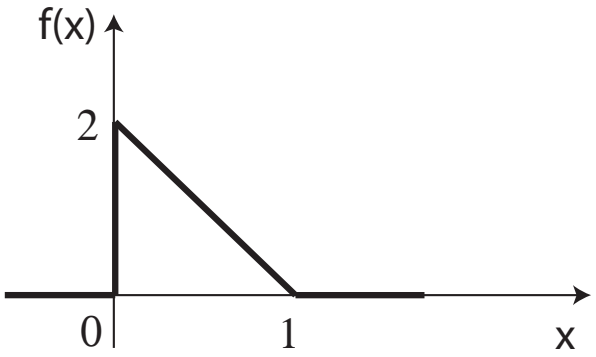
1. Suppose a continuous random variable X has the following probability density function:

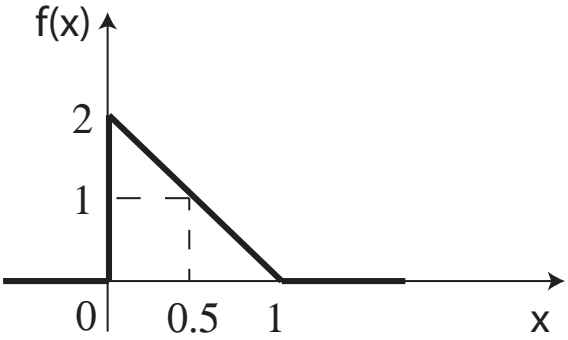


What is the probability P(0≤X≤1) ? What is the probability P(0≤X≤0.5) ?

Solution

(i) The area under the triangle is 1 (the area of the square with side of length 1 is 1, and the area of the triangle is half of that square). Therefore P(0≤X≤1)=1

(ii) Split the area of f(x) between 0≤x≤0.5 into rectangle and a triangle as follows:



The area of the rectangle is 1x0.5 = 0.5. The area of the smaller triangle is 0.25. So the total area is 0.75 and P(0≤X≤0.5)=0.75.

2. Suppose the delay experienced by a packet across a wireless link is Normally distributed with mean 1ms and variance 100ms. For a packet which traverses 10 such links one after the other, with the delay on each link being independent of the delay on the other links, how is the total delay distributed ?

Solution

Let D_i be the delay across link I, $D_i \sim N(0.001,0.1)$. The aggregate delay is $D=\sum_{i=1}^{10} D_i$. $E[D]=10E[D_i]=10x0.001=0.01$. $Var(D)=10Var(D_i)=10x0.1=1$. The sum D is Normally distributed $D \sim N(0.01,1)$.

3. Suppose we operate a large data centre with n servers. Suppose each server fails independently with probability p. Write an exact expression for the PMF of the number of failed servers. What would the Normal approximation to this exact distribution be ? Write a Matlab simulation to compare the measured distribution of failures with the Normal approximation.

Solution

The number X of failed servers is a Bin(n,p) random variable, P(X=x)=$\binom{n}{x} p^x(1-p)^{n-x}$. E[X]=np, Var(X)=np(1-p). So the Normal approximation to Bin(n,p) is N(np,np(1-p)).

Matlab:

```
P=[]; N=10000;

n=1000; p=0.01;

for i=1:N,

        X=sum(rand(1,n)<p);

        P=[P; X];

end

mu=n*p; sigma2=n*p*(1-p);

[n,x]=hist(P,[1:25]);

plot(x,n/(sum(n)),x,exp(-(x-mu).^2/(2*sigma2))/(sqrt(2*pi*sigma2)));
```

4. Write a paragraph comparing and contrasting the use of the Central Limit Theorem, Chebyshev's Inequality and Bootstrapping for estimating confidence intervals.

Solution

Start by explaining what a confidence interval is i.e. given a random variable X a confidence interval [a,b] is an interval such that P(a≤X≤b)≥ c where c is a target probability e.g. 0.95 or 0.99. Chebyshev Inequality allows us to select a and b so that this is true, but the values of a and b are likely larger than strictly necessary (since Chebyshev is only an upper bound). CLT says we can approximate the distribution of X as being Normal provided X is the sum of sufficiently many independent and identically distributed random variables. Main difficulty is in the "sufficiently many" part, so CLT confidence bounds may be overly optimistic i.e select interval [a,b] to be smaller than it ought to be. Bootstrapping uses multiple measurements of X to estimate its distribution and then from this to estimate confidence interval [a,b]. This requires sufficiently many observations

of X that the estimate of its distribution is accurate, but does not require that the distribution is Normal.

5. Suppose a user watches a movie and gives ratings. The rating reported by the user can change each time the user watches the movie, so the rating after viewing k is modeled as random variable $Y_k = X + V_k$ where X captures the user underlying preference which is unknown but stays fixed from viewing to viewing and V_k is ``noise'' which changes at each viewing. Random variable X is Normally distributed with mean 0 and variance σ^2. Noise V_k is also Normally distributed with mean 0 and variance σ_s^2, and so $f_{Y_k|X}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\sigma_s^2}\right)$.

A Matlab simulation of 10 ratings obtained by this process with $\sigma^2=1$ and $\sigma_s^2 =0.1$ is:

x=randn; y=[]; for k=1:10, v=0.1*randn; y=[y, x+v]; end, plot(y,'o')

Use Bayes Rule for PDFs to write an expression for the posterior PDF for X after a observing a single rating by the user. Write the expression in terms of σ^2, σ_s^2 and PDF $f_Y(y)$. Discuss how we might find the value of X that has highest posterior probability for the user. Illustrate using Matlab.

Solution

The model is that

$$f_{Y|X}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\sigma_s^2}\right)$$

and

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

By Bayes Rule for PDFs

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\sigma_s^2}\right) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)}{f_Y(y)}$$

Taking the log to simplify things a little,

$$\log f_{X|Y}(x|y) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(y-x)^2}{2\sigma_s^2} + \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{x^2}{2\sigma^2} - \log f_Y(y)$$

Dropping terms that do not depend on x, leaves us with:

$$-\frac{(y-x)^2}{2\sigma_s^2} - \frac{x^2}{2\sigma^2}$$

So to find the value x that maximizes $f_{X|Y}(x|y)$ its enough to find the x which maximizes $-\frac{(y-x)^2}{2\sigma_s^2} - \frac{x^2}{2\sigma^2}$ where y is the observed rating.

Matlab:

```
x=randn;
v=0.1*randn; y=x+v;  % obtain single rating
xhat=[-2:0.01:2]; plot([x x],[-10 10],xhat,-(xhat-y).^2/(2*0.01)-xhat.^2/2)
```