Statistics I — Solution for Midterm Exam

1. A mathematics department has 25 faculty members excluding the department head. Candidates applying for summer internships are evaluated by all 25 members. To understand if there is any gender bias among the faculty, the department head sends the CV of a female candidate to all 25 members for evaluation, but changes the name and gender of the candidate in the CV before sending it to 10 of the 25 faculty members. The following data are recorded for all 25 members.

`gender` Gender of the faculty member

`age` Age of the faculty member

`evaluation` Their evaluation of the CV, which can be “accept”, “reject”, or “no opinion”

`version` Which version of the CV they were sent

(a) Identify the observational units in the study. [1]

(b) Identify the categorical variables in the study. [1]

(c) Identify the numeric variables in the study. [1]

(d) Is this a case control study? Justify your answer. [2]

(e) Is there any scope for randomization in this study? Justify your answer. [2]

Solution:

(a) The observational units in the study are the 25 faculty members.

(b) The three categorical variables in the study are `gender`, `evaluation`, and `version`.

(c) The only numeric variable in the study is `age`.

(d) YES, this is a case control study, because there are two “treatment” groups, indicated by the `version` of the CV sent to each faculty member. There is gender bias among the faculty only if the distribution of the “outcome”, which in this case is the `evaluation`, changes depending on the `version` of the CV.

(e) YES, there is scope for randomization in this study. The study description above does not mention which faculty member receives which version of the CV. If this assignment is randomized, for example by uniformly choosing one of the $\binom{25}{10}$ subsets of 10 faculty members to receive the modified CV, then the study is randomized.

In fact, randomization is critical in an experiment like this to avoid conscious or unconscious bias. For example, if most of the faculty members who were sent the modified CV were female and most of the faculty members who were sent the origianl CV were male (or the other way around), we would not be able to distinguish between the effects of `gender` and `version`. The same holds for other factors such as age, or other unobserved factors that the experimenter may not even be aware of.

2. Let the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ represent n positive valued observations collected in a survey. The geometric mean of $\boldsymbol{X}$ is defined as [5]

$$\mathrm{GM}(\boldsymbol{X}) = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

Prove that the value of θ which minimizes the loss function $\lambda(\boldsymbol{X} \mid \theta) = \sum_{i=1}^{n} \left[\log \frac{X_i}{\theta}\right]^2$ is $\hat{\theta} = \mathrm{GM}(\boldsymbol{X})$.

Solution 1: Differentiating $\lambda(\boldsymbol{X} \mid \theta)$ w.r.t. θ, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\,\lambda(\boldsymbol{X} \mid \theta) &= 2\sum_{i=1}^{n}\left(\log\frac{X_i}{\theta}\right)\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log\frac{X_i}{\theta}\right] \\
&= 2\sum_{i=1}^{n}\left(\log\frac{X_i}{\theta}\right)\left(\frac{\theta}{X_i}\right)\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{X_i}{\theta}\right] \\
&= 2\sum_{i=1}^{n}\left(\log\frac{X_i}{\theta}\right)\left(\frac{\theta}{X_i}\right)\left(-\frac{X_i}{\theta^2}\right) \\
&= -\frac{2}{\theta}\sum_{i=1}^{n}\left(\log\frac{X_i}{\theta}\right) = -\frac{2}{\theta}\sum_{i=1}^{n}(\log X_i - \log\theta)
\end{aligned}$$

Equating the derivative to 0, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\,\lambda(\boldsymbol{X} \mid \hat{\theta}) = 0 \quad &\Longrightarrow \quad -\frac{2}{\hat{\theta}}\sum_{i=1}^{n}\left(\log X_i - \log\hat{\theta}\right) = 0 \\
&\Longrightarrow \quad \sum_{i=1}^{n}\left(\log X_i - \log\hat{\theta}\right) = 0 \\
&\Longrightarrow \quad \sum_{i=1}^{n}\log X_i = \sum_{i=1}^{n}\log\hat{\theta} = n\log\hat{\theta} = \log\hat{\theta}^n \\
&\Longrightarrow \quad \hat{\theta}^n = \exp\left\{\sum_{i=1}^{n}\log X_i\right\} = \prod_{i=1}^{n} X_i \Longrightarrow \hat{\theta} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}
\end{aligned}$$

Solution 2: Note that $\lambda(\boldsymbol{X} \mid \theta) = \sum_{i=1}^{n}(\log X_i - \log\theta)^2$. Define $Y_i = \log X_i$ and $\beta = \log\theta$. Then the problem is equivalent to minimizing $\sum_{i=1}^{n}(Y_i - \beta)^2$ w.r.t., β, which as we know is solved for

$$\begin{aligned}
\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad &\Longrightarrow \quad \log\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\log X_i \\
&\Longrightarrow \quad \hat{\theta}^n = \exp\left\{\sum_{i=1}^{n}\log X_i\right\} = \prod_{i=1}^{n} X_i \Longrightarrow \hat{\theta} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}
\end{aligned}$$

3. Suppose a dataset consists of observations $(5, 4, 11, 6, x)$, where the value of the last observation, denoted by x, is unknown. Let $\mu(x)$ be the mean of the observations as a function of x, and similarly, $\nu(x)$ be the median of the observations as a function of x. Obtain explicit expressions for $\mu(x)$ and $\nu(x)$, and draw their graphs for $x \in (0, 15)$. [5]

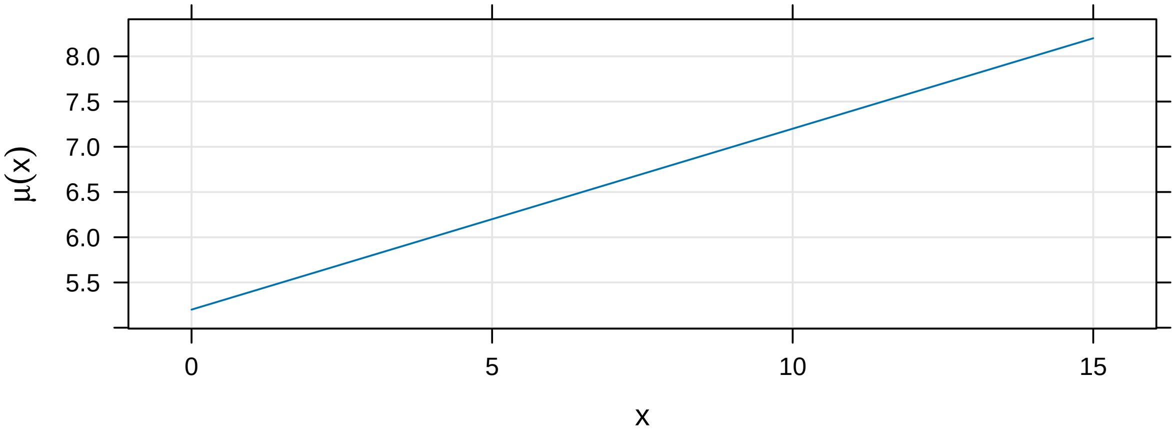
Solution:

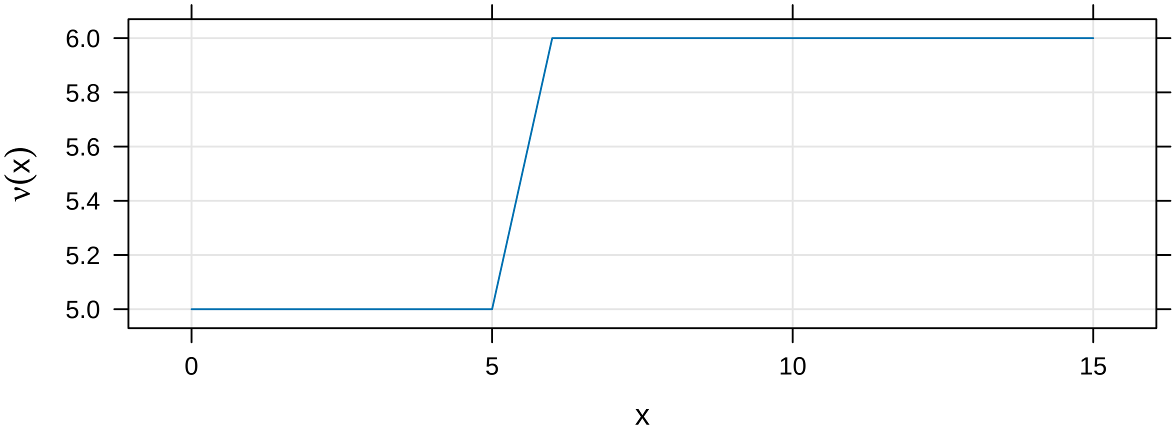
$\mu(x)$ is given by

$$\mu(x) = \frac{5+4+11+6+x}{5} = \frac{26}{5} + \frac{x}{5} \text{ for } x \in \mathbb{R}.$$

$\nu(x)$ is given by

$$\nu(x) = \begin{cases} 5 & x < 5 \\ x & 5 \leq x \leq 6 \\ 6 & x > 5 \end{cases}$$





4. The following table gives binned frequency counts of weight data for a subset of the NHANES data.

bin	midpoint	log2-midpoint	Count	Cumulative Count
(0, 20]	10	3.3	185	185
(20, 40]	30	4.9	187	372
(40, 60]	50	5.6	716	1088
(60, 80]	70	6.1	1253	2341
(80, 100]	90	6.5	561	2902
(100, 120]	110	6.8	218	3120
(120, 140]	130	7.0	52	3172
(140, 160]	150	7.2	15	3187
(160, 180]	170	7.4	4	3191
(180, 200]	190	7.6	1	3192

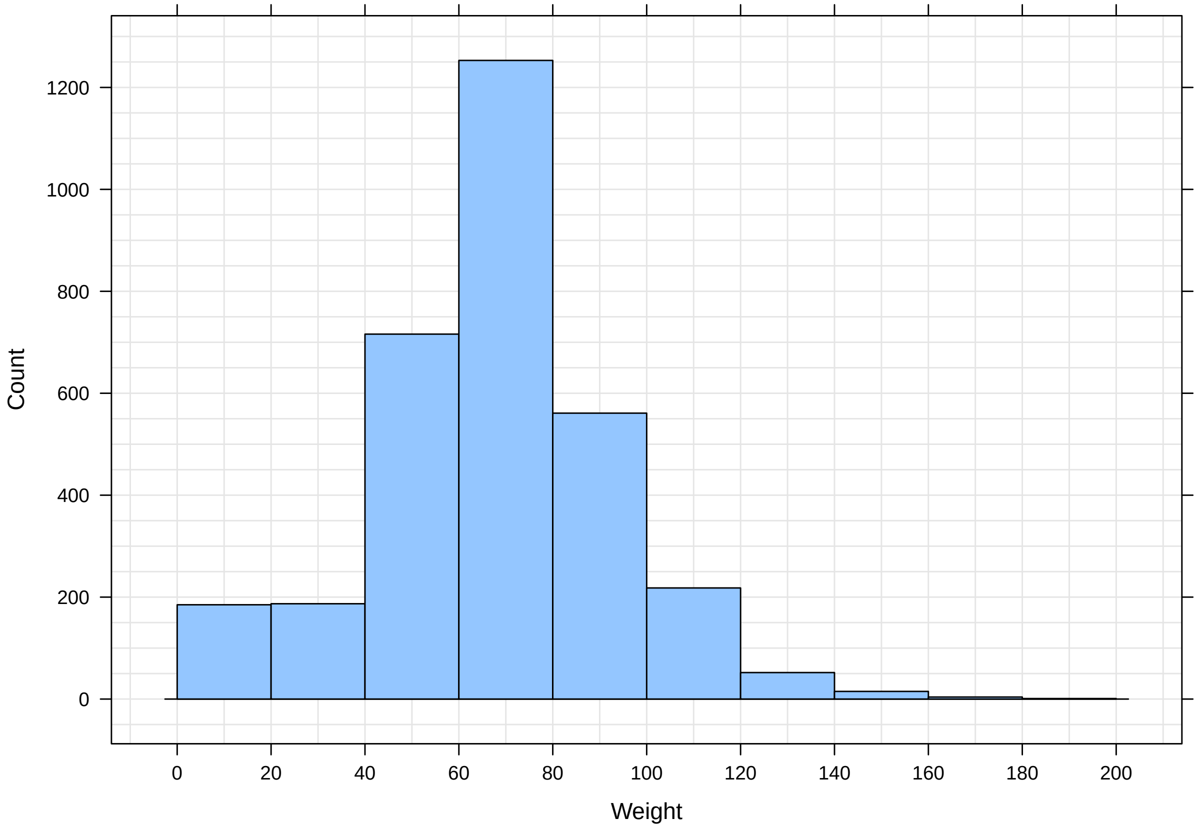
Pretend that all data points in a bin are equal to the midpoint of that bin. Then,

(a) Compute the following summary statistics: [8]
(i) Median, (ii) Inter-Quartile Range, (iii) Arithmetic Mean, and (iv) Geometric Mean.

(b) Using the graphing sheet provided, draw a frequency histogram representing the table above. [5]

Solution:

Median = 70 / Inter-Quartile Range = 40 / Arithmetic Mean = 67.46 / Geometric Mean = 60.38


Count
Weight