Here's how to do the homework online on Courselets:

- We will work through a series of questions applying the concepts from lecture to a genetic problem.
- We suggest you work out your answer on paper just as you would normally do for homework, then finally type your answer in the box at the bottom of the screen.
- Please give a complete answer that fully answers the question and shows your reasoning, just as you would for a standard homework assignment. To get full credit, **you must fully answer the question**. However, **we will not take off points for errors** in your answer.
- After you enter your answer, you will see the correct answer and perform self-assessment as we do in the classroom exercises. Please self-assess honestly; remember that you **will not lose credit** due to getting the answer wrong.
- The due date for completing Homework 1 is noon Friday Oct. 6 (Discussion section). To receive credit, your answers must be entered by that time.
- Answer the questions in the order provided.

In humans, women have two copies of the sex chromosome X, whereas men only have one copy (inherited from their mother). Most disease mutations are "recessive", i.e. having one copy unmutated is sufficient to prevent disease even if the other copy has the disease mutation. Therefore diseases that are caused by recessive mutations on the X chromosome are more likely to cause disease in men (who have only one copy of X) than in women (who have two copies). A woman who has a disease mutation on one X chromosome but not the other is said to be a "carrier" for that disease, because she can pass that disease on to her children, even though she herself shows no symptoms.

Now consider the following scenario. Mrs. X shows no symptoms of disease, but due to her family background, she has 10% probability of being a carrier (C^+) for an X-linked genetic disease. Say that we use the following notation for designating a son of hers who shows no disease symptoms as S^-, vs. a son who shows disease symptoms as S^+.

What are the likelihoods for a son of hers to be free of the disease symptoms if she is a carrier, $p(S^-|C^+)$, vs. if she is not, $p(S^-|C^-)$?

Being a carrier means that **one** of her two X chromosomes has the disease mutation, and her son will randomly receive one of her two X chromosomes with equal probability. If he gets the "normal" copy (i.e. no disease mutation), he will have no symptoms of the disease, whereas if he gets the other copy (i.e. with the disease mutation), he will have disease symptoms. So:

$$p(S^-|C^+) = 1/2$$

Not being a carrier means both of her X chromosomes are normal, so:

$$p(S^-|C^-) = 1$$

Now say Mrs. X has a son, who shows no disease symptoms: S^-. Based on this observation, compute the posterior probability that Mrs. X is a disease carrier, $p(C^+|S^-)$.

Using Bayes' Law for whether she is a carrier vs. is not a carrier:

$$p(C^+|S^-) = \frac{p(S^-|C^+)p(C^+)}{p(S^-|C^+)p(C^+) + p(S^-|C^-)p(C^-)} = \frac{0.5(0.1)}{0.5(0.1) + 1(0.9)} \approx 0.053$$

Now say Mrs. X has a second son, who also shows no disease symptoms: S^-. Based on her two observed sons, compute the posterior probability that Mrs. X is a disease carrier, $p(C^+|S^-, S^-)$.

$$p(C^+|S^-, S^-) = \frac{p(S^-, S^-|C^+)p(C^+)}{p(S^-, S^-|C^+)p(C^+) + p(S^-, S^-|C^-)p(C^-)} = \frac{0.25(0.1)}{0.25(0.1) + 1(0.9)} \approx 0.027$$

Now say Mrs. X has a third son, who **has disease symptoms**: S^+. Based on her three observed sons, compute the posterior probability that Mrs. X is a disease carrier, $p(C^+|S^-, S^-, S^+)$.

$$p(C^+|S^-, S^-, S^+) = \frac{p(S^-, S^-, S^+|C^+)p(C^+)}{p(S^-, S^-, S^+|C^+)p(C^+) + p(S^-, S^-, S^+|C^-)p(C^-)} = \frac{0.125(0.1)}{0.125(0.1) + 0(0.9)} = 1$$

You may have wondered where "prior probabilities" come from -- ideally they are themselves estimated based on observational data, which strictly speaking makes them "posterior" to those observations. Thus a common pattern is to use the *posterior* computed from one set of observations, as a *prior* for a new Bayes Law calculation with a different set of observations.

Now perform a separate calculation in which you use the *posterior* after the *first son* $p(C^+|S^-) \approx 0.053$, as the *prior* for computing a posterior after the *second son* S^-. Do you get the same answer as when you used both sons in a single calculation?

$$p(C^+|S^-,S^-) = \frac{p(S^-|C^+)p(C^+|S^-)}{p(S^-|C^+)p(C^+|S^-) + p(S^-|C^-)p(C^-|S^-)} = \frac{0.5(0.053)}{0.5(0.053) + 1(0.947)} \approx 0.027$$

Yes, same result as when we computed $p(C^+|S^-,S^-)$ in a single step.

Prove that if an observation O is equally likely under all possible models, then the posterior after observing O is simply equal to the prior.

By hypothesis, $p(O|H) = c$, a constant.

$$p(H=h|O) = \frac{p(O|H=h)p(H=h)}{\sum_H p(O|H)p(H)} = \frac{cp(H=h)}{\sum_H cp(H)} = \frac{p(H=h)}{\sum_H p(H)} = p(H=h)$$

A scientist first observes variable X_1 then another variable X_2, and wishes to calculate the posterior probability of a hypothesis $p(H|X_1, X_2)$. Since she already calculated the posterior based on her first observation, i.e. $p(H|X_1)$, she proposes to speed up the calculation of $p(H|X_1, X_2)$ by simply using $p(H|X_1)$ as the prior and X_2 as the *obs*. Is this valid?

A. Yes, this is valid, i.e. it will always give the same result as using the original prior and X_1, X_2 as the *obs*.

B. No, this is never valid.

C. It depends.

It depends on whether the likelihoods of X_1, X_2 are conditionally independent given H. If so, then the scientist's proposed approach is valid. If they are not conditionally independent, then the approach is not valid. It's important that you understand intuitively that this is a question about independence. Whenever you see a calculation being performed separately for one variable and then another in this way, realize that this assumes they're (conditionally) independent.

Say we have two observable variables, X, Y that depend on a hidden variable Θ. Prove that if the observables X, Y are conditionally independent given the hidden variable Θ then the posterior probability can be factored as follows:

$$p(\Theta|X,Y) = \frac{p(Y|\Theta)p(\Theta|X)}{\sum_\Theta p(Y|\Theta)p(\Theta|X)}$$

By conditional independence, we know $p(Y|X,\Theta) = p(Y|\Theta)$

We can simply expand the general chain rule in the order X, Θ, Y, giving us:

$$p(\Theta|X,Y) = \frac{p(X,Y,\Theta)}{p(X,Y)} = \frac{p(Y|\Theta,X)p(\Theta|X)p(X)}{\sum_\Theta p(Y|\Theta,X)p(\Theta|X)p(X)}$$

Using our original conditional independence:

$$= \frac{p(Y|\Theta)p(\Theta|X)p(X)}{\sum_\Theta p(Y|\Theta)p(\Theta|X)p(X)}$$

Canceling p(X):

$$= \frac{p(Y|\Theta)p(\Theta|X)}{\sum_\Theta p(Y|\Theta)p(\Theta|X)}$$

Q.E.D.