Practice questions: Solutions

Statistical Thinking

14/11/2021

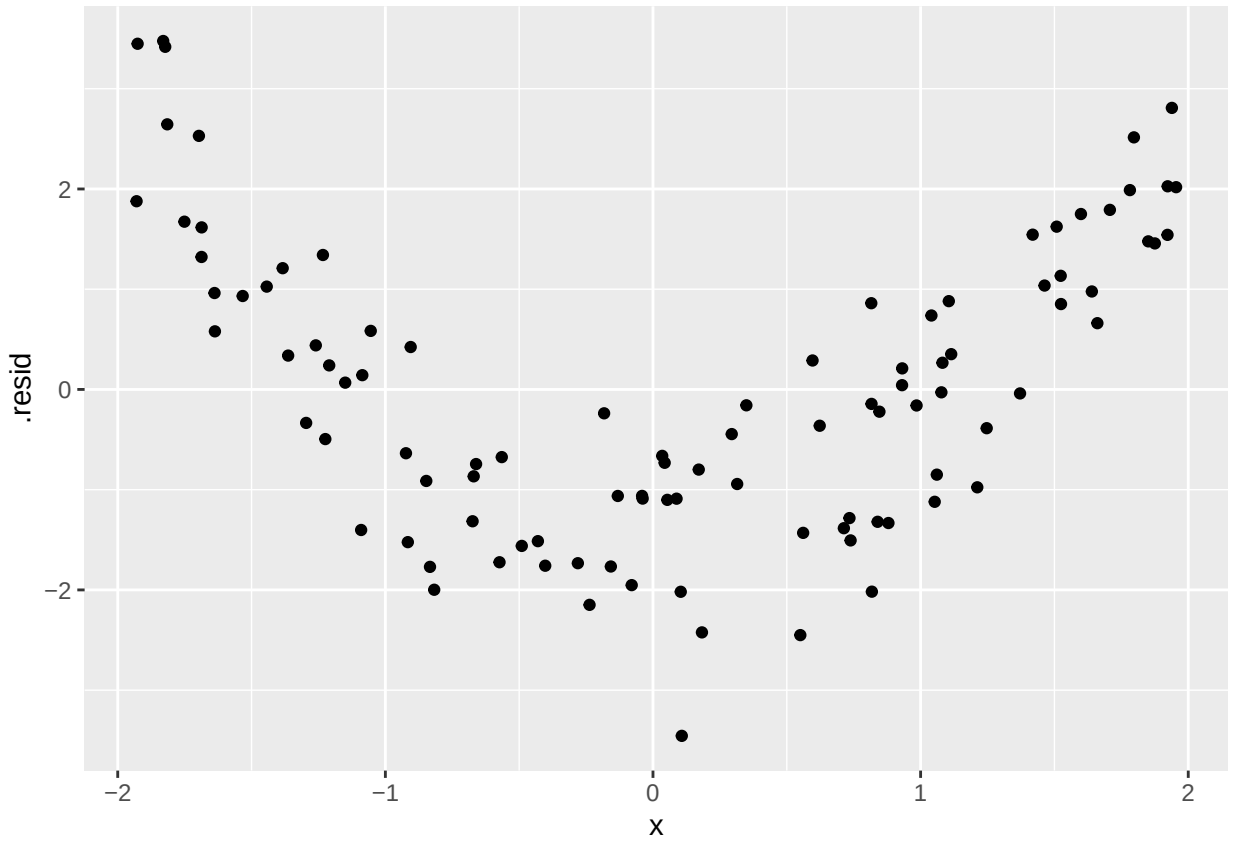
Question 1

After fitting a linear regression using the formula `y ~ x`, you compute the residual and plot the residuals on the y-axis and the covariate `x` on the x-axis. This plot shows a pronounced U shape. Sketch a dataset that would lead to this diagnostic plot.

Answer

A U shape indicates that the data is above the regression line for low and high values of `x` and is below it for intermediate values. An example of a dataset that does this is

```
library(tidymodels)
x <- runif(100, -2,2)
y <- rnorm(100, x^2, 0.8)
fit <- lm(y ~ x)
fit %>% augment() %>%
  ggplot(aes(x, .resid)) +
  geom_point()
```



In an exam you would have to sketch this by hand if instructed.

Question 2

You are interested in producing a confidence interval for the kurtosis[1] of a sample. Your friend Margaret gives you a procedure for computing $L(y)$ and $U(y)$ but she can't remember what the type-1 error is. Describe in words, using bullet point, the procedure to compute the type-1 error of the proposed confidence interval.

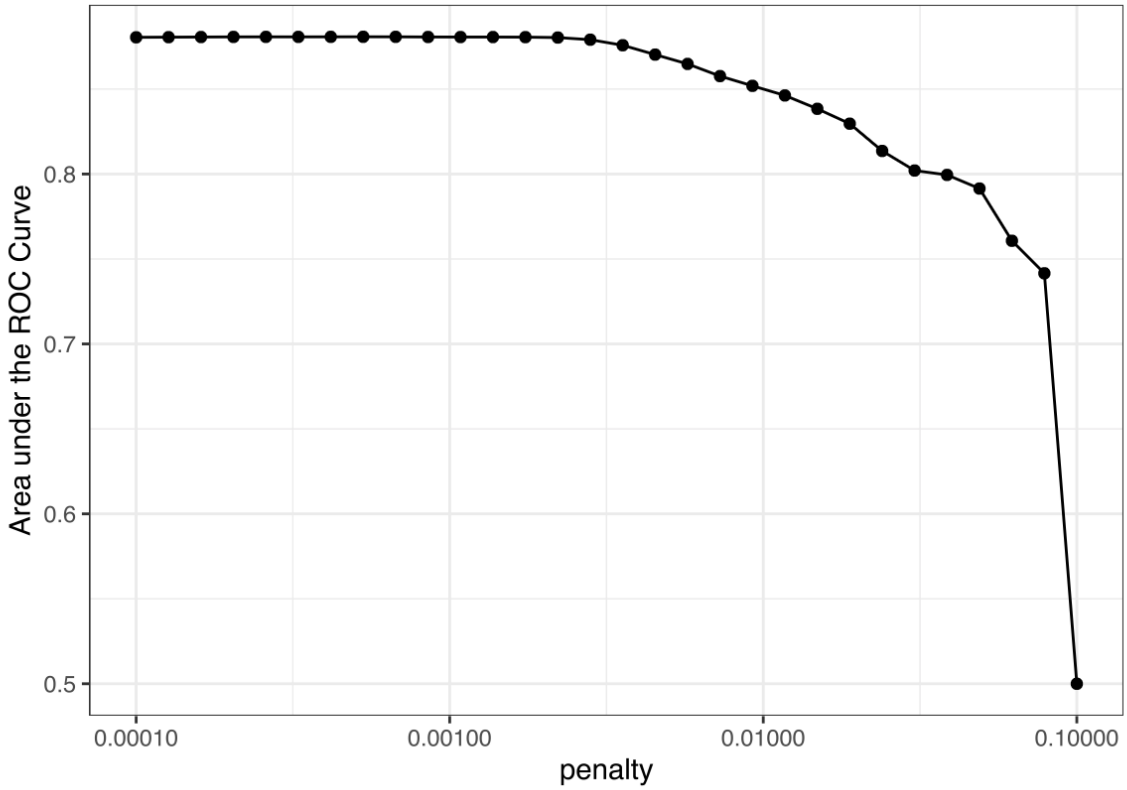
Answer:

There are a bunch of equivalent answers here, basically because I didn't specify the null distribution or the sample size. It would be totally ok to replace "the null distribution" with a reasonable null (like a normal distribution)

- Simulate `n_sim` data sets that are the same size as the data of interest from the null distribution
- For each simualted data set, compute the confidence interval $[L(y), U(y)]$
- Compute the number of times the null value is in the interval and divide it by `n_sim` to get an estimate of the type-1 error.

[1] a measure of how heavy the tails of a distribution are

Question 3



The above graph shows the AUC (bigger is better) for various values of a tuning parameter. If you know that a larger penalty produces a less complex model, choose and justify an appropriate penalty parameter for this problem.

Answer

We want to choose a value of the penalty for which the AUC is large, but we want to bias ourselves towards larger values of the paraemter in order to aid generalisation. Because of this, we should choose a value between 0.001 and 0.0015.

Question 4

Consider a regression problem where you are trying to estimate the causal effect of X on Y in the presence of other variables A, B, C, and D. The appropriate regression for estimating this causal effect is `Y ~ X + A + B`. Draw a DAG that includes at least one each of forks, pipes, and colliders that is consistent with the stated regression estimating the causal effect of X on Y.

Indicative answer (corrected):

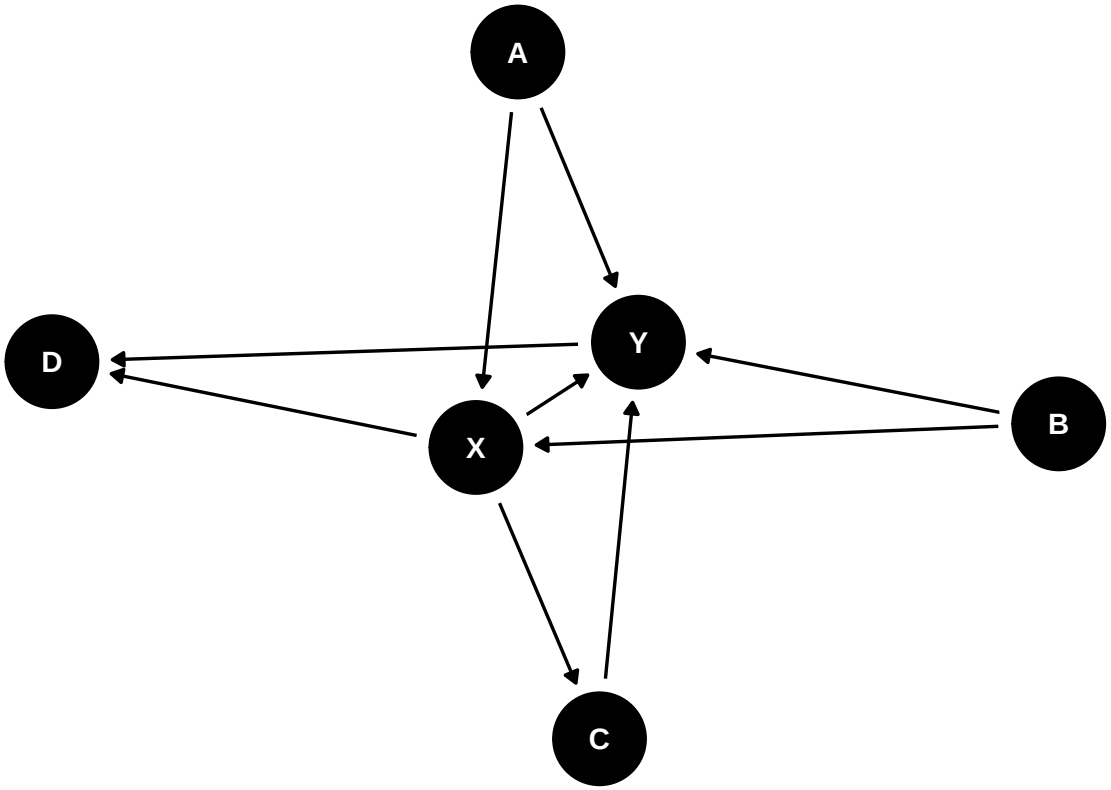
Obviously there are many possible answers here. Here's one. (Obviously you would sketch this by hand and not in R)

Remember: Never condition on colliders, never condition on pipes that block a directed path from X to Y, and condition on forks to close backdoor paths.

```
library(ggdag)

dagify(Y ~ X,
```

```
      Y ~ C, ## PIPE THROUGH C
      C ~ X,
      D ~ Y, ## COLLIDER AT D
      D ~ X,
      Y ~ A, ## FORK AT A
      X ~ A,
      Y ~ B, ## FORK AT B
      X ~ B) %>%
  ggdag() + theme_dag()
```



Question 5

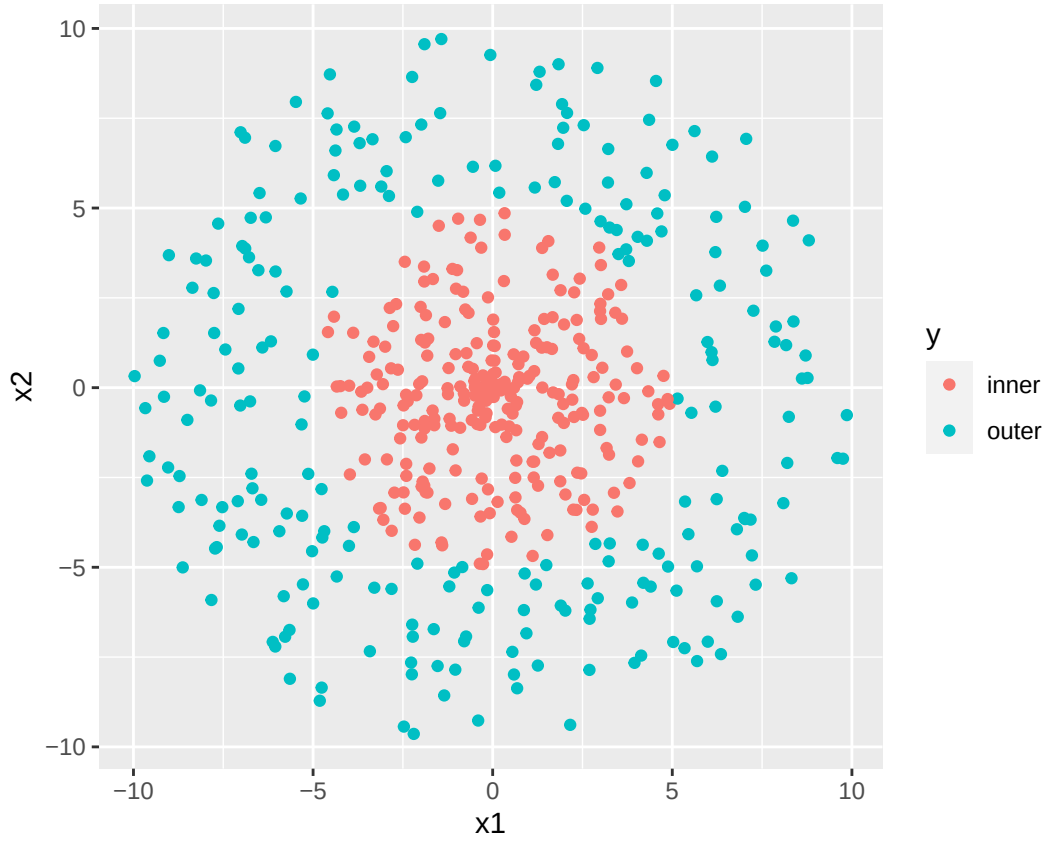
Draw a data set where points in (x_1, x_2)-space are labelled as either $y = 0$ or $y = 1$, where logistic regression using the formula `y ~ x1 + x2` would fail to yield a good classifier.

indicative answer (This would be drawn by hand)

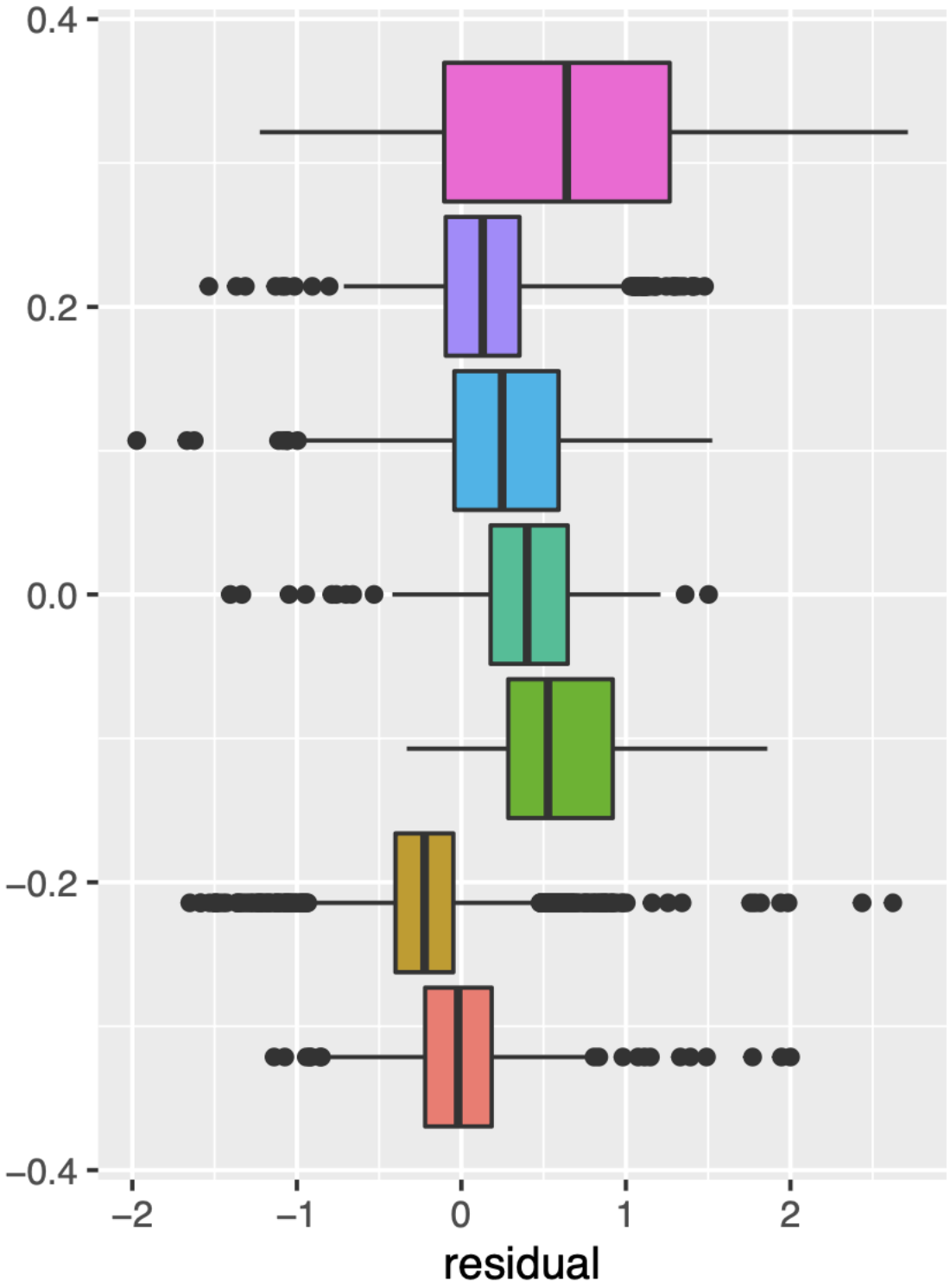
```
n <- 500
dat <- tibble(r = runif(n, max = 10),
              z1 = rnorm(n),
              z2 = rnorm(n),
              x1 = r * z1 / sqrt(z1^2 + z2^2),
              x2 = r * z2 / sqrt(z1^2 + z2^2),
              y = if_else(r < 5, "inner", "outer")) %>%
  mutate(y = factor(y)) %>%
  select(x1, x2, y)

dat %>% ggplot(aes(x1,x2, colour = y)) +
  geom_point() +
```

```
coord_fixed()
```



Question 6



The above boxplot shows the residuals for the linear regression `y ~ x + w` grouped by a categorical variable

z. How could you improve this regression?

Answer

The boxplot shows that the mean of the residuals is different at different levels of the variable z. This suggests that we could improve the quality of the regression by adding z to the model. The resulting model `y ~ x + w + z` should have residuals that have zero mean in each level of z.

Question 7

Consider the data in the image below.

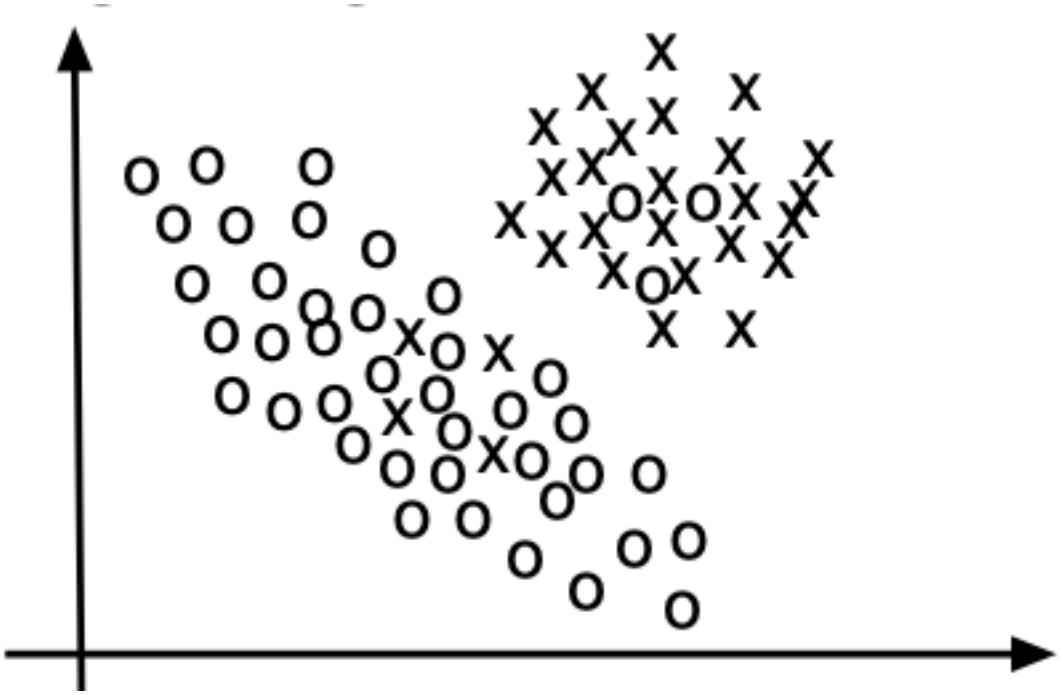


Figure 1: A data set

Which would you expect to have the smallest training error:

- 1-nearest neighbour classification
- 5-nearest neighbour classification
- logistic regression

Answer:

We can compute the accuracy!

- 1-NN will misclassify at least 8 points (possibly more, but the bottom O in the top-right cluster has an X that is nearer to it than any other X)
- 5-NN will misclassify points 7 points (it's possible that it will be more, but visual inspection doesn't show any points that have 3 of the closest values incorrect.)
- Logistic regression will misclassify exactly 7 points.

So the answer is Logistic regression OR 5-NN.

Question 7

Let p denote the probability that a tossed coin will return a *Head* outcome.

Suppose your friend Ella gives you a coin, and tells you that she is 70% sure it is a 'fair' coin. But, she thinks it is also possible that a *Head* outcome could occur with probability $p = 0.6$, but that no other values of the *Head* outcome probability are possible.

You decide to give the coin from Ella to your friend, Wei. You tell Wei that you do not know if it is a 'fair' coin, but you neglect to tell him anything about your belief (or Ella's) regarding the likely values of p, nor do you tell him anything about your previous coin toss.

Knowing only that the coin may not be 'fair', Wei does not consider any individual value of $p \in (0, 1)$ more likely than any other possible value. Wanting to update his belief regarding the value of p, Wei decides to run his own experiment comprised of ten independent tosses the coin. He observes 3 *Head* outcomes.

Part A (TRUE/FALSE)

Ella is acting like a Bayesian by suggesting probabilities for certain possible values of the unknown parameter, p.

Answer: TRUE. (Here Ella has expressed her belief about the possible values of p, using probability statements. This is consistent with a Bayesian approach. Whether she refers to herself as a Bayesian or not, Ella is acting like a Bayesian!)

Part B (TRUE/FALSE)

Given n independent tosses of the coin, the likelihood function associated with the unknown parameter p is equal to the probability (mass) function associated with the distribution for the number of Head outcomes, viewed as a function of the unknown parameter p.

Answer: TRUE. (This is just the definition of the likelihood function. The only tricky thing here is that since the number of head outcomes is a discrete random variable, its distribution is discrete. So we refer to the "density" function as a "mass" function. But this is as we have done during the semester so should be recognized by students.)

Part C (Multichoice)

Given n independent tosses of the coin, the probability distribution associated with a Head outcome, for a given value of p, is...

a. $Normal(\mu, \sigma^2)$

b. t_ν

c. $Binomial(n, p)$

d. $Beta(\alpha, \beta)$

e. $Gamma(\alpha, \beta)$

f. None of the above

Answer: c. (This is because each of the n coin tosses is an independent Bernoulli trial, with the same probability of a *Heads* outcome, p.)

Part D (Multichoice)

Ella's prior distribuiton for the possible values of p can be represented as...

a. $Beta(\alpha = 1,\ \beta = 1)$

b. $\Pr(p = 0.5) = 0.7$ and $\Pr(p = 0.6) = 0.3$

c. $\Pr(p = 0.7) = 0.3$ and $\Pr(p = 0.6) = 0.4$

d. $Beta(\alpha = 0.55,\ \beta = 0.45)$

e. None of the above.

Answer: b. (Because "... Ella ... is 70% sure it is a 'fair' coin" means $\Pr(p = 0.5) = 0.7$, and "it is also possible that a $Head$ outcome could occur with probability $p = 0.6$" means $\Pr(p = 0.6) = 1 - 0.7 = 0.3$, since "no other values of the $Head$ outcome probability are possible" and the two probabilities must add up to one.)

Part E (Short answer)

Report both Wei's *prior distribution* and his *posterior distribution*, and briefly explain how the two distributions are related to each other.

Indicative Answer: Since Wei doesn't have a preference for any of the possible values of p, his prior is a $Beta(1, 1)$ which is also known as a continuous uniform distribution on the interval (0,1). Then, since the number of *Heads* given p has a $Binomial(n, p)$ distribution, and $Beta$ prior and $Binomial$ likelihood form a *conjuage pair*, we know via Bayes theorem that the posterior distribution for p is a $Beta(1 + 3, 1 + 10 - 3) = Beta(4, 8)$.

Part F (Short answer)

Using only the outcomes from Wei's ten coin tosses, briefly explain how a Frequentist would attempt to determine the value of p.

Indicative Answer: Actually there is more than one way to answer this questions, as there are many estimation approaches from a Frequentist perspective. One way is to use a maximum likelihood approach, which results in a point estimate equal to the number of successes over the number of trails, or $\hat{p}_{MLE} = x/n = 3/10 = 0.3$.