461 HW 1-22-14 Simulation, kernel densities.
Reminder, there is no class Monday 1-20-14 due to MLK obser-
vance.

To prepare read pp. 296-318 and the material of this handout
which can appear on exam 1 set for Wednesday 1-29-14.

1. In R create a discrete probability distribution from which you
can sample.

**Step one:** We’ll do this by merging data of different types. Pre-
pare an x consisting of 100000 iid samples from a continuous distri-
bution. Prepare a y consisting of an iid sample of 50000 from a
second discrete or continuous distribution. Merge x, y into w <-
c(x, y). Prepare a population “pop” consisting of a random sample
of 500000 drawn from w with equal probability and **WITH** replace-
ment. This will allow some repeats.

**Step two.** Although pop is partly continuous and partly discrete
plot the “density” of pop using the defaults. This density under-
takes a smoothing of the population scores, such as is used when
preparing to view a “density” representation of sample scores
which are always discrete. Understand that the density of sam is
an estimate of the sample kernel density method applied to pop.
Strictly speaking pop has no continuous density of the kind being
prepared and viewed.

**Step three.** Prepare an iid sample “sam” of 1000 from pop. Plot
the default “density” from this sample. Comment on the degree to
which sam has revealed the characteristics of pop. To facilitate
your comparison plot the two densities together in a single plot.
Be sure to indicate by hand which is which on your submission.
**Step four.** Repeat step three 10 times showing a plot of all ten sample densities together with the properly identified density of pop. This plot tells us how much sampling variation there is from one density(sam) to another density(sam). When we draw a sam and prepare its density plot we naturally would like to know what the sampling variation of the plot looks like. But we can’t draw lots of samples of 1000 just to get a look at it.

**Step five.** Bootstrap. To get a look at the sampling variations of density(sam) around density(pop) we have recourse to a bootstrap principle that for nearly all samples sam, under some conditions assures that

\[ \text{sampling variations of density(sam) about density(pop)} \]

can be estimated by

\[ \text{sampling variations of density(sam*) about density(sam)} \]

where sam* denotes an iid sample of 1000 from sam.

We have learned to trust sample sd s as a suitable substitute for unknown population sd \( \sigma \) when constructing CI. In the bootstrap this is amped waaaaaay up. We are saying something like

*Nearly all random samples are so good at representing the population that sampling from our sample (sam* from sam), which we can repeat without much cost using our original sample, provides distribution of \( \theta \text{sam*}-\theta \text{sam} \) that approximates the distribution of \( \theta \text{sam}-\theta \text{pop} \) only available with fresh sam.*

**Prepare a plot of 10 such replicated densities** density(sam*) together with the properly identified density(sam). Comment on the degree to which this plot reveals the general character of your plot of Step five. This is of course a statistical comparison based on only 10 replications of each. In practice we might use 100 instead of 10 plots to see a more revealing comparison.

**Step six. (Optional).** A “butterfly” plot can be used to compare the plot of 10 density(sam*) & density(sam) with the plot of 10 density(sam) & density(pop). To do it you plot side by side

| left to right flipped | 10 density(sam*) − density(sam) | 10 density(sam) − density(pop) |

Just as you can immediately appreciate the symmetry of a butterfly you will immediately appreciate the general comparison of these two plots. A left to right flipped plot can be prepared by replacing x with -x when plotting.