# STATISTICS 641 - ASSIGNMENT 3

**DUE DATE: NOON (CDT), MONDAY, SEPTEMBER 27, 2021**

Name ______________________________________________________

Email Address _______________________________________________

**Please TYPE your name and email address. Often we have difficulty in reading the handwritten names and email addresses. Make this cover sheet the first page of your Solutions.**

## STATISTICS 641 - ASSIGNMENT #3 - Due NOON (CDT) Monday - 9/27/2021

- Read: Handouts 4 & 5
- Supplemental Reading in Devore book: Chapters 1 & 4

- Submit for grading the following problems:

P1. ( 10 Points) Let Y have a double exponential distribution, that is, Y has pdf and cdf in the following form with parameters $\theta,\ \beta > 0$:

$$f(y) = \frac{1}{2\beta} e^{-\left(\left|\frac{y-\theta}{\beta}\right|\right)} \text{ for } \quad -\infty \le y \le \infty \qquad F(y) = \begin{cases} \frac{1}{2} e^{-\left(\frac{\theta-y}{\beta}\right)} & \text{for} \quad y < \theta \\ 1 - \frac{1}{2} e^{-\left(\frac{y-\theta}{\beta}\right)} & \text{for} \quad y \ge \theta \end{cases}$$

( a.) Derive the quantile function for $Y$

( b.) Derive the survival function for $Y$

( c.) Derive the hazard function for $Y$

P2. ( 10 Points) A researcher is studying the relative brain weights 1000 times the ratio of brain weight to body weight for 51 species of mammal whose average litter size is less than 2 and for 44 species of mammal whose average litter size is greater than or equal to 2. The researcher was interested in determining what evidence that brain sizes tend to be different for the two groups. (Data from *The Statistical Sleuth* by Fred Ramsey and Daniel Schafer). The data is in the Homework Assignment Folder: Assign3-BrainSize

BRAINSIZE - SMALL LITTER SIZE

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.42 | 0.86 | 0.88 | 1.11 | 1.34 | 1.38 | 1.42 | 1.47 | 1.63 |
| 1.73 | 2.17 | 2.42 | 2.48 | 2.74 | 2.74 | 2.79 | 2.90 | 3.12 |
| 3.18 | 3.27 | 3.30 | 3.61 | 3.63 | 4.13 | 4.40 | 5.00 | 5.20 |
| 5.59 | 7.04 | 7.15 | 7.25 | 7.75 | 8.00 | 8.84 | 9.30 | 9.68 |
| 10.32 | 10.41 | 10.48 | 11.29 | 12.30 | 12.53 | 12.69 | 14.14 | 14.15 |
| 14.27 | 14.56 | 15.84 | 18.55 | 19.73 | 20.00 | | | |

BRAINSIZE - LARGE LITTER SIZE

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.94 | 1.26 | 1.44 | 1.49 | 1.63 | 1.80 | 2.00 | 2.00 | 2.56 |
| 2.58 | 3.24 | 3.39 | 3.53 | 3.77 | 4.36 | 4.41 | 4.60 | 4.67 |
| 5.39 | 6.25 | 7.02 | 7.89 | 7.97 | 8.00 | 8.28 | 8.83 | 8.91 |
| 8.96 | 9.92 | 11.36 | 12.15 | 14.40 | 16.00 | 18.61 | 18.75 | 19.05 |
| 21.00 | 21.41 | 23.27 | 24.71 | 25.00 | 28.75 | 30.23 | 35.45 | |

A software package uses the estimator $\widehat{Q}(u) = Y_{((n-1)u+1)}$ as the estimator of Q(u).

- Calculate the estimates of the Quartiles: of $Q(.25),\ Q(.5),\ Q(.75)$ for just the **Large Litter Size** using the software package's formula.

P3. ( 24 points) Using the data from Problem 2 for just the **Large Litter Size**, we want to estimate the pdf $f(y)$ for the relative brain weights of the 44 species of mammal.

The kernel density estimate of $f(y)$ is given by

$$\widehat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-Y_i}{h}\right),$$

Suppose we use the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ and a bandwidth of $h = 3$.

( a.) Estimate $f(3)$ and $f(16)$ using the kernel density estimator.

( b.) Using a relative frequency histogram with bin width of 5, estimate the values of $f(3)$ and $f(16)$.

( c.) Which data value provides the smallest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

( d.) Which data value provides the largest contribution to the kernel density estimator at y=16, $\widehat{f}(16)$?

P4. ( 28 points ) Using the relative Brain Weight data, answer the following questions:

( a.) Produce the following plots of the data: estimates of the pdf, cdf, and quantile function for both Small and Large litter sizes.

( b.) Describe the underlying distribution of the relative brain weights for both Small and Large litter sizes.

( c.) Based on the graphs, what are your conclusions about the relationship between litter size and relative brain weights?

P5. ( 28 Points) **Select** the letter of the **best** answer for each question. No explanation is needed for your selection.

1. The function which provides the most detailed description for the realizations of a random variable is

   A. the probability density (mass) function, pdf $f(\cdot)$

   B. the quantile function, $Q(\cdot)$

   C. the survival function, $S(\cdot)$

   D. the cumulative distribution function, cdf $F(\cdot)$

   E. all the above functions are equivalent

2. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. This plot will not be an appropriate estimator of the population pdf because

   A. all the intervals are not the same width.

   B. the relative frequency varies greatly by class width.

   C. the area under the curve is not proportional to one.

   D. the area under the curve for each class is not an estimator of the probability of that class.

   E. In fact it is an unbiased estimator of the pdf.

3. A relative frequency histogram having classes of greatly different class widths was used as an estimator of a continuous population pdf. The relative frequency was plotted versus the class intervals. The plot will result in a graphical distortion. The plot can be corrected by

   A. making all the intervals have the same width.

   B. plotting the relative frequency divided by class width.

   C. making sure that the area under the curve adds to one

   D. increasing the sample size.

   E. In fact there will not be a distortion since it is an unbiased estimator of the pdf.

4. A kernel density estimator was used as an estimator of a continuous population pdf, $f(y)$. The kernel density estimator is generally a vastly improved estimator over a density histogram (plot of
$\frac{N_i/n}{h_i}$ vs Class $i$) because

   A. in using the histogram, it is necessary to select the number of bins, bin widths, and their location.

   B. there are too many spurious modes using the histogram

   C. the area under the curve adds to 1 for the kernel density estimator.

   D. the kernel density estimator makes use of all the data in estimating $f(y)$ whereas the histogram only uses those data values in the same bin as $y$.

   E. all of the above

5. In using a kernel density estimator to estimate a population pdf based on a random sample $Y_1, \cdots, Y_n$, the design factor which is **least** crucial in determining the effectiveness of the estimator is

   A. the sample size, $n$

   B. the number of plotting points, $m$ provided $m > 50$

   C. the bandwidth, $h$

   D. the kernel $k(\cdot)$

   E. all four factors are equally crucial

6. A kernel density estimator is an estimator of a population pdf, $f(y)$. The bandwidth of the kernel density estimator is selected by

   A. using the uniform distribution to randomly select a value between 0 and 1.

   B. taking the value which minimizes the asymptotic integrated mean square error.

   C. taking the value which produces a curve having area closest to 1.

   D. taking the value of the bandwidth which yields maximum entropy.

   E. asking Dr. Sheather.

7. A random sample of n data values is obtained from a process having an absolutely continuous cdf of unknown shape. The metallurgist wants to select the best fitting distribution amongst several candidate cdfs. She decides to select the distribution which has mean and variance most closely matching the corresponding sample mean and variance. The major weakness in this approach is

   A. the mean and variance may be highly inflated by outliers

   B. the empirical distribution function contains more information about the tails of the distribution than does the mean and variance

   C. she should have used robust estimators of the location and scale parameters

   D. there are many distribution having the same mean and variance but very different shapes

   E. the moments of a distribution determine the distribution, hence there is no weakness in the approach

8. The skewness and kurtosis parameters are generally thought to represent the following characteristics of the population cdf, respectively,

   A. the center and spread in the distribution
   B. the heaviness of the tails and deviation from normality of the distribution
   C. the deviation from symmetry and concentration in the tails of the distribution
   D. the deviation from symmetry and concentration in the tails and/or the peakedness of the distribution
   E. none of the above

9. The median is a trimmed mean with level of trimming equal to

   A. 0%
   B. 25%
   C. 50%
   D. 75%
   E. none of the above

10. The standard deviation is preferred to MAD as a measure of population dispersion when the population distribution

    A. has absolutely no outliers.
    B. has a skewed but short-tailed distribution.
    C. has a lognormal distribution.
    D. has a normal distribution.
    E. cannot be determined with the given information.

11. The coefficient of kurtosis

    A. measures the dispersion of the distribution about two values $\mu \pm \sigma$.
    B. measures the peakedness of a distribution.
    C. measures the concentration of the mass of a distribution in the tails of the distribution.
    D. measures the difference between a distribution and the normal distribution.
    E. all of the above

12. Alternatives to $\sigma$ for measuring the dispersion in a distribution are $SIQR$ and $MAD$. Which of the following statements about these measures are **TRUE**?

    A. All three measures are equal if the pdf for the distribution is symmetric.
    B. $SIQR$ is preferred to $MAD$ if the distribution has very heavy tails
    C. For the normal distribution, $SIQR$ is preferred to $MAD$
    D. all of the above
    E. none of the above

13. A government study of the average monthly nitrate levels in the Mississippi river, $N_t$, just prior to its entry into the Gulf of Mexico is modeled as

    $N_t = 22.3 + .6N_{t-1} + e_t$ where $e_t's$ are iid r.v.s, $E[e_t] = 0,\ Var[e_t] = 2.8,\ e_t's$ are independent of $N_t's$

    The mean and variance of $N_t$ are given by

    A. $\mu = 22.3,\ \sigma^2 = 2.8$
    B. $\mu = 55.75,\ \sigma^2 = 4.375$
    C. $\mu = 34.84,\ \sigma^2 = 2.8$
    D. $\mu = 22.3,\ \sigma^2 = 4.375$
    E. The values of $\mu$ and $\sigma^2$ would change from month to month.

# Stat 641 Fall 2021

## Solutions for Assignment 3

P1. ( 10 points) Let $Y$ have a double exponentiall distribution.

( a.) The quantile function

$$Q(u) = \begin{cases} \theta + \beta \; log(2u) & \text{for} \quad u \le .5 \\ \\ \theta - \beta \; log(2(1-u)) & \text{for} \quad u \ge .5 \end{cases}$$

( b.) The survival function is given by

$$S(y) = P(Y > y) = 1 - F(y) \;\; \Rightarrow \qquad S(y) = \begin{cases} 1 - \frac{1}{2} e^{-\left(\frac{\theta - y}{\beta}\right)} & \text{for} \quad y < \theta \\ \\ \frac{1}{2} e^{-\left(\frac{y-\theta}{\beta}\right)} & \text{for} \quad y \ge \theta \end{cases}$$

( c.) The hazard function is given by

$$h(y) = \frac{f(y)}{S(y)} \;\; \Rightarrow \qquad h(y) = \begin{cases} \frac{\frac{1}{\beta} e^{\left(\frac{y-\theta}{\beta}\right)}}{2 - e^{\left(\frac{y-\theta}{\beta}\right)}} & \text{for} \quad y \le \theta \\ \\ \frac{1}{\beta} & \text{for} \quad y > \theta \end{cases}$$

P2. ( 10 points) $n = 44 \;\; \Rightarrow \;\; \widehat{Q}(u) = Y_{(43u+1)} \;\; \Rightarrow$

- $\widehat{Q}(.25) = Y_{(11.75)} = .25Y_{(11)} + .75Y_{(12)} = .25(3.24) + .75(3.39) = 3.35$
- $\widehat{Q}(.5) = Y_{(22.5)} = .5Y_{(22)} + .5Y_{(23)} = .5(7.89) + .5(7.97) = 7.93$
- $\widehat{Q}(.75) = Y_{(33.25)} = .75Y_{(33)} + .25Y_{(34)} = .75(16.00) + .25(18.61) = 16.65$

P3. (24 points) Using the R code:

```
y = c(0.94,    1.26,    1.44,    1.49,    1.63,    1.80,    2.00,    2.00,    2.56,
    2.58,    3.24,    3.39,    3.53,    3.77,    4.36,    4.41,    4.60,    4.67,
    5.39,    6.25,    7.02,    7.89,    7.97,    8.00,    8.28,    8.83,    8.91,
    8.96,    9.92,   11.36,   12.15 ,  14.40,   16.00,   18.61,   18.75,   19.05,
   21.00,   21.41,   23.27,   24.71,   25.00,   28.75,   30.23,   35.45   )
h=3
n=length(y)
deni <- function(x){
  (1/sqrt(2*pi))*exp(-((x-y)/h)^2/2)/(n*h)
}
f3 = sum(sapply(3,deni))
f16 = sum(sapply(16,deni))
f16i = sapply(16,deni)
min = min(f16i)
imin = which(f16i==min)
ymin=y[imin]
max = max(f16i)
imax=which(f16i==max)
ymax=y[imax]
```

(a.) The value for $\widehat{f}(3)$ is f3 = 0.059703 and for $\widehat{f}(16)$ is f16 = 0.01669353

(b.) Using a relative frequency histogram with a bin width of 5, with

$n_j = \#Y_i$'s in $[0.94 + 5(j-1), 0.94 + 5j)$, we have $n_1 = 19, \; n_2 = 10, \; n_3 = 3, \; n_4 = 4, \; n_5 = 5, \; n_6 = 2, \; n_7 = 1$.

Therefore, the estimates are $\widehat{f}(3) = 19/(44 \times 5) = 0.08636364$ and for $\widehat{f}(16) = 4/(44 \times 5) = 0.01818182$. A fairly close agreement between the estimates obtained by the two methods.

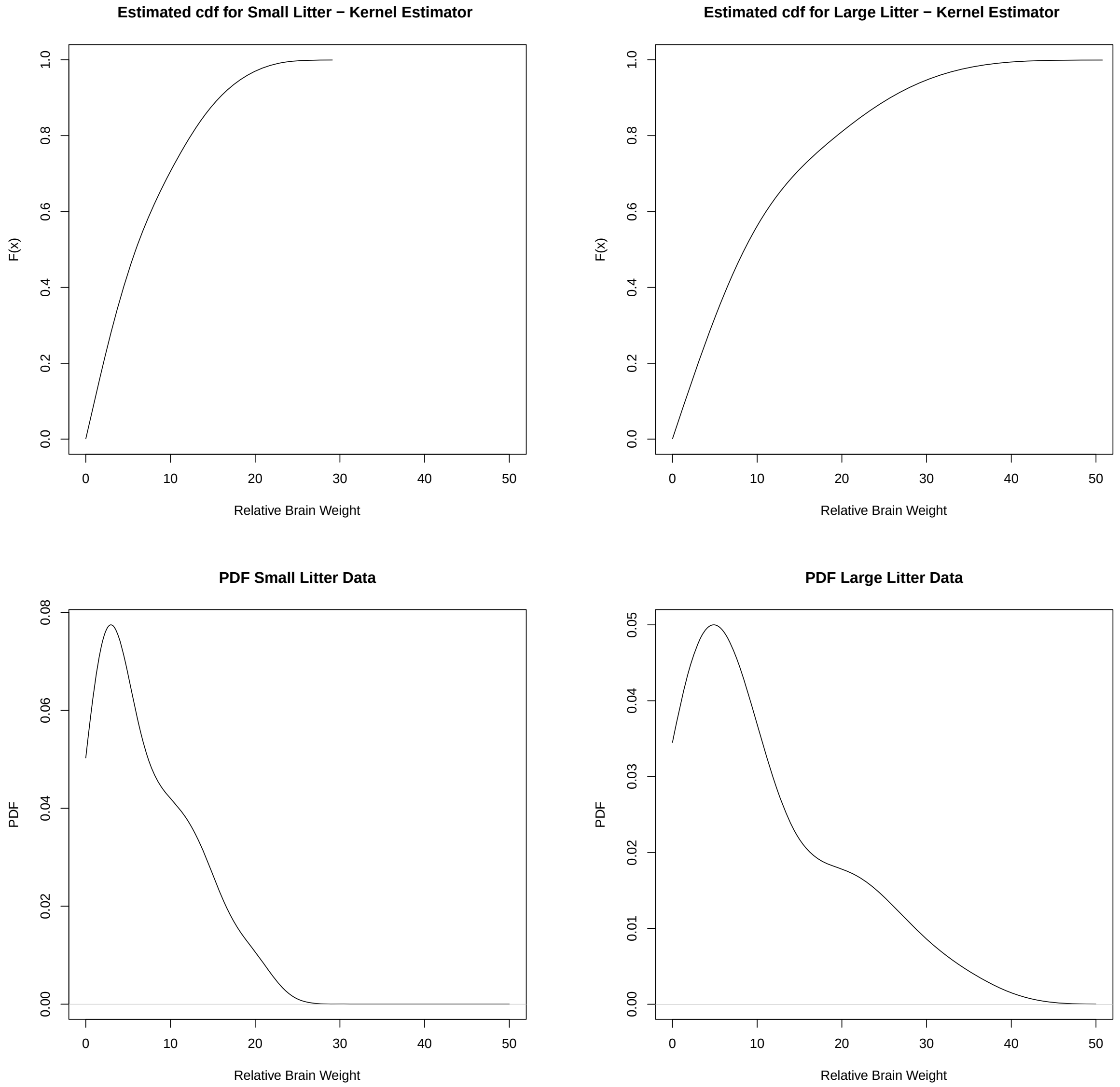
(c.) The data value provides the smallest contribution to the estimator at y=16, $\widehat{f}(16)$ is the data value furthest from 16, which is y = 35.45 with a contribution of 2.253479e-12 to $\widehat{f}(16)$=0.01669353. This is obtained by computing by hand:

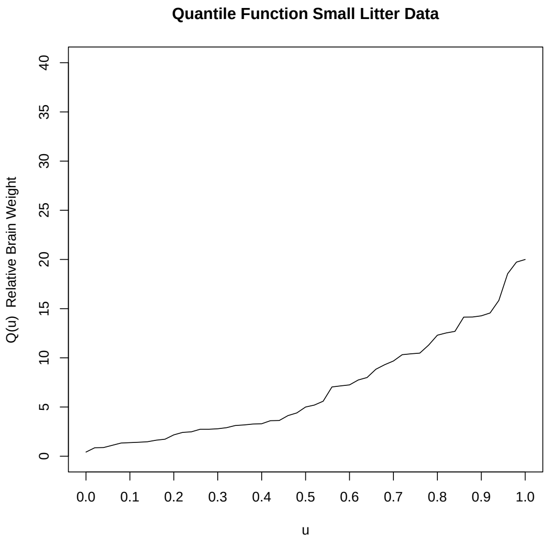
$$\frac{1}{nh}K\left(\frac{y-Y_i}{h}\right) = \frac{1}{44*3}K\left(\frac{16-35.45}{3}\right) = \left(\frac{1}{132}\right)\left(\frac{1}{2*\pi}\right)e^{-\left(\frac{16-35.45}{3}\right)^2/2} = 2.253479e^{-12}$$

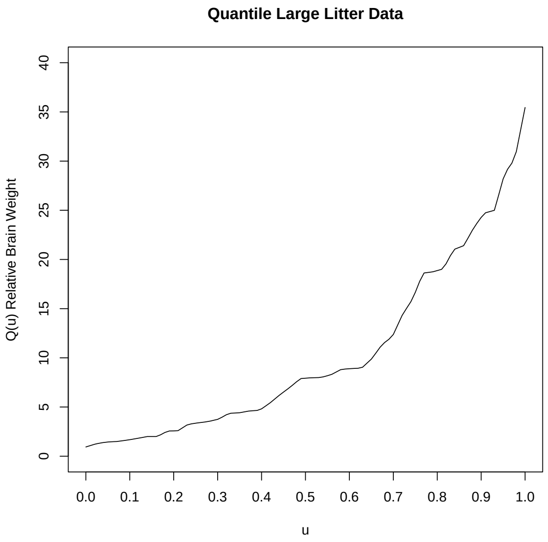
(d.) The data value provides the largest contribution to the estimator at y=16, $\widehat{f}(16)$ is the data value closest to 16, which is y = 16 with a contribution of 0.00302229 to $\widehat{f}(16)$=0.01669353. This is obtained by computing by hand:

$$\frac{1}{nh}K\left(\frac{y-Y_i}{h}\right) = \frac{1}{44*3}K\left(\frac{16-16}{3}\right) = \left(\frac{1}{132}\right)\left(\frac{1}{2*\pi}\right)e^{-\left(\frac{16-16}{3}\right)^2/2} = .00302229$$

P4. (28 points) (a.) Plots of pdfs (kernel density estimator), edf (smoothed), and quantile (smoothed):







(a)

See code at end of document.

( b.) Small Litter: Relative brain weights are somewhat right skewed which indicates that a few species of mammals with small average litters have large brains relative to their body weights.

Large Litter: Relative brain weights are highly right skewed which indicates that sizeable proportion of the species of mammals with large average litters have large brains relative to their body weights.

( c.) Based on the graphs, I would conclude that there is a positive relation ship between average litter size and relative brain weights. However, it would be more informative to have the actual litter sizes associated with each species to draw a more concrete conclusion.

P5. ( 28 points) Multiple Choice Questions:

1. **E** Given any one of the four functions then you can derive the other three from the given function
2. **D** See page 24 in Handout 4
3. **B** See page 24 in Handout 4
4. **D** See pages 30 & 32 in Handout 4
5. **B** See page 50 in Handout 4
6. **B** See page 37 in Handout 4
7. **D** See pages 16 & 17 in Handout 5
8. **C or D** See page 13 & 14 in Handout 5
9. **C** See page 23 in Handout 5
10. **A or D** See page 27 in Handout 5
11. **E** See page 14 in Handout 5
12. **E** See pages 25-27 in Handout 5

A. is false because $\sigma$ does not exist for Cauchy which is symmetric whereas both SIQR and MAD exist and are equal

B. is false because MAD is nearly always preferred to SIQR

C. is false because for a normal distribution MAD=SIQR

13. **B** See page 33 in Handout 5:

$\theta = 22.3,\ \rho = .6,\ \sigma_e^2 = 2.8\ \Rightarrow \mu_X = \frac{\theta}{1-\rho} = \frac{22.3}{1-.6} = 55.75$

$\sigma_X^2 = \frac{\sigma_e^2}{1-\rho^2} = \frac{2.8}{1-.36} = 4.375$

```
##
## (2)
##

dta <- read.csv("Assign3_BrainSize.csv")

y <- dta[, 2]
y <- y[!is.na(y)]
n <- length(y)

y_s <- sort(y)

## 0.25: (n - 1) * 0.25 + 1 = 11.75
(n - 1) * 0.25 + 1
y_s[11] + 0.75 * (y_s[12] - y_s[11])

## 0.5: (n - 1) * 0.5 + 1 = 22.5
(n - 1) * 0.5 + 1
y_s[22] + 0.5 * (y_s[23] - y_s[22])

## 0.75: (n - 1) * 0.5 + 1 = 33.25
(n - 1) * 0.75 + 1
y_s[33] + 0.25 * (y_s[34] - y_s[33])

##
## (3)
##

dd <- density(y_s)
plot(dd, type = "l")

h <- 3

## (a)
K_u <- function(u) {
  return(dnorm(u))
}

## f(3)
sum(K_u((3 - y_s) / h)) / (n * h)

## f(16)
sum(K_u((16 - y_s) / h)) / (n * h)

## (b)
brks <- y_s[1] + 5 * (0:7)
n_j <- hist(y_s, prob = FALSE, breaks = brks)$counts
R_j <- n_j / n
f_hat_j <- R_j / 5

hist(y_s, prob = TRUE, breaks = brks)$density

## (c)
kk <- K_u((16 - y) / h) / (n * h)
y[which.min(kk)]

## (d)
```

```
y[which.max(kk)]

##
## (4)
##

y_small <- dta[, 1]
y_large <- dta[, 2]
y_large <- y_large[!is.na(y_large)]

## (a)

## PDFs
par(mfrow = c(1, 2))
plot(density(y_small), xlab = "y", ylab = "f", main = "Small Litters", xlim = c(-10, 50),
  cex.axis = 0.75)
plot(density(y_large), xlab = "y", ylab = "f", main = "Large Litters", xlim = c(-10, 50),
  cex.axis = 0.75)

## EDFs
qq_small <- quantile(y_small, probs <- seq(0, 1, by = 0.01))
qq_large <- quantile(y_large, probs)

plot(qq_small, probs, type = "s", xlab = "y", ylab = "F", main = "Small Litters",
  xlim = c(0, 35), cex.axis = 0.75)
plot(qq_large, probs, type = "s", xlab = "y", ylab = "F", main = "Large Litters",
  xlim = c(0, 35), cex.axis = 0.75)

## Quantile functions
plot(probs, qq_small, type = "s", xlab = "Q(u)", ylab = "u", main = "Small Litters",
  ylim = c(0, 35), cex.axis = 0.75)
plot(probs, qq_large, type = "s", xlab = "Q(u)", ylab = "u", main = "Large Litters",
  ylim = c(0, 35), cex.axis = 0.75)
```