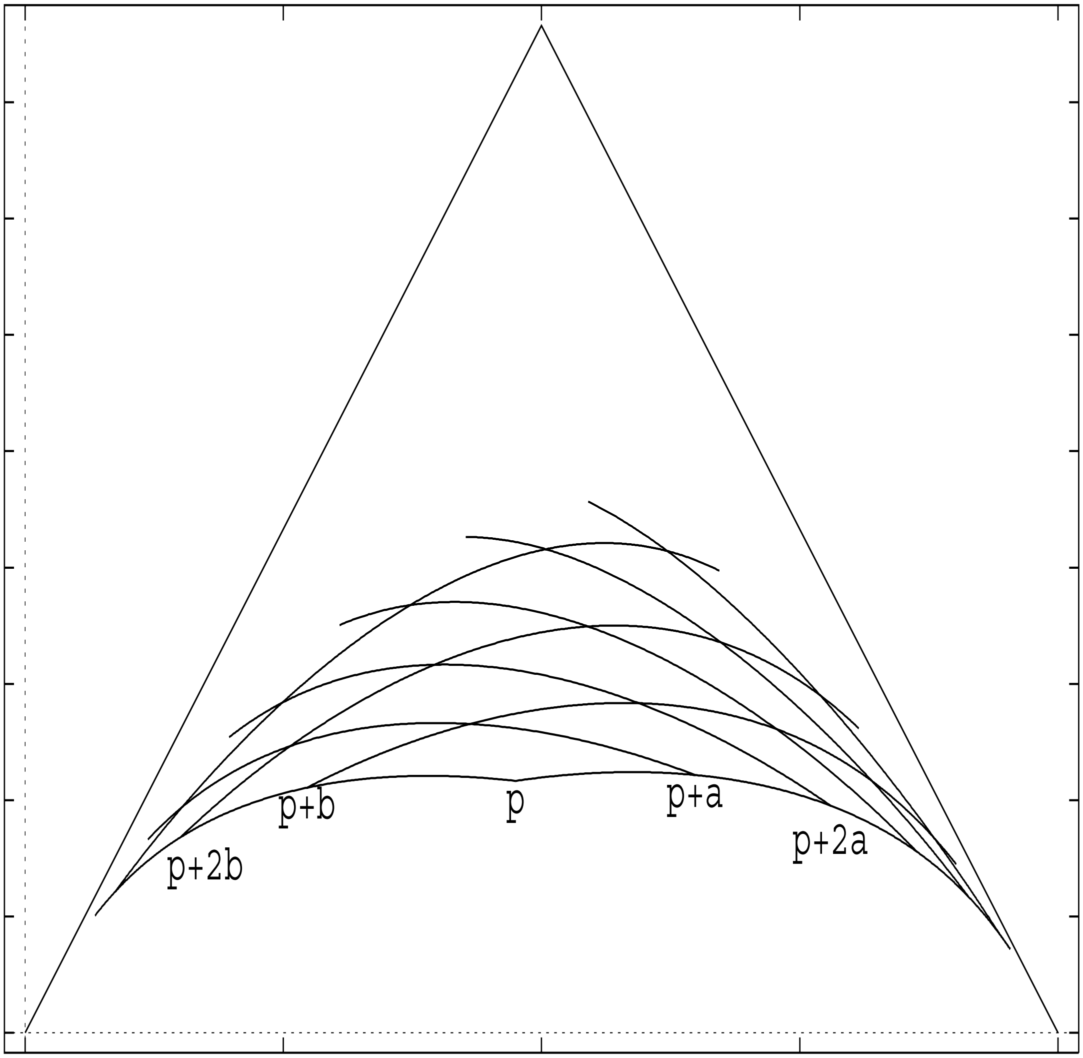

Figure 2: