```
x_star = normrnd(mu_4,sigma_4);


>> r = rand(1);  % a single uniform draw in (0,1)
k_star = find(r <= cum_weights, 1)

k_star =

     3
% once more, we sampled the third component
% so let's simulate accordingly from the third Gaussian:
x_star = normrnd(mu_3,sigma_3);
```

In summary, we obtained independently three `x_star` values, the (x_1^*, x_2^*, x_3^*) we wanted:. Notice that since the mixture components are 1-dimensional Gaussians we used `normrnd()`, which wants a standard deviation as second argument (see the last page of this document).

Exercise 1 (to be solved with MATLAB), 6 points

You are given simulated data as a file `diabetes` on Canvas. This synthetic dataset consists of 500 blood glucose concentrations (mg/dL) each obtained from 500 hypothetical subjects.

```
% load data
load('diabetes.txt');
```

We assume that measurements have been obtained on three groups of subjects, healthy, prediabetics and diabetics, where "healthy" subjects have smaller blood glucose measurements while the diabetic ones have the largest measurements. You can easily see the data distribution and notice the three groups by looking at an histogram[6] of the data.

(i) (a)[0.5 point] Fit a Gaussian mixture model with three components on the given data, and report the obtained maximum likelihood estimates for all the parameters $(\mu_k, \sigma_k, \pi_k)_{k=1}^K$ (notice I wrote σ_k and not σ_k^2). It is important to obtain stable estimates by providing suitable starting values for the parameters so not to obtain the "switching" problem.
(b) [0 points] What is the probability that a subject from the population is prediabetic?

(ii) [1 points] Simulate a vector of $n = 500$ observations from a three-components Gaussian mixture using the estimated parameters you obtained in (i). Compare the histogram of the simulated data with the histogram of the original `diabetics` dataset. Do you observe a good agreement? *[Place rng(123) in the code before simulating data so we get the same results.]*

(iii) [2 points] Run a nonparametric bootstrap procedure to infer the distribution of all parameters $(\mu_k, \sigma_k, \pi_k)_{k=1}^K$, using $B = 2,000$. For each parameter, report the histogram for the bootstrap distribution, and the 95% confidence intervals based on the percentile method. *[Place rng(123) in the code before simulating data so we get the same results.]*

[6] By the way, `histogram()` and `hist()` both work, the outputs are slightly different due to different algorithms to bin the data. Whatever you use is ok, but the former one is nicer and recommended.

(iv) [2 points] The same as in (iii) but using the parametric bootstrap. This time though, report the histograms only (not the percentiles). Do you see any appreciable difference or is the method giving similar results to the nonparametric one? And if you were told that the diabetes.txt data were indeed generated by a Gaussian mixture model with three components, and not by some other type of model, what would it be your appreciation of the performance of the nonparametric bootstrap at this point?

(v) [0.5 points] Using the results from the nonparametric bootstrap, what is the probability that the estimated mean glucose concentration is larger than 85 for healthy subjects? Also, what is the probability that the estimated mean glucose concentration is larger than 85 for prediabetic subjects?