18.338J/16.394J: The Mathematics of Infinite Random Matrices
Histogramming

Professor Alan Edelman

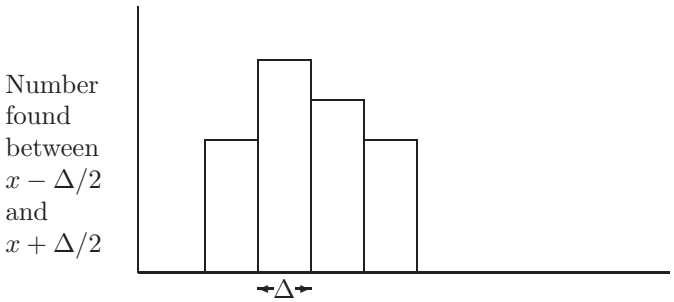
Handout #2, Tuesday, September 14, 2004

1 Random Variables and Probability Densities

We assume that the reader is familiar with the most basic of facts concerning continuous random variables or is willing to settle for the following sketchy description. Samples from a (univariate or multivariate) experiment can be histogrammed either in practice or as a thought experiment. Histogramming counts how many samples fall in a certain interval. Underlying is the notion that there is a probability distribution which precisely represents the probability of falling into an interval.

If $x \in \mathbb{R}$ is a real random variable with *probability density* $p_x(t)$, this means that the probability that x may be found in an interval $[a, b]$ is $\int_a^b p_x(t)\mathrm{d}t$. More generally if S is some subset of $\mathbb{R}$, the probability that $x \in S$ is $\int_S p(t)\mathrm{d}t$. Later on, we may be more careful and talk about sets that are Lebesgue measurable, but this will do for now.

The probability density is roughly the picture you would obtain if you collected many random values of x and then histogrammed these values. The only problem is that if you have N samples and your bins have size Δ, then the total area under the boxes is $N\Delta$ not 1:



Therefore a normalization must occur for the total area under the boxes to equal to 1 so that the (normalized) histogram and probability densities can line up.

Normal distribution

The normal distribution with mean 0 and variance 1 (standard normal, Gaussian, "bell shaped curve") is the random variable with probability density $p_x(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. It deserves its special place in probability theory because of the *central limit theorem* which states that if $x_1, \ldots, x_n, \ldots$ are iid random variables (iid = independent and identically distributed) with mean μ and variance σ then

$$\lim_{n\to\infty} \mathrm{Prob}\Big(a < \frac{x_1 + \cdots + x_n - n \cdot \mu}{\sqrt{n}\,\sigma} < b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\,\mathrm{d}t\,.$$

The central limit theorem roughly states that a large collection of identical random variables behaves like the normal distribution. Many investigations into the eigenvalues of random matrices suggest experimentally that this statement holds, i.e., the eigenvalues of matrices whose elements are not normal behave, more or less, like the eigenvalues of normally distributed matrices.
It is of value to note that the normal distribution with mean μ and variance σ^2 has $p_x(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.

2 Univariate Histograms

In Figure 2, we plot the normal distribution as well as a histogram obtained from 5000 samples from the normal distribution We see in the second line of the code below that we divide the counts `n` by the total number times the bin size: `5000*0.2`. This guarantees that the total area of the boxes over the whole line is normalized to 1.

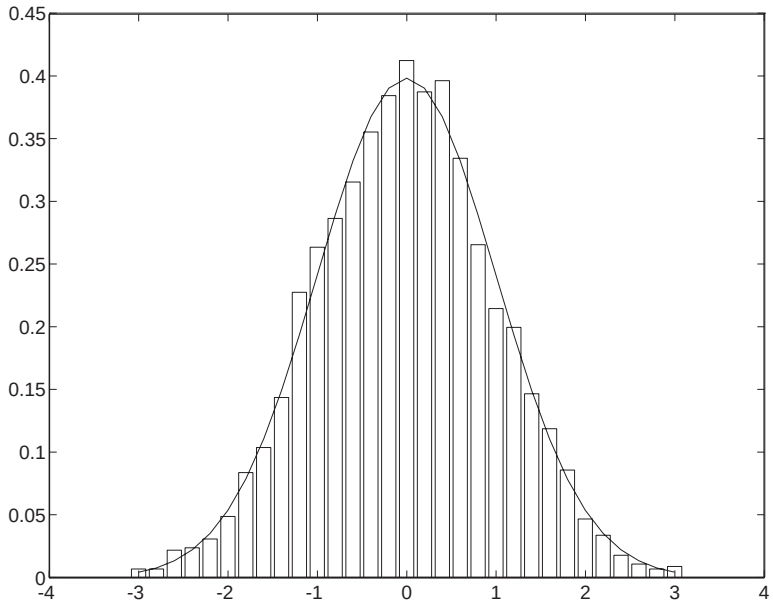


Figure 1: This figure illustrates the idea that the probability density is a histogram

Code 1 is our MATLAB code to obtain this figure.

```
>>  a=randn(1,5000);[n,x]=hist(a,[-3:.2:3]);
>>  bar(x,n/(5000*.2));
>>  hold  on,plot(x,exp(-x.^2/2)/sqrt(2*pi)),hold  off
```

```
%bellcurve.m
%Code 1.1 of Random Eigenvalues by Alan Edelman

%Experiment:  Generate random samples from the normal distribution.
%Observation: Histogram the random samples.
%Theory:      Falls on a bell curve.

trials=100000; dx=.2;

v=randn(1,trials);[count,x]=hist(v,[-4:dx:4]);
hold off, b=bar(x,count/(trials*dx),'y'); hold on

x=-4:.01:4;
plot(x,exp(-x.^2/2)/sqrt(2*pi),'LineWidth',2)
axis([-4 4 0 .45]);
```

Code 1

3 How Accurate Are Histograms?

When playing with Code 1, the reader will happily see that given enough trials the histogram is close to the true bell curve. One can press further and ask how close? Multiple experiments will show that some of the bars may be slightly too high while others slightly too low. There are many experiments which we explore in the exercises to try to understand this more clearly. We will discuss these as the course progresse

4 HISTN: Normalized Histogram

We can incorporate the ideas discussed above into the following MATLAB code.

```
function  [h,hn,xspan]=histn(data,x0,binsize,xf);
%HISTN  Normalized  Histogram.
%     [H,HN,XSPAN]  =  HISTN(DATA,X0,BINSIZE,XF)  generates  the  normalized
%     histogram  of  area  1  from  the  values  in  DATA  which  are  binned  into
%     equally  spaced  containers  that  span  the  region  from  X0  to  XF
%     with  a  bin  width  specified  by  BINSIZE.
%
%      X0,  BINSIZE  and  XF  are  all  scalars  while  DATA  is  a  vector.
%      H,  HN  and  XSPAN  are  equally  sized  vectors.
%
%      References:
%      [1]  Alan  Edelman,  Handout  2:  Histogramming,
%                        Fall  2004,  Course  Notes  18.338.
%      [2]  Alan  Edelman,  Random  Matrix  Eigenvalues.
%
%      Alan  Edelman  and  Raj  Rao,  Sept.  2004.
%      $Revision:  1.1  $   $Date:  2004/09/10   17:11:18  $

xspan=[x0:binsize:xf];

h=hist(data,xspan);          %  Generate  histogram
hn=h/(length(data)*binsize);  %  Normalize  histogram  to  have  area  1

bar(xspan,hn);               %  Plot  histogram
```

Code 2

We will use this code throughout the remainder of the course to corroborate theoretical predictions with experimental data.