Department of Statistics
STATS 782 Statistical Computing
Assignment 3 Solutions (2019FC)

1. The following lines produce Figure 1. The coordinates of the reddish rectangle background were chosen to be $[-1, 1] \times [-1, 1]$.

```
> x <- 0.6    # A measure of how long the horizontal lines are
> y <- 0.38   # A measure of distance between the 'arrows'
> plot.new()
> plot.window(xlim = c(-1, 1), ylim = c(-1, 1),
              xaxs = "i", yaxs = "i", mar = rep(0, 4))
> usr <- par("usr")
> par(mar = c(0, 0,  0, 0), oma = rep(0, len = 4), mfrow = c(1, 1))
> rect(usr[1], usr[3], usr[2], usr[4], border = NA,
       col = "coral2")  # Big reddish rectangle background
> arrows(-x,  y, x,  y, col = "white", angle = 135, lwd = 4.5,
         code = 3, length = 0.4)  # Add the top 'arrow'
> arrows(-x, -y, x, -y, col = "white", angle =  45, lwd = 4.5,
         code = 3, length = 0.4)  # Add the bottom 'arrow'
```

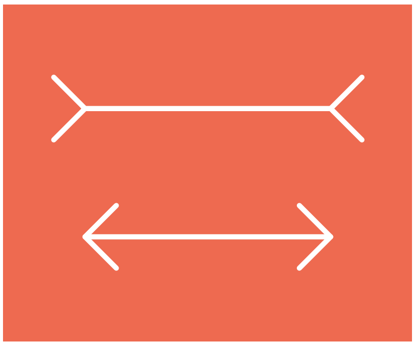

Figure 1: Hopefully identical to Fig. 2 but not perfect! The figure has `fig.width=3.7` and `fig.height=3.05` before scaling.

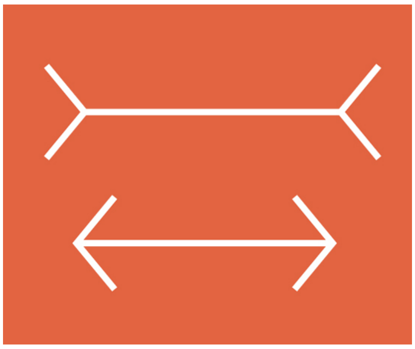

Figure 2: (Original) A figure that deceptively makes two parallel lines of equal length look of unequal length.

2. (a) Some of the following code is based on some statistical theory that you won't understand, nevertheless they produce the coordinates stated in the original question. The VGAM package is needed here too.

```
> my.cols <- c("darkgreen", "blue", "purple")  # For triangles
> my.cex <- 1.2
> mycol <- c("blue", "darkorange")
> mylwd <- 1.4
> my.tau <- 0.10
> my.exp <- qenorm(my.tau)  # The parent distribution is N(0, 1)
>
> myfoo <- function(x) x * dnorm(x) / pnorm(my.exp)  # LHS pdf only
> cen1 <- integrate(f = myfoo, lower = -Inf,   upper = my.exp)$value
> cen2 <- (my.exp + pnorm(my.exp)*(my.exp - cen1) / my.tau -
          pnorm(my.exp) * (2 * my.exp - cen1)) / (1-pnorm(my.exp))
> mu1 <- 0  # Mean of the parent distribution
> Cen1 <- my.exp + (my.tau / (2 * my.tau - 1)) *
          (mu1 - my.exp) / pnorm(my.exp)
> Cen2 <- my.exp + (mu1 - my.exp) * (my.tau - 1) / ((2 * my.tau - 1) *
          pnorm(my.exp, lower.tail = FALSE))
>
> # Now plot the curve etc.
> all.yyy <- seq(-3.3, 3.3, len = 201)
> ygrid <- seq(my.exp, max(all.yyy), len = 201)
> plot(ygrid, dnorm(ygrid), type = "l", col = mycol[1],
       ylab = "", xlab = "",
       axes = FALSE, xlim = range(all.yyy),
       lwd = mylwd, cex.lab = my.cex, cex.axis = my.cex,
       cex.main = my.cex,
       ylim = c(0, 0.4), main = "", lty = "solid")
>
> ygrid <- seq(min(all.yyy), my.exp, len = 201)
> lines(ygrid, dnorm(ygrid), col = "darkgreen", lwd = mylwd)
>
> lines(c(my.exp, my.exp), c(0-0.02, dnorm(my.exp)), col = my.cols[3],
        lty = "dashed", xpd = TRUE, lwd = mylwd)
>
> # Now annotate the figure with text, etc.
> abline(h = 0, col = my.cols[2])
>  triangle <-  2  # hollow triangle
> striangle <- 17  # solid triangle
> cofprop <- 3.5  # Constant of proportionality
> points(cen1, 0-0.02,   pch =  triangle,
         cex = cofprop * sqrt(  my.tau), col = my.cols[1], xpd = TRUE)
> points(cen2, 0-0.04,   pch =  triangle,
         cex = cofprop * sqrt(1-my.tau), col = my.cols[2], xpd = TRUE)
> points(my.exp, 0-0.04, pch = striangle, cex = cofprop * sqrt(0.5),
         col = my.cols[3], xpd = TRUE)
>
> mylab <- bquote(paste(mu, "(", omega == .(my.tau), ")"))
> text(my.exp, 0-0.09, labels = mylab, xpd = TRUE,
       cex = 1.2, col = my.cols[3])
> text(Cen1,   0+0.02, labels = expression(c[1]),
       cex = 1.2, col = my.cols[1])
> text(Cen2,   0+0.02, labels = expression(c[2]),
       cex = 1.2, col = my.cols[2])
```

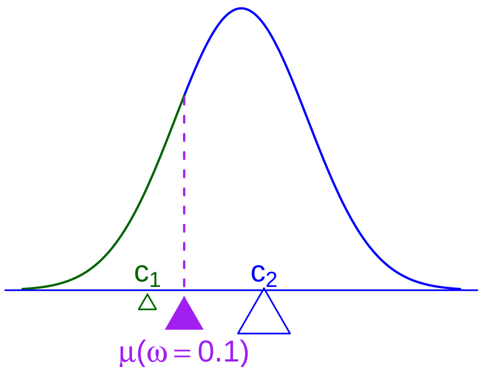

Figure 3: Illustration of the interpretation of expectiles in terms of centres of balance: the hollow triangles at positions c_1 and c_2. Here, the vertical dashed line is at the 0.1-expectile, and the solid triangle at $\mu(\omega = 0.1)$. The figure has `fig.width=3.5` and `fig.height=2.5` before scaling.

(b) I may look at your plots to get some good ideas, for the second edition of my book. Please let me know if you have any objections. . .

3. The following produces Fig. 4.

```
> par(mar = c(0, 0,  0, 0), oma = rep(0, len = 4), mfrow = c(1, 1))
> swirls <- 3  # Any positive number really
> # Angle (use polar coordinates):
> theta <- seq(0, 2 * pi * swirls, length = round(swirls * 73))
> x <- cos(theta)  # Outer circle(s) starting 3 o'clock going
> y <- sin(theta)  # anti-clockwise.
> x2 <- (theta / (2 * pi * swirls)) * cos(theta)  # Swirls
> y2 <- (theta / (2 * pi * swirls)) * sin(theta)  # going in.
>
> x3 <- pmin(1, 1.2 * x2)  # Expand out and trim off excess
> y3 <- pmin(1, 1.2 * y2)
>
> r3 <- sqrt(x3^2 + y3^2)  # Choose subset within outer circle
> x3[r3 >= 1] <- x[r3 >= 1]
> y3[r3 >= 1] <- y[r3 >= 1]
>
> plot.new()
> plot.window(xlim = c(-1, 1), ylim = c(-1, 1), asp = 1)
> polygon(x, y, col = "darkseagreen")  # Outer circle
> polygon(c(x2, rev(x3)), c(y2, rev(y3)), col = "white",
          border = NA)  # Colour in the swirl white
> lines(x2, y2, col = "red", lwd = 2)  # Boundary of swirl
> lines(x3, y3, col = "red", lwd = 2)  # The other boundary
```

4. (a)

```
> sydneytemp <- read.csv(paste0(foldername,
    "IDCJAC0002-066062-Data12edited.csv"), header = TRUE)
> # head(sydneytemp)  # Quick look and check
> summary(sydneytemp[, -1])

 StationNumber        Year           Jan             Feb              Mar
 Min.   :66062   Min.   :1859   Min.   :23.1   Min.   :23.60   Min.   :22.4
 1st Qu.:66062   1st Qu.:1899   1st Qu.:25.2   1st Qu.:25.00   1st Qu.:24.1
```

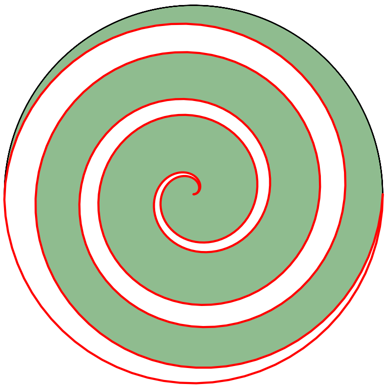

Figure 4: A figure for you to mimic. The figure has `fig.width=4` and `fig.height=4` before scaling.

```
 Median :66062   Median :1939   Median :25.9   Median :25.70   Median :24.7  
 Mean   :66062   Mean   :1939   Mean   :26.0   Mean   :25.82   Mean   :24.8  
 3rd Qu.:66062   3rd Qu.:1979   3rd Qu.:26.8   3rd Qu.:26.30   3rd Qu.:25.5  
 Max.   :66062   Max.   :2019   Max.   :29.6   Max.   :29.00   Max.   :27.1  
                                               NA's   :1       NA's   :1     
      Apr             May             Jun             Jul             Aug       
 Min.   :20.00   Min.   :16.60   Min.   :14.10   Min.   :13.40   Min.   :14.90  
 1st Qu.:21.60   1st Qu.:18.60   1st Qu.:16.20   1st Qu.:15.38   1st Qu.:17.00  
 Median :22.45   Median :19.50   Median :17.00   Median :16.50   Median :17.90  
 Mean   :22.49   Mean   :19.51   Mean   :16.99   Mean   :16.39   Mean   :17.88  
 3rd Qu.:23.40   3rd Qu.:20.32   3rd Qu.:17.90   3rd Qu.:17.40   3rd Qu.:18.80  
 Max.   :26.10   Max.   :23.20   Max.   :20.30   Max.   :19.90   Max.   :21.30  
 NA's   :1       NA's   :1       NA's   :1       NA's   :1       NA's   :1      
      Sep             Oct             Nov             Dec             Annual     
 Min.   :17.40   Min.   :19.20   Min.   :20.60   Min.   :22.50   Min.   :20.10  
 1st Qu.:19.00   1st Qu.:21.30   1st Qu.:22.70   1st Qu.:24.30   1st Qu.:21.30  
 Median :19.85   Median :22.00   Median :23.60   Median :25.25   Median :21.70  
 Mean   :20.08   Mean   :22.16   Mean   :23.67   Mean   :25.24   Mean   :21.75  
 3rd Qu.:21.10   3rd Qu.:23.10   3rd Qu.:24.52   3rd Qu.:26.10   3rd Qu.:22.20  
 Max.   :24.40   Max.   :26.20   Max.   :26.50   Max.   :28.60   Max.   :23.80  
 NA's   :1       NA's   :1       NA's   :1       NA's   :1       NA's   :1      

> range(rowMeans(sydneytemp[, month.abb]) -
        sydneytemp[, "Annual"], na.rm = TRUE)  # A quick check

[1] -0.05  0.05
```

There is a little bit of round-off error, but it is negligible and to be expected. The data seems okay with respect to that check.

(b) The following lines produce Figure 5.

```
> smallst <- subset(sydneytemp, 1910 <= Year & Year <= 2018)
> use.mean <- with(subset(sydneytemp, 1961 <= Year & Year <= 1990),
                   mean(Annual))  # See bottom caption of the figure
> smallst <-
    transform(smallst, cen.Annual = Annual - use.mean)
>
> # Set up the overall plot
> plot(cen.Annual ~ Year, data = smallst, xpd = TRUE,
       xlim = c(1904, 2022), ylim = c(-1.29, 1.34),
      xlab = "", ylab = "", xaxs = "i", yaxs = "i",
      col = ifelse(cen.Annual > 0, "red3", "turquoise4"),
      bty = "n", las = 1, axes = FALSE, lwd = 3.5, type = "h")
>
> # Add axes
> xgrid <- seq(1910, 2018, by = 18)
> ygrid <- seq(-1, 1, by = 0.5)
> mymgp <- c(8, 2.1, 0)  # Default: mgp = c(3, 1, 0)
> mycex.axis <- 1.2
> axis(1, at = xgrid, mgp = mymgp, xpd = TRUE, col = "gray90",
      cex.axis = mycex.axis, line = NA, pos = -1, lwd = 0.4)
> axis(2, at = ygrid, mgp = mymgp, xpd = TRUE, col = "gray90",
      labels = c("-1", "-0.5", "0", "0.5", "1"), lwd = 0.4,
      cex.axis = mycex.axis, las = 1, line = NA, pos = 1910)
>
> # Add gridlines
> abline(v = xgrid, h = ygrid, lwd = 0.4, col = "gray90")
>
> # Repeat plotting the bars
> lines(cen.Annual ~ Year, data = smallst,
      col = ifelse(cen.Annual > 0, "red3", "turquoise4"),
      lwd = 3.5, type = "h")
>
> abline(h = 0, lwd = 1.8, col = "gray50")  # Thicker darker grid line
>
> # Top captions
> mycex <- 1.46
> text(1897, 1.89, "Australia has been getting warmer",
      adj = 0, cex = mycex + 0.34, xpd = TRUE, font = 2)
> text(1897, 1.6, adj = 0, cex = mycex, xpd = TRUE,
      expression("Annual mean temperature above or below average " *
                 (degree*C)))
>
> lines(c(1897, 2022), c(-1.79, -1.79), lwd = 0.4, col = "gray50",
       xpd = TRUE)  # Separates the two bottom lines
>
> # Bottom captions
> mycex <- 0.9
> text(1897, -1.7, "Note: Average is calculated from 1961-1990 data",
      adj = 0, cex = mycex, xpd = TRUE)
> text(1897, -1.9, "Source: Australia Government Bureau of Meteorology",
      adj = 0, cex = mycex + 0.1, xpd = TRUE)
```

(c) Adding all the data (1859 is the first year) means that the captions have to be moved around. This is quite labour-intensive and needs manual work; the code below could have been improved in that sense.

It was pretty cold way back then, and the big mass of green at the LHS bottom means that some of the labelling has been obscured; this means that the plot should be redesigned really.

```
> use.mean <- with(subset(sydneytemp, 1961 <= Year & Year <= 1990),
                   mean(Annual))  # See bottom caption of the figure
> sydneytemp <-
    transform(sydneytemp, cen.Annual = Annual - use.mean)
>
> # Set up the overall plot
> plot(cen.Annual ~ Year, data = sydneytemp, xpd = TRUE,
       xlim = c(1850, 2022),
       ylim = c(-1.29, 1.34),
       xlab = "", ylab = "", xaxs = "i", yaxs = "i",
       col = ifelse(cen.Annual > 0, "red3", "turquoise4"),
       bty = "n", las = 1, axes = FALSE, lwd = 3.5, type = "h")
>
> # Add axes
> xgrid <- seq(1856, 2018, by = 18)
> ygrid <- seq(-1, 1, by = 0.5)
> mymgp <- c(8, 2.1, 0)  # Default: mgp = c(3, 1, 0)
> mycex.axis <- 1.2
> axis(1, at = xgrid, mgp = mymgp, xpd = TRUE, col = "gray90",
       cex.axis = mycex.axis, line = NA, pos = -1, lwd = 0.4)
> axis(2, at = ygrid, mgp = mymgp, xpd = TRUE, col = "gray90",
       labels = c("-1", "-0.5", "0", "0.5", "1"), lwd = 0.4,
       cex.axis = mycex.axis, las = 1, line = NA, pos = 1860)
>
> # Add gridlines
> abline(v = xgrid, h = ygrid, lwd = 0.4, col = "gray90")
>
> # Repeat plotting the bars
> lines(cen.Annual ~ Year, data = sydneytemp,
       col = ifelse(cen.Annual > 0, "red3", "turquoise4"),
       lwd = 3.5, type = "h")
>
> abline(h = 0, lwd = 1.8, col = "gray50")  # Thicker darker grid line
>
> # Top captions
> mycex <- 1.46
> text(1857, 1.89, "Australia has been getting warmer",
       adj = 0, cex = mycex + 0.34, xpd = TRUE, font = 2)
> text(1857, 1.6, adj = 0, cex = mycex, xpd = TRUE,
       expression("Annual mean temperature above or below average " *
                  (degree*C)))
>
> lines(c(1857, 2022), c(-1.79, -1.79), lwd = 0.4, col = "gray50",
        xpd = TRUE)  # Separates the two bottom lines
>
> # Bottom captions
> mycex <- 0.9
> text(1857, -1.7, "Note: Average is calculated from 1961-1990 data",
       adj = 0, cex = mycex, xpd = TRUE)
> text(1857, -1.9, "Source: Australia Government Bureau of Meteorology",
       adj = 0, cex = mycex + 0.1, xpd = TRUE)
```

(d) There are many things one could try. Here, we fit a simple linear regression to all the data from 1964 and onwards, because that subset appears to be linear by eye. Furthermore, prediction intervals are very easily obtained from them. Using all the data from 1910 onwards would be a mistake because it appears reasonably constant prior to 1950.

```
> fit1 <- lm(cen.Annual ~ Year, data = smallst, subset = Year >= 1964)
> predict(fit1, data.frame(Year = 2030), interval = "prediction")

       fit      lwr      upr
1 1.316277 0.421176 2.211378

> fit2 <- lm(Annual ~ Year, data = smallst, subset = Year >= 1964)
> predict(fit2, data.frame(Year = 2030), interval = "prediction")
```

```
       fit      lwr      upr
1 23.46961 22.57451 24.36471
```

The advantage of this method is that it is simple; one can obtain prediction intervals easily too. The results suggest it may be a lot hotter in the future—however we are making the assumption that the trend will continue unchanged into the next decade—and this is a strong assumption. The prediction interval is wide, which reflects the small amount of data used here.

Something to be done later though: look at the residuals to check that the model is okay.

(e) Use the following result. Let $M_N = \max(Y_1, \ldots, Y_N)$ where Y_i are i.i.d. random variables from a continuous cumulative distribution function F. Then

$$\Pr(M_N \le y) \;=\; \prod_{j=1}^{N} \Pr(Y_j \le y) \;=\; [F(y)]^N.$$

Suppose that we can find normalizing constants a_N and $b_N > 0$ such that (this is convergence in distribution)

$$P\left(\frac{M_N - a_N}{b_N} \le y\right) \;\longrightarrow\; G(y) \tag{1}$$

as $N \to \infty$, where G is some proper distribution function. Then G becomes useful as it is a nondegenerate limiting distribution function. For us, suppose that $Y_j \sim \text{Exp}(\lambda = 1)$ independently, hence $F(y) = 1 - e^{-y}$. For $b_N = 1$ and $a_N = \log N$,

$$\begin{aligned}
\Pr(M_N - \log N \le y) \quad &= \quad [F(y + \log N)]^N \;=\; \left[1 - e^{-y - \log N}\right]^N \\
&= \quad \left[1 - \frac{e^{-y}}{N}\right]^N \;\to\; \exp\left\{-e^{-y}\right\}.
\end{aligned}$$

Now we want N so that $\Pr(M_N \le 6.5) \le 0.95$, which means that

$$\Pr(M_N - \log N \le 6.5 - \log N) \;\approx\; \exp\left\{-e^{-[6.5 - \log N]}\right\}.$$

Solving this gives

```
> (N <- exp(6.5 + log(-log(0.95))))

[1] 34.11731

> ceiling(N)  # Approximate waiting time, in years

[1] 35
```

We round upwards because the probability of an exceedance increases over time.

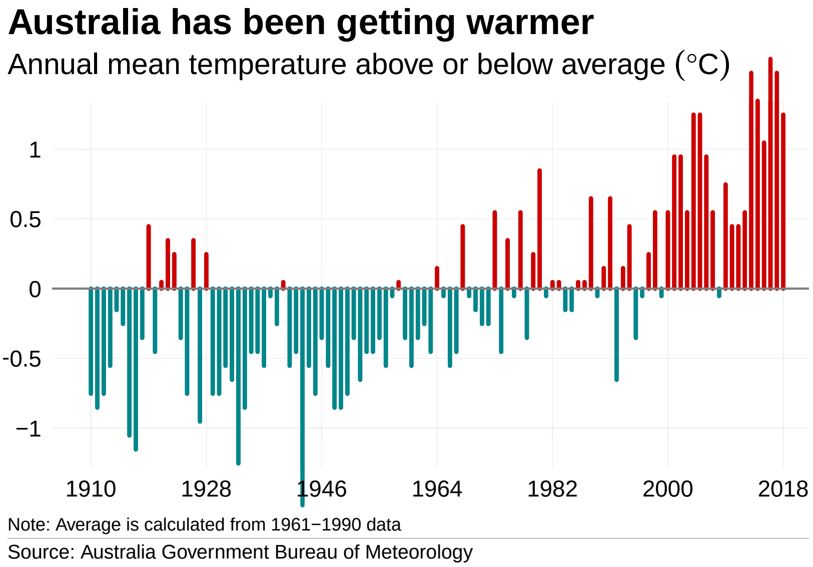

Figure 5: Hopefully identical to Fig. 6 but not perfect! The figure has `fig.width=6.95` and `fig.height=5.3` before scaling.

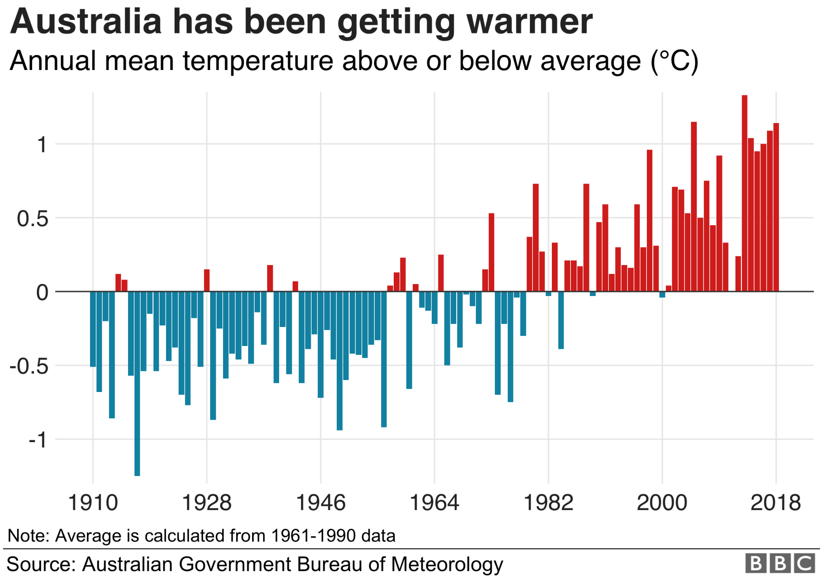

Figure 6: A figure from BBC News.

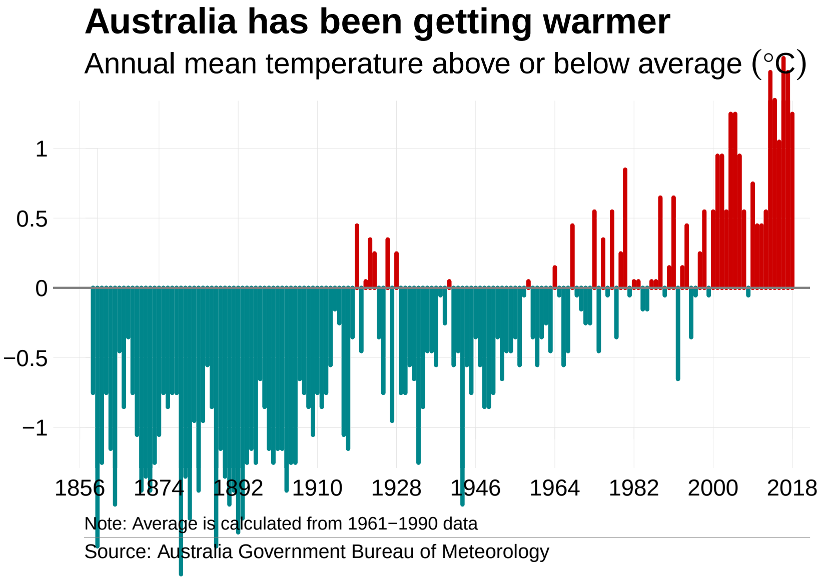

Figure 7: Same as Fig. 5 but having all the data. Some of the labelling should be moved a little.