Beatriz Hernandez, Jimmy Nguyen, Ryan Quigley
MATH 267A
Section 2
Group D

Problem Set 9

1.

a) The mystery distribution (MD) has thinner left and right tails than the normal distribution (ND), which suggest that the MD has higher quartile values than the ND in regions 1 and 3. Though both the MD and ND peak in region 2, the peak of the MD is higher than the ND, with larger portion of area under the PDF curve in this region compared to ND. This suggests a smaller variance of the MD compared to the ND.

b) In region 1, the points do describe the difference between the left tails of the MD and ND. In this region, the MD has higher quantile values associated with the quantiles of the ND; that is, the MD requires higher quantile values to achieve the same area under the PDF curve as the ND. The points in the QQ plot in region 1 do describe the thinner left tail of the MD depicted in region 1 on the left panel.

c) The points in region 3 of the right panel suggest that the quantiles of the MD are linearly related to the quantiles of the ND, but in this case, we cannot tell whether the right tail of the MD will be thinner or thicker than the ND; only that the quantiles are proportional in a linear fashion. Furthermore, the points in region 2 in the QQ plot do not give us any information that could help us determine if the MD will have thicker or thinner right tail compared to ND. All we can determine is that the MD behaves like a normal distribution in regions 2 and 3.

d) Q-q plots cannot be solely relied on to make accurate inferences about the underlying distribution of sample data. Though q-q plots can be a useful piece in determining whether the unknown probability distribution of the sample data grossly deviates from a known family of distributions, these plots do not allow us to catch key but more subtle differences, or even not so subtle ones as shown in region 2 of the left panel in this exercise.

2.

$$p = 0 \qquad F^-(0) = inf\{x \in \mathbb{R} : F(x) \geq 0\}$$
$$= inf(-\infty, \infty)$$
$$= -\infty$$
$$:= 0$$

$$0 < p < \frac{1}{4} \qquad F^-(p) = F^{-1}(p) \quad \text{by prop. 2}$$
$$p = \frac{x}{4} \Rightarrow x = 4p$$

$$p = \frac{1}{4} \qquad F^-(\frac{1}{4}) = inf\{x \in \mathbb{R} : F(x) \geq \frac{1}{4}\}$$
$$= inf[1, \infty)$$
$$= 1$$

$$\frac{1}{4} < p < \frac{1}{2} \qquad F^-(p) = inf\{x \in \mathbb{R} : F(x) \geq p\}$$

If $x < 1$, $F(x) < \frac{1}{4}$; if $x \geq 1$, $F(x) \geq \frac{1}{2}$

$$\Rightarrow = inf[1, \infty)$$
$$= 1$$

$$p = \frac{1}{2} \qquad F^-(\frac{1}{2}) = inf\{x \in \mathbb{R} : F(x) \geq \frac{1}{2}\}$$
$$= inf[1, \infty)$$
$$= 1$$

$$\frac{1}{2} < p \leq \frac{2}{3} \qquad F^-(p) = inf\{x \in \mathbb{R} : F(x) \geq p\}$$

If $x < 2$, $F(x) \leq \frac{1}{2}$; if $x \geq 2$, $F(x) \geq \frac{2}{3}$

$$\Rightarrow = inf[2, \infty)$$
$$= 2$$

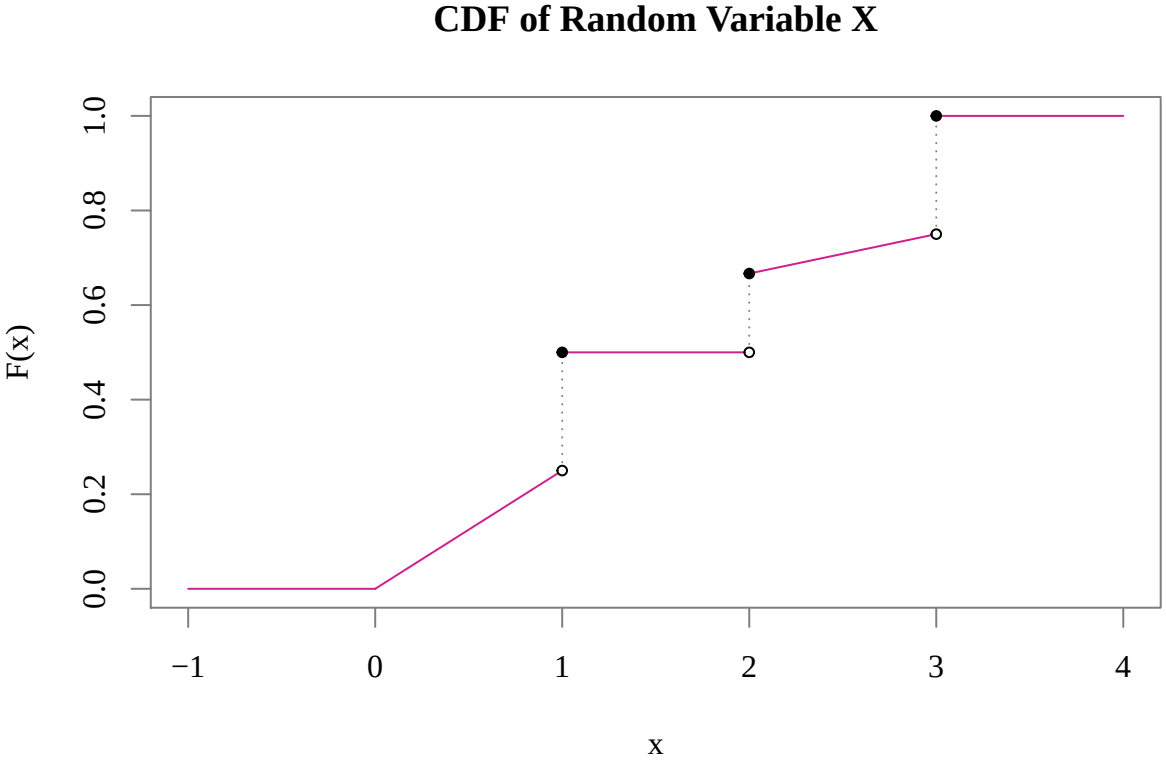
$$\frac{2}{3} < p \leq \frac{3}{4} \qquad F^-(p) = F^{-1}(p) \quad \text{by prop. 2}$$
$$p = \frac{x}{12} - \frac{1}{2} \Rightarrow x = 12(p - \frac{1}{2})$$

$$\frac{3}{4} < p \leq 1 \qquad F^-(p) = inf\{x \in \mathbb{R} : F(x) \geq p\}$$

If $x < 3$, $F(x) < \frac{3}{4}$; if $x \geq 3$, $F(x) = 1$

$$\Rightarrow = inf[3, \infty)$$
$$= 3$$

$$F^{-}(p) = \begin{cases} 4p & 0 \leq p \leq \frac{1}{4} \\ 1 & \frac{1}{4} < p \leq \frac{1}{2} \\ 2 & \frac{1}{2} < p \leq \frac{2}{3} \\ 12\left(p - \frac{1}{2}\right) & \frac{2}{3} < p \leq \frac{3}{4} \\ 3 & \frac{3}{4} < p \leq 1 \end{cases}$$



3. The dataset x has 150 observations which is a reasonably large sample size. Thus, using the kernel density method should give us a good estimate of the uknown distribution of the data. After plotting several estimates using different kernel functions, it is clear that the resulting estimate is not noticeably changed based on the choice of kernel. Therefore, we proceed with the Epanechnikov kernel because it has properties (albeit uknown to us) that make it optimal. After testing several bandwidths in the range [0.01, 0.10], h = 0.05 gives a good idea of the shape and appears to be a reasonable middle group between under and over smoothing. With the finalized kernel density estimate plotted (violet red line in the figure below), the underlying distribution appears to be symmetric, bell-shaped, and centered around 0.50. As a result, the normal distribution is an obvious candidate for the unknown distribution; furthermore, it has the additional desireable property that good parameter estimate can be easily obtained from the data by taking $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2$. With $\bar{x} = 0.52$ and $s^2 = 0.02$, we get the dashed blue line in the plot below. The plot shows that the density of $\mathrm{N}(\bar{x}, s^2)$ follows the kernel density very closely, so we can conclude the model is a good fit.

   Another distribution that can be symmetric and bell-shaped is the beta distribution. The beta distribution is an appropriate candidate for this dataset because the support of the distribution is the interval [0, 1]. It is not completely clear from the kernel density estimate what the range of the dataset is, but a quick summary call in R gives the data range: [0.054, 0.868]. The beta parameters are not as easily estimated from the data, so we proceed with trial and error to get

a close fit. The resulting parameterized beta distribution is plotted in the figure below using a block dotted line. This candidate model is also a very good fit for the unknown distribution.

FINAL CHOICE??