Swarthmore Honors Examination 2022: Statistics

Instructions. This is a three-hour exam with seven questions. Budget your time wisely. It is best to answer every problem at least with partial solutions, not leaving any unanswered.

This is a closed-materials exam. You may not refer to any books, notes, online sources, software, or other resources, except for a simple calculator to do arithmetic. You will not need any probability tables.

Show your work and explain your reasoning with sufficient justification. Use clearly identified steps, calculations, formulas, or well-labeled graphs as necessary. Write your solutions neatly on separate sheets of paper. Label your responses clearly (1a, 1b, 1c, 2a, etc.).

Good luck!

---

1. PDFs. Suppose that

$$f_{Y|X}(y|x) = \begin{cases} 1, & x < y < x+1 \\ 0, & \text{otherwise} \end{cases}$$

and that X has the Uniform(0,1) distribution.
(Hint: Sketch the 2-dimensional joint support of x and y.)

a. Find $\mathbb{E}(Y)$.
b. Find $f_{X|Y}(x|y)$.
c. Find $\mathbb{P}(X + Y < 1)$.

2. Confidence intervals for the median. Say that we wish to estimate the median θ of a population, and that we have taken a random sample of iid draws $X_1, X_2, \ldots, X_n$. Assume a continuous population where the probability of ties is 0.

a. Let S_1 be the smallest sample observation and L_1 be the largest. Show that

$$\mathbb{P}(S_1 \leq \theta \leq L_1) = 1 - \frac{1}{2^{n-1}}$$

b. We can use (S_1, L_1) as a very simple confidence interval for the median. What's the smallest sample size n for which this interval's confidence level is over 95%? And what is the value of this confidence level?

c. More generally, let S_d be the d^{th} smallest sample observation and L_d be the d^{th} largest. Then we could consider confidence intervals of the type (S_d, L_d).
In principle we could derive a general formula for $\mathbb{P}(S_2 \leq \theta \leq L_2)$ that is a function of d (but don't actually do it on this exam!). Then we could collect larger samples and choose larger values d that would make the interval (S_d, L_d) achieve an approximate 95% confidence level.
Now let n_b be the sample size from your answer to part (b). What might be the benefits, if any, of collecting bigger samples with $n > n_b$ and using a larger $d > 1$ to get approx. 95% confidence? In other words, why not always just use $n = n_b$ and $d = 1$?
(Don't work out any derivations or proofs this time—just give brief intution.)

3. Communicating about statistics. Explain a confidence interval, in two different ways:

a. State a formal definition for a 95% confidence interval for a population proportion.
b. Give a brief explanation that's suitable for a journalist to use when writing for a general, non-technical audience. For example, fill in the gap here:
"Using 2011 US Census Bureau data, demographers reported that the average travel time to work for commuters in Maine is around 23.4 minutes, with a 90% confidence interval of 22.6 to 24.2 minutes. A '90% confidence interval' means that..."

4. Small area estimation. When a national statistical office carries out a large national survey, asking respondents about their household income, the office may wish to publish US county-level estimates for the average household income in each county. But if a small county has few respondents, and we only use the few survey respondents in that county to calculate an unbiased "direct estimate" of average income, then this estimate has small n and high variance.

In one approach to this kind of "small area estimation" problem, the statisticians might use a hierarchical Bayesian model to pool information across all the small areas.

For each area $i \in 1, \ldots, m$, let's model the true (unobserved) mean household income in each county θ_i as conditionally iid Normal draws around an unknown (hyperparameter) mean μ:

$$\theta_i | \mu \sim N(\mu, \tau^2)$$

To keep it simple, let's assume that τ^2 is known, and let's put a flat improper prior on μ so that $f(\mu) \propto 1$.

Next, assume that we don't get to see the raw data from each survey respondent, but only the aggregated direct estimates $\bar{Y}_i$, the sample means of all survey responses in each county i. Assume that their respective standard errors SE_i are known, but may vary from county to county (depending on sample size etc.). In other words, treat $SE_i^2 \equiv \text{Var}(\bar{Y}_i)$ as a known constant for each i.

Let's model these direct estimates $\bar{Y}_i$ as conditionally independent Normal draws around their respective true means:

$$\bar{Y}_i | \theta_i, \mu \sim N(\theta_i, SE_i^2)$$

a. Show that each $\theta_i | \bar{\boldsymbol{Y}}, \mu \sim N(\hat{\theta}_i, V_i)$, where

$$\hat{\theta}_i = \frac{\tau^2 \bar{Y}_i + SE_i^2 \mu}{\tau^2 + SE_i^2} \quad \text{and} \quad V_i = \frac{1}{1/SE_i^2 + 1/\tau^2}$$

It's also possible to show (but don't actually do it on this exam!) that $\mu | \bar{\boldsymbol{Y}} \sim N(\hat{\mu}, V_\mu)$, where

$$\hat{\mu} = V_\mu \sum_{i=1}^{m} \frac{1}{\tau^2 + SE_i^2} \bar{Y}_i \quad \text{and} \quad V_\mu = \frac{1}{\sum_{i=1}^{m} \frac{1}{\tau^2 + SE_i^2}}$$

b. We call $\hat{\theta}_i$ a "composite estimate": We take each direct estimate $\bar{Y}_i$, and we shrink it towards $\hat{\mu}$ which is a sort of pooled mean.
   i. Under what conditions would a particular $\hat{\theta}_i$ be very close to its direct estimate $\bar{Y}_i$ (almost no shrinkage)? When would it be very close to $\hat{\mu}$ (almost complete shrinkage)?
   ii. Likewise, under what conditions would a particular V_i be very close to its SE_i^2? When would it be substantially smaller than its SE_i^2?
   iii. Likewise, under what conditions would $\hat{\mu}$ be very close to a simple mean of the $\bar{Y}_i$ values?
   iv. Assume the sample in each county is iid so that SE_i depends on the county's sample size n_i in the usual way, and each n_i is proportional to its county population.
   Then which counties will tend to have more shrinkage: large-population or small-population counties?

5. Randomized response. Say we want to carry out an in-class poll of all N students in our course, but the yes/no poll question is about a sensitive topic such as a medical condition, past illegal behavior, embarrassing habits, etc. Students might not feel safe answering honestly unless they are sure their answer cannot be traced back to them.

One approach to preserving privacy is "randomized response," in which each student independently chooses one of two different yes/no questions at random. This way, the instructor cannot know which question any given student answered. For instance, let Question A be "Did you brush your teeth this morning?" and let Question B be the opposite: "Did you fail to brush your teeth this morning?"

We are interested in estimating the unknown count θ of the students who would have honestly said "Yes" to Question A (how many students actually brushed their teeth)—but we cannot know who got A vs B.

Method: Choose a probability $p \in [0, 1]$. Tell each student to independently use a random number generator to choose what to answer: Each student should answer either Question A with probability p or Question B with probability $1 - p$. Let X be the total number of "Yes" answers from the N students.

We can treat $X = X_1 + X_2 + \ldots + X_N$ as the sum of independent **but not identical** Bernoulli random variables. Then each X_i comes from one of two distributions—we just won't know which one:

- θ of these random variables are Bernoulli(p), for those who could truthfully say "Yes" to Question A.
- The other $N - \theta$ of them are Bernoulli($1 - p$), for those who could truthfully say "Yes" to Question B.

In short, let $X = V_\theta + W_{N-\theta}$, where $V_\theta \sim \text{Binomial}(\theta, p)$ and $W_{N-\theta} \sim \text{Binomial}(N - \theta, 1 - p)$.

a. Derive the Method of Moments estimator for θ, assuming $p \neq 1/2$. It should only depend on the (random) observed value of X and the (fixed) known values of N and p.
(Hint: We only get to observe X once. So as you set up your MoM, what should you use for $\bar{X}$?)

b. Derive $\text{Var}(\hat{\theta})$. What values of p would lead to the lowest variance? Would it make sense to use them in practice for randomized response? Why or why not?

c. Would it make sense to use a fair coin flip as the random number generator (such as "if heads, answer Question A; if tails, answer Question B")? Why or why not?

d. Can you think of another physical artifact we could use as the random number generator? Describe an example of how you'd use it and what p would be.

6. Power analysis for dice. Imagine trying to make your own six-sided dice out of modeling clay. Consider how you could check whether such a handmade die is fair.

a. Write down the null hypothesis that a six-sided die is fair, in terms of several parameters p_i and specific numbers.
Also, write down a specific alternative hypothesis. Choose some concrete p_i values that make the die "unfair enough" to seem worth detecting.
(This is a matter of taste. I'm not looking for any particular answer—just choose something specific and feasible.)

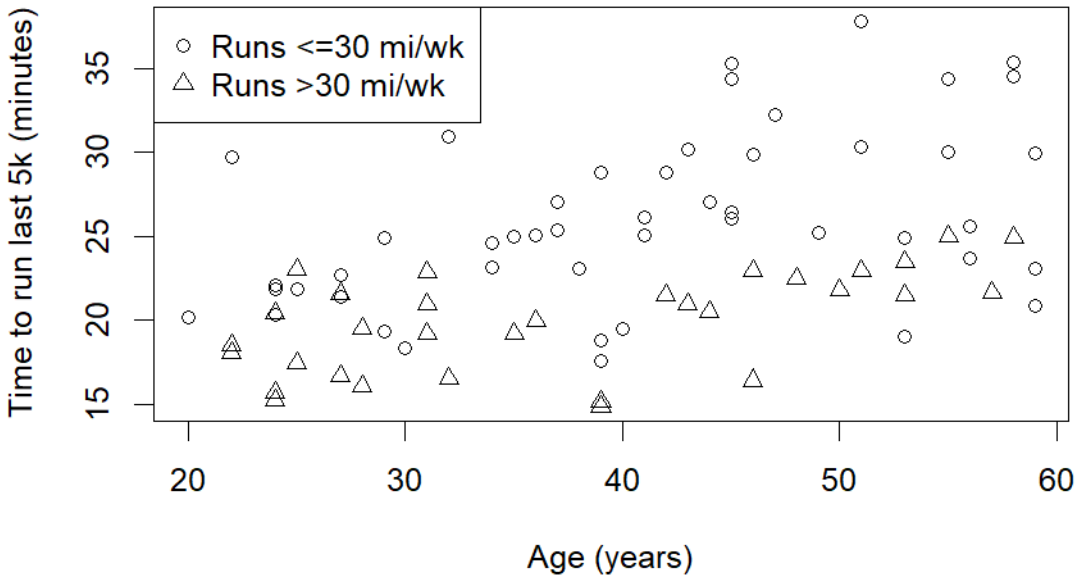
b. Next, imagine that you plan to record the results from n independent rolls of your handmade die. What test statistic will you use to test the hypotheses in (a)? Describe how you would choose a critical value, assuming the usual significance level $\alpha = 0.05$. Draw a sketch of any relevant distributions if it helps.

c. Now say we want 80% power to detect the degree of unfairness in your alternative hypothesis from (a). Describe how you could use computer simulations to choose a sample size n to achieve 80% power. You don't need to write any code. Just give a detailed but brief description of what you would do. Draw a sketch of any relevant distributions if it helps.

7. Regression with 5k race times. *The next two pages provide some problem context and regression output. All questions for Problem 7 are on the final page.*

When running a 5k race, is the Time (in minutes) taken to complete the race related to the runner's Age? Also, does this potential relationship differ by the runner's training habits? We asked each runner whether they typically run 30 miles/week or less, or whether they run over 30 miles/week.

A scatterplot of race time by age appears below, along with some regression output. Diagnostic plots appear on the following page. Assume the data were obtained via a simple random sample of runners and that all observations are independent.

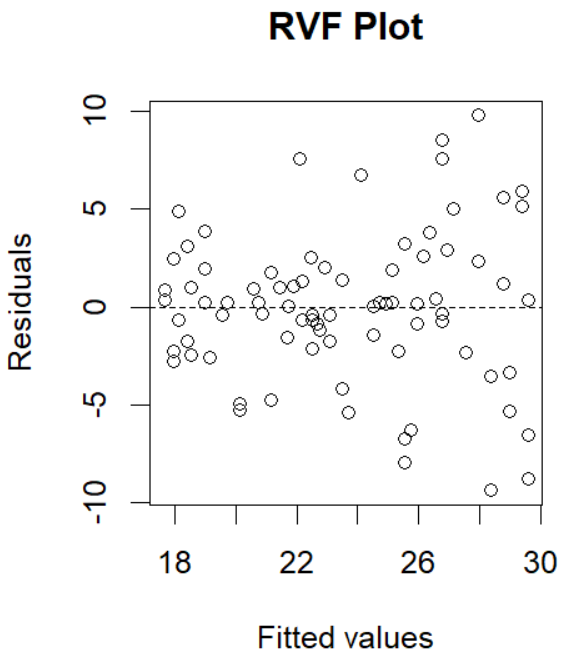


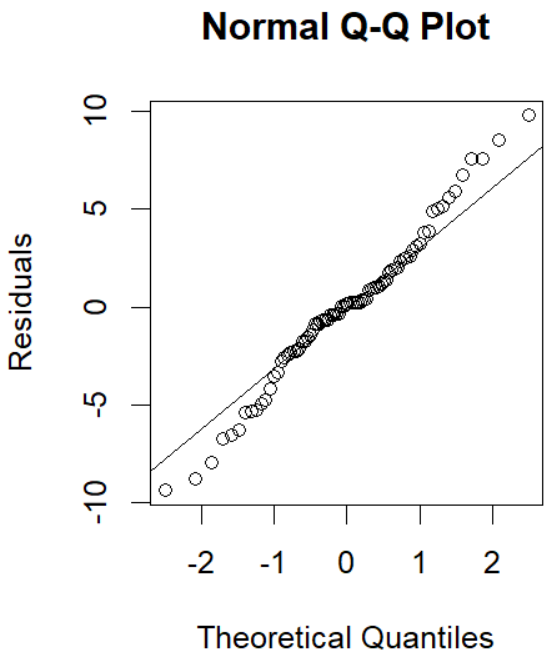

```
Call:
lm(formula = Time ~ Age + runsOver30MiWk + Age:runsOver30MiWk)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3769 -2.1663  0.1511  1.9765  9.7790

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       17.61974    2.14011   8.233 3.94e-12 ***
Age                0.20297    0.05012   4.050 0.000122 ***
runsOver30MiWk    -3.13437    3.18996  -0.983 0.328935
Age:runsOver30MiWk -0.05778    0.07858  -0.735 0.464393
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.915 on 76 degrees of freedom
Multiple R-squared:  0.4733,    Adjusted R-squared:  0.4525
F-statistic: 22.76 on 3 and 76 DF,  p-value: 1.29e-10
```





Next, a simple linear regression was re-fit to this dataset, using only Age to predict Time. The regression output appears below.

```
Call:
lm(formula = Time ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5386 -3.3072 -0.2441  2.3858 11.7402

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.95477    1.88374   7.939 1.24e-11 ***
Age          0.21677    0.04562   4.752 9.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.689 on 78 degrees of freedom
Multiple R-squared:  0.2245,    Adjusted R-squared:  0.2145
F-statistic: 22.58 on 1 and 78 DF,  p-value: 9.03e-06
```

For parts (a)-(e), only consider the **first** model—the multiple linear regression.

a. Using this MLR model, estimate and compare the predicted 5k race times for two 40-year-old runners, if one runs fewer than 30 mi/wk and the other runs more than 30 mi/wk.

b. Briefly interpret all the coefficients associated with `Age`.

c. Which of these would be your first step in simplifying this MLR model? Why?

- remove the term associated with `runsOver30MiWk`
- remove the term associated with `Age:runsOver30MiWk`
- remove **both** `runsOver30MiWk` **and** `Age:runsOver30MiWk`
- remove neither term—the model needs no further simplification

Next, consider conducting a hypothesis test to determine if a parallel-lines model is a good enough model choice for these data, or if a different-slopes model is significantly better.

d. Do you think the assumptions underlying the usual test are met here?
Why or why not?

e. Regardless of your answer to (d), assume for a moment that the assumptions are met. Now use the regression output to conduct the hypothesis test. Be sure to state the parameter of interest, state appropriate null and alternative hypotheses, report a test statistic and a p-value, and state your conclusion in the context of the study.

Now, **also** consider the simple linear regression model.

f. If you assume for a moment that all conditions for both models are met, which model do you prefer: the simple linear regression, or the multiple linear regression from parts (a)-(e)? Why?

Finally, it turns out that we have more detailed info on each runner's training habits. We know whether they typically run 10, 20, 30, 40, or 50 mi/wk.

g. Write down a new parallel-lines model that could be used to predict Time, using Age and indicators for training mileage. Be sure to define your indicator variables clearly.
Write your answer as one equation including the words "Time," "Age," your choice of indicator variables, and some β coefficients.

h. Sketch out (roughly) what the form of the model in (g) would look like geometrically. Assume that runners with higher training mileage tend to have shorter race times at every age.
Start by drawing and labelling a set of axes. Your sketch should be a single figure with 5 labeled lines.