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

Problem Set 9

1.

a) The mystery distribution (MD) is thinner than the normal distribution (ND) in both the tails, which results in larger area under the density than the ND in the middle. However, the relative thinness is not the same for both tails; the MD left-tail is much thinner on the left than the right when compared to the corresponding ND tail, which suggests the MD is slightly skewed right. Another indicator of the skewness is that on the left the increase in the MD density is much sharper than the ND density, but the subsequent decrease on the right is closer to the ND in that it is more gradual. The rapid increase in the MD density results in the peak occuring slightly earlier than the peak of the ND.

b) In region 1, the points do describe the difference between the left tails of the MD and ND. The points do not fall on the dotted straight line indicating that the quantiles of the MD do not have a linear relationship with the quantiles of the ND. More specifically, the MD requires larger quantile values to achieve the same area under the PDF curve as the ND as graphically depicted by the thinner left tail of the MD 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. All we can conclude is that in this region the rate at which the MD is accumulating area is proportional to the rate that the ND is accumulating area.

d) We have previously shown mathematically that Q-Q Plots can be reliably used to determine if an unknown distribution is identical to a particular parameterization of a theoretical distribution by checking whether or not the points fall on the 45 degree line. This is not all that useful because we rarely know the exact parameters of the theoretical distribution that we want to compare the unknown distribution to. We have also seen that for particular dsitribution families, such as normal and exponential, any linearity (not just the 45 degree line) in the Q-Q plot suggests the unknown distribution belongs to the family of distributions that contains the theoretical distribution being compared.

These cases above are special. In general, the shape of the Q-Q plot should not be used to reconstruct an unknown distribution's density. More specifically, one should not plot the density of the theoretical distribution and then attempt to estimate the uknown density by extraploting the differences from the reference line observed in the Q-Q plot. Parts (a) through (c) show that using this approach on the Q-Q plot in the right panel would not reproduce the density of the unknown distribution plotted in the left panel.

2.

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

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

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

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

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

$$\begin{aligned} \Rightarrow &= inf[1, \infty) \\ &= 1 \end{aligned}$$

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

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

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

$$\begin{aligned} \Rightarrow &= inf[2, \infty) \\ &= 2 \end{aligned}$$

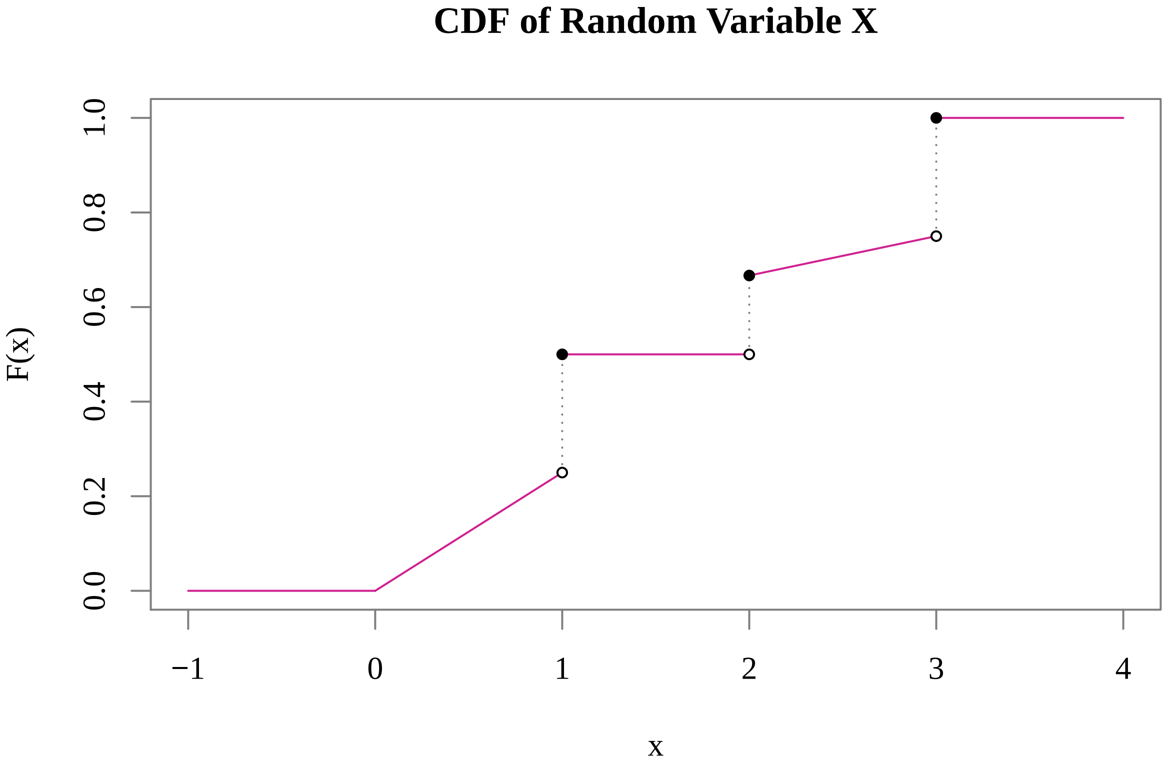
$$\frac{2}{3} < p \leq \frac{3}{4} \qquad \begin{aligned} F^{-}(p) &= F^{-1}(p) \quad \text{by prop. 2} \\ p &= \frac{x}{12} - \frac{1}{2} \Rightarrow x = 12(p - \frac{1}{2}) \end{aligned}$$

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

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

$$\begin{aligned} \Rightarrow &= inf[3, \infty) \\ &= 3 \end{aligned}$$

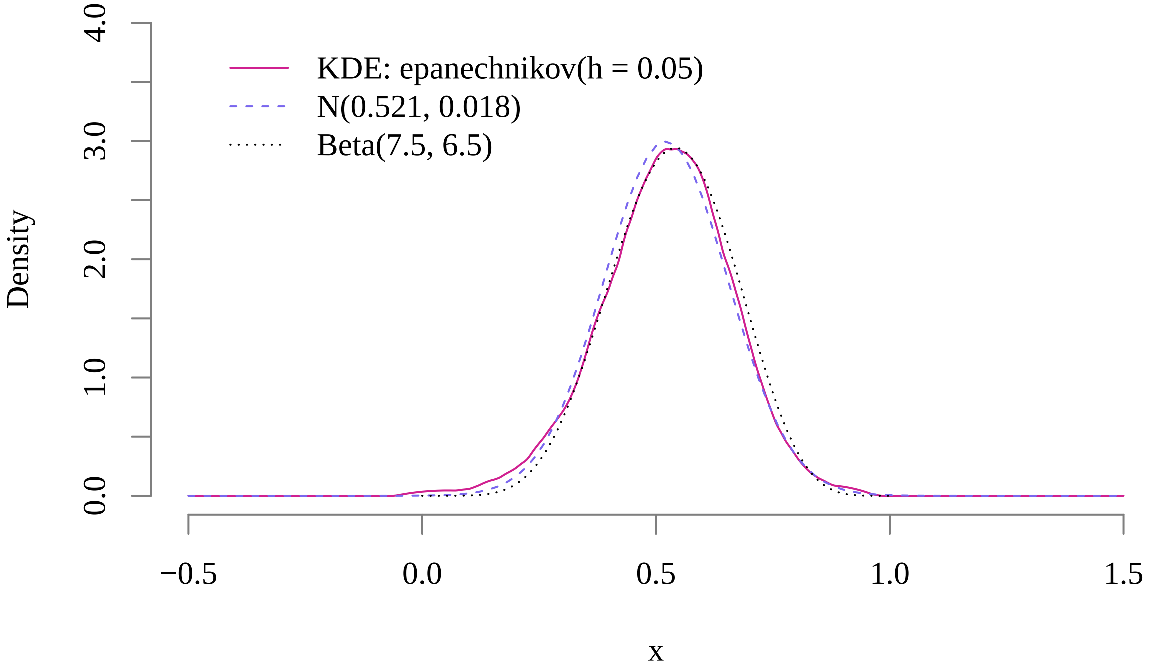
$$F^{-}(p) = \begin{cases} 4p & 0 \le p \le \frac{1}{4} \\ 1 & \frac{1}{4} < p \le \frac{1}{2} \\ 2 & \frac{1}{2} < p \le \frac{2}{3} \\ 12\left(p - \frac{1}{2}\right) & \frac{2}{3} < p \le \frac{3}{4} \\ 3 & \frac{3}{4} < p \le 1 \end{cases}$$



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 desirable 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 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 also 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 black dotted line. Based on the closeness of the theoretical distribution to the kernel density estimate, this candidate model also apperas to be a very good fit for the unknown distribution.

Because we have no information about the underlying population that this sample is drawn from we select the normal distribution, $\mathrm{N}(\bar{x} = 0.52, s^2 = 0.018)$, to allow for the possibility that the underlying data could be negative. We do not want to rule out this possibility by choosing a distribution with a restricted support.


KDE: epanechnikov(h = 0.05)
N(0.521, 0.018)
Beta(7.5, 6.5)
Density
x