Truncated distributions

---

35 marks

The questions here are designed to explore some basic characteristics of, and differences between, probability distributions and the random realizations from them. See `help("Distributions")` for those built into R .

1. Suppose we have a continuous random variable X with distribution function $F_X(x) = Pr(X \leq x)$ and quantile function $Q_X(p) = F_X^{-1}(p)$. That is $p = F_X(x) = Pr(X \leq x)$ and $p = Pr(X \leq Q_X(p)) = F_X(Q_X(p)) = F_X(F_X^{-1}(p)) = p$.

   We can define a random variable Y having cumulative distribution function

$$G_Y(y) = \begin{cases} 0 & y < a \\ \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)} & a \leq y \leq b \\ 1 & y > b \end{cases}$$

   where $-\infty \leq a < b \leq \infty$ and X is a continuous random variable as above.

   That is, Y has the same distribution as X **except** that it is **truncated** on the left at a and on the right at b. Unlike X, Y cannot take values less than a or larger than b.

   a. *(3 marks)* Mathematically show that the random variable W defined

$$W = Q_X(F_X(a) + U \times (F_X(b) - F_X(a)))$$

      where U is a uniform random variable $U \sim U(0, 1)$ has the same distribution as Y.

   b. *(15 marks)* Here you are to write a function `truncate()` of the form

```
truncate <- function(ddist = dnorm, pdist = pnorm,
                     qdist = qnorm, a = -inf, b = inf) {
                      # your code here
                      }
```

      where `ddist`, `pdist`, and `qdist` refer to functions which calculate the density $f_X(x)$, distribution (cumulative probability) $F_X(x)$, and quantiles $Q_X(p)$, for the input distribution of the random variable X. The arguments a and b ($a < b$) are the truncation points. Note that `-inf` and `inf` are representations in R of $-\infty$ and $+\infty$, respectively. Your code will need to be able to handle all cases correctly.

      The function `truncate()` is to return a list with components named `ddist`, `pdist` and `rdist` containing functions which can be called to produce the density $g_Y(y)$, distribution $G_Y(y)$, and any number of pseudo-random observations from the distribution of Y.

      That is, the following should work for the half-normal distribution.

```
half_normal <- truncate(a = 0)
xsample <- half_normal$rdist(300)
x <- seq(-3, 3, 0.01)
fx <- half_normal$ddist(x)
Fx <- half_normal$pdist(x)
```

```
oldPar <- par(mfrow = c(1,3))
plot(x, fx, type = "l", main = "Half normal density")
plot(x, Fx, type = "l", main = "Half normal distribution")
hist(xsample, main = "Half normal sample")
par(oldPar)
```

Hand in the above plots with your code.

c. A 2011 article by Gil Greengross and Geoffrey Miller of the University of New Mexico was entitled "Humor ability reveals intelligence, predicts mating success, and is higher in males" and appeared in Volume 39 of the journal *Intelligence* (pp. 188-192).

They tested the sense of humour of 400 university students (200 men, 200 women) using a standardized method and found that their measures of humour had about the same standard deviation (0.49) but differed in means with the men scoring an average of 0.09 and the women an average of -0.09 (the higher the score the greater the "humour ability"). The difference in means was found to be statistically significant ($p < 0.001$).

The distributions of humour between men and women seems to be significantly different but what does that actually say? To get some idea, suppose we take the results to mean that the measure of humour ability for men is $Y \sim N(0.09, (0.49)^2)$ and the same for women is $X \sim N(-0.09, (0.49)^2)$.

i. *(3 marks)* On a single (nicely labelled with a legend) draw the densities (in different colours) for both men and women.

ii. *(4 marks)* Generate a random sample of 1000 scores from each of these distributions and save the values on `x` for women, `y` for men, and `results <- data.frame(women = x, men = y)`. We now have paired results as if in each row, we randomly drew one women and one man and measured their "humour ability".

Based on your sample, estimate the following

- the average humour ability of the men
- the average humour ability of the women
- the probability that the man will be funnier than the woman (at least as measured by this scale).

iii. *(4 marks)* Are women funny? Suppose that to be really funny (e.g. professional standup comedian) requires a humour ability measure of at least 1.07 (two standard deviations past the mean of the males)

Generate 1000 pseudo random scores `y` from the truncated distribution for men and another 1000 `x` from the truncated distribution for women. Form the data frame `funny <- data.frame(women = x, men = y)` and based on this paired sample, estimate the following

- the average humour ability of the men
- the average humour ability of the women
- the probability that the man will be funnier than the woman (at least as measured by this scale).

iv. *(2 marks)* What conclusions do you draw about the differences between the humour of men and women?

v. *(4 marks)* Repeat part iii, again conditioning on considering only individuals with a "humour ability" score of at least 1.07. Except now, assume that the means of the female and male score distributions are identical at 0.09 **but** that the standard deviation for the men is 10% larger for than that for the women (0.049).