

1 Univariate Histograms

In Figure 1, 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.

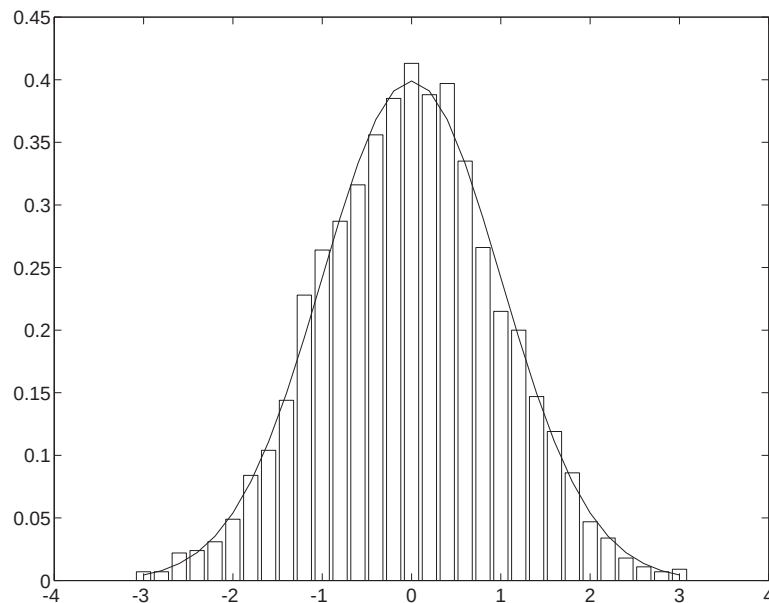


Figure 1: This figure illustrates the idea that the probability density is a histogram

Code 0.1 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
```

It is of interest to see the deviations from the true histograms. The following MATLAB code illustrates this

```
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([-440.45]);
```

Code 0.1

```
n=1000; trials=400; dx=.05; x=-6:dx:6; bell=exp(-x.^2/2)/sqrt(2*pi); cov = zeros(length(x)); dev =
[];
hold off for i=1:trials, a=randn(1,n); y=hist(a,x)/(n*dx); z=sqrt(n)*(y-bell)/sqrt(bell); dev=[dev
z]; cov=cov+z'*z; end cov=dx*cov/trials; cov=cov(2:end-1,2:end-1); dev=dev*sqrt(dx); dx=.2;
[count,x]=hist(dev,[-6:dx:6]);
figure(1) hold off, b=bar(x,count/(length(dev)*dx),'y'); hold on
x=-4:.01:4; plot(x,exp(-x.^2/2)/sqrt(2*pi),'LineWidth',2)axis([-440.45]);
figure(2) hold off plot(diag(cov(10:(end-10),10:(end-10))), 'r'); hold on; plot(diag(cov,1), 'b');
```

Code 0.2

2 How Accurate Are Histograms?

When playing with Code 0.1, 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.

In Code 0.2, we take many histograms and at each x we plot the average square deviation. This is not a histogram! The Exercises do explore histogramming the deviation. Here we just are looking at the variance at each point x .

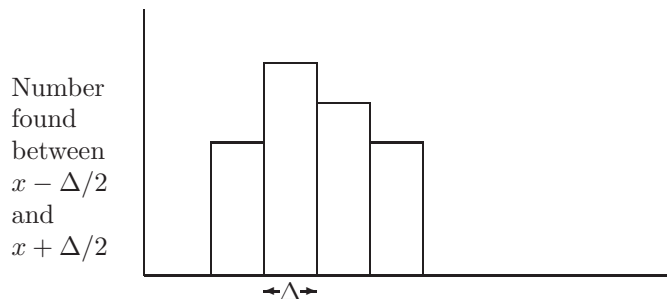
The experiment shows that the variance behaves like the height of the bell curve divided by $n * dx$. This means that as the number of samples gets larger, the variance in the height of the bar goes down proportionally. Also if one tries to get too greedy by making the boxwidth tiny, one runs the risk of having a very high variance in the bar heights.

3 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 $x \in R$ 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 $\int_a^b p_x(t) dt$. More generally if S is some subset of R , the probability that $x \in S$ is $\int_S p(t) dt$. 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 histogram and probability densities to 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 $x_1, \dots, x_n, \dots$ are iid random variables (iid = independent and identically distributed) with mean μ and variance σ then

$$\lim_{n \rightarrow \infty} \text{Prob} \left(a < \frac{x_1 + \dots + x_n - n \cdot \mu}{\sqrt{n} \sigma} < b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

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 $p_x(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$.

4 A computational realization

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

```
function [Hist,Histnorm] = normalized_hist(data, x_start, bin_size, x_end);  
x=[x_start : bin_size : x_end];  
Hist=hist(data,x);  
length_data = length(data); normalization_factor = length_data * bin_size;  
Histnorm = Hist/normalization_factor;
```

Code 0.3

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