NAME (Please print clearly):

Cheating on exams, whether for your own benefit or for the benefit of other students, is in violation of the Duke Honor Code and will be referred to the Duke Undergraduate Judicial Board. Please acknowledge this by signing the honor code pledge:

I have neither given nor received aid on this examination.

Signature:

- There are 4 questions, but note that credit for the exam is NOT equally distributed across questions.
- Write solutions to each question separately on a new page. Hand in completed solutions AND question sheets.
- As earlier advised, you WILL need a calculator.
- You have (up to) 3 hours. Many of you will not need 3 hours – please leave quietly without disturbing others.
- Use notes, homeworks, slides from class as needed. I do NOT want derivations of results that have been covered in notes, homeworks, class slides, etc. during the course. Simply state the needed results and use them – don’t derive them.
Exercise (1)
An analyst in a commercial inventory department is responsible for monthly updates to demand forecasts for company products. On one product line, his forecast from last month of this month’s demand level $\theta$ was $\theta \sim N(\mu_0, \sigma_0^2)$ where $\mu_0 = 115$ and $\sigma_0 = 10$ (the units are in $\$000$s).

He receives recorded sales of the product that provide information about actual demand levels; the sales data are imperfect, however, due to delays in reporting, demand that was unmet, and so forth. Past experience has led to the sampling model $(x|\theta) \sim N(\theta, 25)$ where $x$ is an adjusted sales figure.

(a) When $x$ is recorded, what is the analyst’s posterior $p(\theta|x)$ for the underlying demand? What is this posterior when he sees the specific sales record $x = 129$?

(b) Find the 99% posterior central interval for demand $\theta$ based on $p(\theta|x)$ given $x = 129$.

(c) The manager of the product line says he was almost sure that the demand level would exceed 135 in this month. Does this seem reasonable under the analyst’s posterior distribution?

\textbf{Useful facts:}

- The upper and lower 2.5% quantiles of the standard normal distribution are $\pm 1.96$
- The upper and lower 0.5% quantiles of the standard normal distribution are $\pm 2.58$
Exercise (2)
A motor company measures battery voltages to assess the average voltage level of new batches of batteries. They are interested in whether or not the batch average is within specified tolerance limits quoted by the battery manufacturer. In a large batch of new batteries, write $\beta_0$ for the average voltage. Individual batteries are randomly sampled from the batch and their measured voltages $y_i$ are assumed normally distributed with mean $\beta_0$ and unknown variance $\sigma^2$.

Using the results of Exercise (3) Part (A) of Homeworks 8-9, answer the following:

(a) What is the reference posterior for $(\beta_0|Y)$ where $Y = \{y_1, \ldots, y_n\}$? The data from a small sample of size $n = 10$ give summary statistics $\bar{y} = 11.65$ and $s = 0.20$. Give a 95% central posterior interval for the average voltage $\beta_0$.

(b) Let $y_{n+1}$ be the voltage of a future battery from the batch. What is the predictive distribution $p(y_{n+1}|Y)$ for the voltage of such a future battery? Give a 95% interval for $y_{n+1}$.

(c) The battery supplier claims that: (A)† the probability that a battery voltage exceeds 12.25 is less than 0.01; and (B)† the probability that a battery voltage is lower than 11.30 is also less than 0.01. Use the predictive distribution $p(y_{n+1}|Y)$ to assess the validity of these claims.

(d) One analyst in the company suggests that future battery voltages should be predicted using the distribution $(y_{n+1}|Y) \approx N(\bar{y}, s^2)$. Explain why and how this underestimates realistic uncertainty about future voltage levels.

†Useful facts:

- The upper and lower 2.5% quantiles of the Student T distribution with 9 degrees of freedom are ±2.26, respectively.

- The upper and lower 1% quantiles of the Student T distribution with 9 degrees of freedom are ±2.82, respectively.
Exercise (3)

Astronomers record neutrino particle counts in a fixed section of sky where a new supernova is hypothesised as generating an increase in such radiation. The known background rate of neutrinos (per minute) is \( b = 0.73 \) and so the number of particles recorded in one minute is Poisson with mean \( b \). If a new supernova exists, it increases the rate to \( b + \mu \) where \( \mu > 0 \) represents the increase. Hence future measurements generate independent Poisson counts \( (x_i|\mu) \sim \text{Pois}(b + \mu) \) in minutes \( i = 1, \ldots, n \). As a result, the recorded total counts \( y = \sum_{i=1}^{n} x_i \) has the sampling distribution \( (y|\mu) \sim \text{Pois}(n(b + \mu)) \) with density

\[
p(y|\mu) = (n(b + \mu))^y \exp(-n(b + \mu))/y!
\]

over \( y = 0, 1, \ldots \).

The existence of the new source is questionable, and this leads to the following hypotheses:

\[
H_0: \mu = 0 \quad \text{versus} \quad H_1: \mu > 0.
\]

Assuming a prior probability \( \pi_0 = \Pr(H_0) \) and a prior density \( p(\mu|H_1) \) for \( \mu \) under \( H_1 \), the posterior probability of \( H_0 \) may be computed using the odds form

\[
\frac{(1 - \pi_0^*)}{\pi_0^*} = \frac{(1 - \pi_0)}{\pi_0} B_{1,0}
\]

where \( B_{1,0} \) is the (generalised) Bayes' factor comparing \( H_1 \) to \( H_0 \).

We now see the data: \( n = 3600 \) minutes of recordings generate a total of \( y = 2725 \) recorded counts.

(a) One astronomer has a prior \( p(\mu|H_1) \) that leads to \( B_{1,0} = 3.53 \) based on this data. What is the implied posterior probability \( \pi_0^* \) of no supernova, assuming that \( \pi_0 = 0.5 \)?

(b) Another astronomer has a prior \( p(\mu|H_1) \) that leads to \( B_{1,0} = 5.8 \) based on this data. How small must his prior probability \( \pi_0 \) be in order that \( \pi_0^* = 0.03 \)?

We know that whatever prior is assigned for \( \mu \) under \( H_1 \), the Bayes' factor always satisfies

\[
B_{1,0} \leq B
\]

where

\[
B = \frac{p(y|\hat{\mu})}{p(y|\mu = 0)}
\]

and where \( \hat{\mu} \) is the MLE. It can be shown that \( \hat{\mu} = \max\{0, y/n - b\} \).

(c) Show that, based on this data, \( \hat{\mu} = 0.0239 \) and therefore \( B = 5.86 \).

(d) Assume \( \pi_0 = 0.5 \). Show that, since \( B_{1,0} \leq 5.86 \), then \( \pi_0^* \geq 0.146 \).

(e) Interpret the result in (d) in connection with the strength of evidence this data provides in favour of a supernova.

(f) A classical statistician remarks that, under \( H_0 \), the probability of a \( y \) value at least as large as 2725 is 0.029. From a significance testing perspective, this small \( p \)-value indicates that the null hypothesis (of no supernova) can be rejected at the 3\% level. His astronomer friends then publicise the view that “the data indicate a probability of only 0.03 that there is no supernova”. What is wrong with this statement? Compare this significance testing result with the Bayesian results above.
Exercise (4)
The Sex Bias in Salaries? data set is summarised on the accompanying sheets. This data arises in a study of potential differentials in annual professional salaries, taking into account other factors. Here there are \( n = 44 \) observations classified by three factors: Sex, a two-level factor (Sex=0 for Men, Sex=1 for Women); Man, a management level factor with two levels indicating the job level of the person (0=Lower, 1=Upper management); and Ed, a factor at three levels indicating the level of education of the person (0=High School, 1=Bac. Degree, 2=Graduate Degree). The other variables are Salary, the annual salary, and Years, the years in the job. Three separate graphs of Salary against Years indicate the salaries classified according to each of the three factors (points are labelled by the level of the factor in each of the three graphs).

The accompanying sheets give summaries of parameter estimates from the Coefficients table output by S-Plus using summary for four different linear regression models to the logs of salaries. The sheet summarises four model fits, labelled fit.1 to fit.4.

Notes: remember that adding a factor variable to a model simply changes the intercept by a parameter for each level of the factor relative to level 0; as in our Mercedes prices data analyses, for example. In each of these four models, the slope parameter on predictor Years does not vary with the factors. For example, in fit.2 the model is estimated as

\[
9.5171 + 0.0334\text{years for management level 0}, \quad \text{and} \quad 9.5171 + 0.2142 + 0.0334\text{years} = 9.7313 + 0.0334\text{years for management level 1}. \quad \text{Similarly, in fit.3 the fitted model is } 9.4769 + 0.0339\text{years for educational level 0, but } 9.4769 + 0.0681 + 0.0339\text{years} = 9.545 + 0.0339\text{years for educational level 2}; \quad \text{and so forth.}
\]

(a) Does the Sex graph suggest likely salary differences between Men and Women? Briefly discuss why this graph alone may give a misleading indication of possible differentials.

(b) Which of fit.1 and fit.2 give a better predictive fit to the data, and why?

(c) Why does fit.3 have 40 degrees of freedom, whereas fit.1 and fit.2 have 41?

(d) Does fit.4 appear to be a useful model in explaining variations in salaries? Why or why not?

(e) What does fit.4 imply about differences in salary levels between men and women, taking into account the other factors? Do your conclusions seem reasonable in connection with the graph?

(f) Remembering that the response is salary on the log scale, what does the estimated coefficient of 0.092 for Ed1 imply for differences in salaries between people with Bac. degrees relative to those with just high school education?

(g) What do you make of the fact that the estimated Ed2 coefficient is significantly smaller than that for Ed1? Do the graphs give you any indication of why this perhaps surprising result arises?