nurdan.kar@bilkent.edu.tr

4. Continuous Random Variables

4.1. Probability density functions and cumulative distribution functions

Recall that a rv is continuous if

1. Its set of possible values consists either of all numbers in a single interval on the number line, $(-\infty, \infty)$, or all numbers in a disjoint union of such intervals, $[0,5] \cup [5,10]$
2. $P(X=c)=0$ for any number c that is a possible value of X.

Example. Consider the velocity of a vehicle travelling along the highway. If this velocity is measured by a digital speedometer, the speedometer's reading is a discrete rv. To model the exact velocity, we need a continuous rv.

Some benefits of continuous rv can be listed as:

1. Finer-grained and possibly more accurate modeling
2. Powerful tools from calculus
3. Insightful analysis that would not be possible under a discrete model

In principle, variables such as height, weight and temperature are continuous. However, in practice, the measuring instruments restrict us to a discrete world.

Probability distributions for continuous variables

Definition. Let X be a continuous rv. Then, a probability distribution or probability density function (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

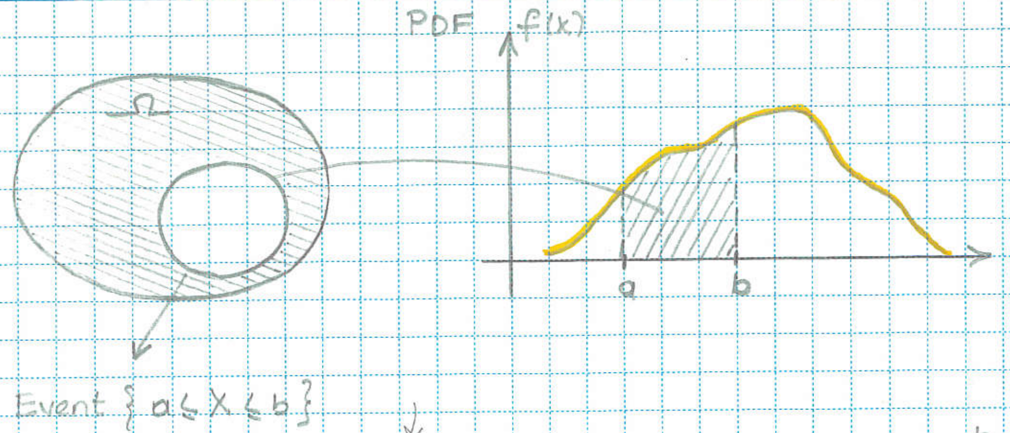
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Here, the graph of $f(x)$ is often referred to as the density curve.

The PDF above can be considered as:

$$P(X \in B) = \int_B f(x)\,dx \quad \rightarrow \quad P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

B is the union of a finite or countable numb. of intervals.



The probability that X takes value in an interval $[a,b]$ is $\int_a^b f(x)dx$, which is the shaded area in the figure.

Now, for any single value a, we have $P(X=a) = \int_a^a f(x)dx = 0$.

Since we have $P(X=a)=0$ and $P(X=b)=0$, including or excluding the endpoints of an interval has no effect on its probability. Namely,

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b).$$

A PDF must satisfy the following two conditions:

1. $f(x) \geq 0 \rightarrow f(x)$ is nonnegative
2. $\int_{-\infty}^{\infty} f(x)dx = 1 = P(-\infty < X < \infty) \rightarrow$ area under the entire graph of $f(x)$ is 1.

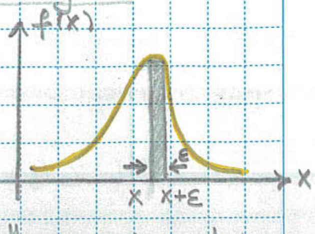
<u>Support of a PDF</u>

The <u>support</u> of a PDF is the set of all x-values where $f(x) > 0$. In other words, it shows where the rv can actually take values. Altough a PDF is defined for $-\infty < x < \infty$, we will consider the PDF in its support; otherwise $f(x) = 0$.

<u>Interpretation of the PDF as "probability mass per unit length"</u>

Note that for $[x, x+\varepsilon]$ with very small length ε, we have

$$P([x, x+\varepsilon]) = \int_x^{x+\varepsilon} f(t)dt \approx f(x) \cdot \varepsilon$$



so we can view $f(x)$ as the "probability mass per unit length" near x. Note that $f(x)$ is not the probability itself; that is it is not restricted to be less than or equal to 1.

Example. The amount of time in hours that a computer functions before breaking down is a continuous rv with PDF given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

What is the probability that

a) a computer will function between 50 and 150 hours before breaking down?

b) it will function for fewer than 100 hours?

Solution.

a) Since $1 = \int_{-\infty}^{\infty} f(x)dx = \lambda \int_{0}^{\infty} e^{-\frac{x}{100}} dx = -100\lambda e^{-\frac{x}{100}}\Big|_0^{\infty} = -100\lambda \frac{1}{e^{x/100}}\Big|_0^{\infty} = 100\lambda$

→ The support is $[0,\infty)$. $f(x) > 0$ in $[0,\infty)$.

we have $1 = 100\lambda \Rightarrow \lambda = \frac{1}{100}$.

We wish to find $P(50 < X < 150)$. Then,

$$P(50 < X < 150) = \int_{50}^{150} \frac{1}{100} e^{-\frac{x}{100}} dx = -\frac{1}{100} 100 e^{-\frac{x}{100}}\Big|_{50}^{150} = -e^{-\frac{3}{2}} + e^{-\frac{1}{2}} \approx .383$$

b) We wish to find $P(X < 100)$.

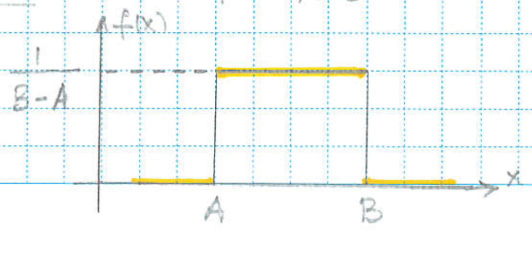
$$P(X < 100) = \int_{0}^{100} \frac{1}{100} e^{-\frac{x}{100}} dx = -e^{-\frac{x}{100}}\Big|_0^{100} = -\frac{1}{e} + 1 \approx .632$$

In other words, approximately 63.2 percent of the time, a computer will fail before registering 100 hours of use.

Definition. A continuous rv X is said to have uniform distribution on the interval $[A, B]$ if the PDF of X is

$$f(x; A, B) = \frac{1}{B-A}, \quad A \leq x \leq B.$$

NOTATION. $X \sim \text{Unif}[A, B]$: X has a uniform distribution on $[A, B]$.



→ The PDF of a uniform rv.

Example. A gambler spins a wheel of fortune, continuously calibrated between 0 and 1, and observes the resulting number. Assuming that all subintervals of $[0,1]$ of the same length are equally likely. This exp. can be modeled with PDF

$$f(x) = \begin{cases} c & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant c. To find c,

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_0^1 c\,dx = cx \Big|_0^1 = c \Rightarrow c = 1$$

More generally, we can consider

$$f(x) = \begin{cases} c & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

and this gives

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_a^b c\,dx = cx \Big|_a^b = c(b-a) \Rightarrow c = \frac{1}{b-a},$$

which represents the PDF of a uniform rv.

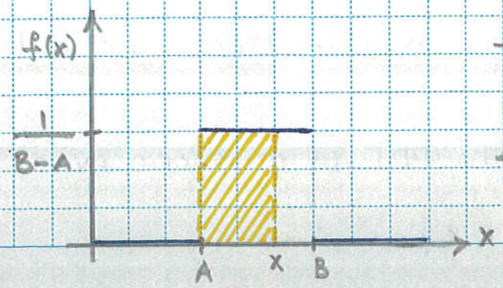
The cumulative distribution function

Definition. The cumulative distribution function $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy.$$

For each x, $F(x)$ is the area under the density curve to the left of x.

Example. Let X, the thickness of a membrane, have a uniform distribution on $[A, B]$. The density function is shown in Figure below:



→ For $x < A$, $F(x) = 0$ (there is no area under the graph to the left of x.)

→ For $x > B$, $F(x) = 1$ (all the area is accumulated to the left of x.)

NEXT PAGE

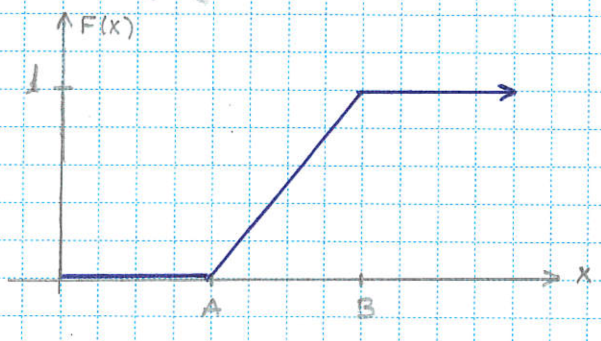
→ For $A < x < B$, $F(x) = \int_{-\infty}^{x} f(y)dy = \int_{A}^{x} \frac{1}{B-A} dy = \frac{1}{B-A} y \Big|_{A}^{x} = \frac{x}{B-A} - \frac{A}{B-A} = \frac{x-A}{B-A}$.

Function is zero outside its support
In other words, for $(-\infty, A)$, area is 0.

Then the cdf is written as

$$F(x) = \begin{cases} 0 & , x < A \\ \frac{x-A}{B-A} & , A \le x < B \\ 1 & , x \geq B \end{cases}$$

and the corresponding graph is



Using $F(x)$ to compute probabilities

The probabilities of varius intervals can be computed from a formula.

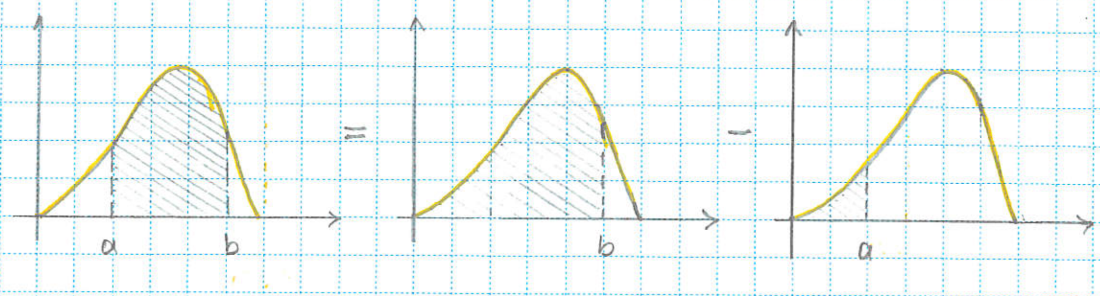
Proposition. Let X be a continuous rv with PDF $f(x)$ and CDF $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \le X \le b) = F(b) - F(a). \quad \dots (*)$$

Equation $(*)$ can be illustrated as below:



<u>Obtaining f(x) from F(x)</u>

The following result is based on the Fundamental Theorem of Calculus.

<u>Proposition.</u> If X is a continuous rv with PDF $f(x)$ and CDF $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$.

<u>Example.</u> If X is continuous rv with CDF $F(x)$ and PDF $f(x)$, find the PDF of $Y = 2X$.

<u>Solution.</u> Let us derive and then differentiate :

$$F(y) = P(Y \le a) = P(2X \le a) = P\left(X \le \frac{a}{2}\right) = F\left(\frac{a}{2}\right)$$

By differentiation,

$$F'(y) = \frac{d}{da} F\left(\frac{a}{2}\right) = \frac{1}{2} F'\left(\frac{a}{2}\right) = \frac{1}{2} f\left(\frac{a}{2}\right) = f(y).$$

<u>Percentiles of a continuous distribution</u>

<u>Definition.</u> Let p be a number between 0 and 1. The (100p)th percentile (or <u>pth quantile</u>) of the dist. of a cont. rv X, denoted by η_p, is defined by

$$p = F(\eta_p) = \int_{-\infty}^{\eta_p} f(y)\,dy.$$

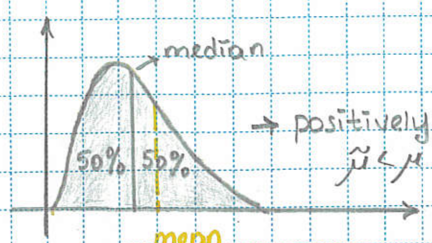
If the inverse of $F(x)$ exists, we can write

$$\eta_p = F^{-1}(p).$$

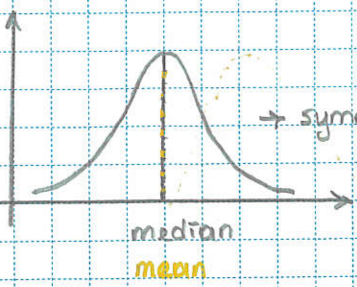
In particular, the <u>median</u> is the 50th percentile ($p = .5$); that is

$$\tilde{\mu} = \eta_{.5} = F^{-1}(.5)$$

This means that, half of the total area under the density curve lies to the left of the median, and half lies to the right.



→ positively skewed dist.
$\tilde{\mu} < \mu$



→ symmetric dist.
$\mu = \tilde{\mu}$

<u>Example.</u> Suppose X takes values $\{1, 2, 3, 30\}$. Then, $\mu = \frac{1+2+3+30}{4} = 9$,

and $\tilde{\mu} = \frac{2+3}{2} = 2.5$.

4.2. Expected values and moment generating functions

Expected values

Definition. The expected value (or mean) of a continuous rv X with PDF $f(x)$ is

$$\mu = \mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx.$$

(exp. val.)

This mean will exist provided that $\int_{-\infty}^{\infty} |x| \cdot f(x)\,dx < \infty$.

Example. The density function of X is given by

$$f(x) = \begin{cases} 1 & , 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $E[e^X]$.

Solution. $E[e^X] = \int_{-\infty}^{\infty} e^x \cdot f(x)\,dx = \int_0^1 e^x \cdot 1\,dx = e^x \Big|_0^1 = e - 1.$

Law of the unconscious statistician. If X is a continuous rv with PDF $f(x)$ and $h(x)$ is any function of X, then

$$\mu_{h(X)} = E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$$

(exp. val.)

This mean will exist provided that $\int_{-\infty}^{\infty} |h(x)| f(x)\,dx < \infty$.

Example. The variation in a certain electrical current source X (in milliamps) can be modeled by the PDF $f(x) = 1.25 - .25x$, $2 \le x \le 4$.

a) Find the average current.

b) If this current passes through a 220-ohm resistor, the resulting power (in microwatts) is given by $h(X) = (\text{current})^2(\text{resistance}) = 220X^2$. Find the expected power.

Solution. a) $E(X) = \int_2^4 x \cdot (1.25 - .25x)\,dx = \left[1.25\frac{x^2}{2} - .25\frac{x^3}{3}\right]_2^4 = \left[\frac{5}{8}x^2 - \frac{1}{12}x^3\right]_2^4$

$= \left(10 - \frac{16}{3}\right) - \left(\frac{5}{2} - \frac{2}{3}\right)$

$= \frac{15}{2} - \frac{14}{3} = \frac{45-28}{6} = \frac{17}{6}$

$= 2.833\text{mA}$

b) $E[h(X)] = E[220X^2] = \int_2^4 220x^2 \cdot (1.25 - .25x)\,dx$

$= 220\left[1.25\frac{x^3}{3} - .25\frac{x^4}{4}\right]_2^4$

$= \frac{5500}{3} = 1833.3$ microwatts.

......./......./..........

Definition. The variance of a continuous rv X with PDF $f(x)$ and mean μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx = E\left[(X-\mu)^2\right]$$

The standard deviation of X is $SD(X) = \sigma_X = \sqrt{V(X)}$.

Example. The time elapsed, in minutes, between the placement of an order of pizza and its delivery is random with the density function

$$f(x) = \begin{cases} 1/15 & \text{if } 25 < x < 40 \\ 0 & \text{otherwise} \end{cases}$$

a) Determine the mean and standard deviation of the time it takes for the pizza shop to deliver pizza.

b) Suppose that it takes 12 minutes for the pizza shop to bake pizza. Determine the mean and standard deviation of the time it takes for the delivery person to deliver pizza.

Solution. a) Assume X: the time elapsed between the placement and the delivery.

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{25}^{40} x \cdot \frac{1}{15}\,dx = \frac{x^2}{30}\Big|_{25}^{40} = \frac{40.40}{30} - \frac{25.25}{30} = \frac{160}{3} - \frac{125}{6} = \frac{320-125}{6} = \frac{195}{6} = 32.5$$

$$E(X^2) = \int_{25}^{40} x^2 \frac{1}{15}\,dx = \frac{1}{15}\frac{x^3}{3}\Big|_{25}^{40} = \frac{1}{45}\left[64000 - 15625\right] = \frac{48375}{45} = 1075.$$

$$V(X) = E(X^2) - \left[E(X)\right]^2 = 1075 - (32.5)^2 = 18.75$$

$$\sigma_X = SD(X) = \sqrt{V(X)} = \sqrt{18.75} = 4.33.$$

b) The time between the placement and the delivery was X in (a). Then, the time between delivery person take and deliver it is $X-12$. Then,

$$E(X-12) = E(X) - 12 = 32.5 - 12 = 20.5$$

$$V(X-12) = V(X) = 18.75$$

$$SD(X-12) = \sqrt{18.75} = 4.33$$

........./........./...........

Proposition. Let X be a continuous rv with PDF $f(x)$, mean μ, and standard deviation σ. Then the following properties hold.

1. **Variance shortcut**:

$$V(X) = E(X^2) - [E(X)]^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - \left(\int_{-\infty}^{\infty} x \cdot f(x)\,dx\right)^2$$

2. **Linearity of expectation.** For any functions $h_1(X)$ and $h_2(X)$ and any constants a_1, a_2, and b,

$$E[a_1 h_1(X) + a_2 h_2(X) + b] = a_1 E_1[h_1(X)] + a_2 E_2[h_2(X)] + b$$

3. **Rescaling.** For any constants a and b,

$$E(aX+b) = aE(X) + b$$

$$V(aX+b) = a^2 V(X)$$

$$SD(aX+b) = |a| \cdot SD(X)$$

Approximating the mean value and standard deviation

Next result is about the mean and the variance of a nonlinear function.

Proposition (The delta Method). Suppose $h(x)$ is differentiable and that its derivative evaluated at μ satisfies $h'(\mu) \neq 0$. If $V(X)$ is small (X is concentrated near μ), the mean and variance of $Y = h(X)$ can be approximated as follows:

$$E[h(X)] \approx h(\mu) \qquad \cdots (1)$$

$$V[h(X)] \approx [h'(\mu)]^2 V(X) \qquad \cdots (2)$$

To obtain (1), consider a first order Taylor series expansion of $h(X)$ about μ:

$$Y = h(X) \approx h(\mu) + h'(\mu)(X-\mu) \qquad \cdots (3)$$

$$\Rightarrow E[h(X)] \approx E[h(\mu) + h'(\mu)(X-\mu)]$$

$$\approx h(\mu)$$

$h(\mu)$ and $h'(\mu)$ are const. That is, $E[h(\mu)] = h(\mu)$, $E[h'(\mu)] = h'(\mu)$, $E[X] = \mu$

To obtain (2), considering the expansion (3),

$$V[h(X)] \approx V[h(\mu) + h'(\mu)(X-\mu)]$$

$$\approx \underbrace{V[h(\mu)]}_{0} + [h'(\mu)]^2 V(X-\mu)$$

$$\approx [h'(\mu)]^2 V(X).$$

(by shortcut formula)
$V(X-\mu) = E[(X-\mu)^2] - [\underbrace{E(X-\mu)}_{0}]^2 = V(X)$

Example. A chemistry student measures the mass m and volume X of an aluminum chunk and calculates the density as $Y = h(X) = \frac{m}{X}$. The mass is measured very accurately and can be considered constant. The volume X has some measurement uncertainty with standard deviation $\sigma_X = .3\,cm^3$.

a) Estimate the σ_Y in terms of m, μ_X and σ_X.

b) For a chunk with measured values $m = 18.19\,g$ and $X = 6.6\,cm^3$, compute

i) the estimated density Y

ii) the estimated standard deviation of Y

c) Compare the estimated density and standard deviation with the official density of aluminum, $2.70\,g/cm^3$.

Solution. We are given $Y = h(X) = \frac{m}{X}$. Then, $h'(X) = -\frac{m}{X^2}$.

a) First, find the variance. By the delta method: $V(Y) \approx [h'(\mu)]^2 \cdot \sigma_X^2$

Then, we have $V(Y) \approx \left(-\frac{m}{\mu_X^2}\right)^2 \cdot \sigma_X^2$, and $\sigma_Y \approx \frac{m}{\mu_X^2} \cdot \sigma_X$.

b) The estimated density : $Y = \frac{18.19}{6.6} = 2.76\,g/cm^3$.

The estimated standard deviation : $\sigma_Y \approx \frac{m}{\mu_X^2} \cdot \sigma_X = \frac{18.19}{(6.6)^2}(.3) = .125$

c) Estimated density is $2.76 \pm .125$, which means that the true density will be within the interval $[2.63, 2.88]$. In fact, the official density, $2.70\,g/cm^3$ lies within this range.

Moment generating functions

Definition. The mgf of a cont. rv. X is

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx.$$

* We also have, when $t = 0$,

$$M_X(0) = E(e^{0.X}) = \int_{-\infty}^{\infty} e^{0.x} f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

* (Uniquiness property). Two distributions have the same PDF iff they have the same mgf.

* The rth moment of a continuous rv with mgf $M_X(t)$ is given by

$$E(X^r) = M_X^{(r)}(0),$$

the rth derivative of the mgf w.r.t. t evaluated at $t=0$, if the mgf exists.

* If X has the mgf $M_X(t)$, then $M_Y(t) = e^{bt} M_X(at)$, where $Y = aX+b$.

Example. Let X have a uniform dist. on the interval $[A, B]$, so its PDF is $f(x) = \frac{1}{B-A}$, $A \le x \le B$, $f(x)=0$ otherwise. Show that the MGF of X is

$$M_X(t) = \frac{e^{Bt} - e^{At}}{(B-A)t}, \quad t \neq 0.$$

Solution. For $t=0$, $M_X(0) = E[e^{0.X}] = \int_{-\infty}^{\infty} e^{0.x} f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 1.$

For $t \neq 0$,

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx = \int_A^B e^{tx} \cdot \frac{1}{B-A}\,dx = \frac{1}{B-A} \frac{e^{tx}}{t}\Big|_A^B = \frac{1}{B-A} \frac{e^{Bt} - e^{At}}{t} = \frac{e^{Bt} - e^{At}}{(B-A)t}.$$

Example. If the PDF of a measurement error X is $f(x) = .5e^{-|x|}$, $-\infty < x < \infty$, show that $M_X(t) = \frac{1}{1-t^2}$ for $|t| < 1$.

Solution. We know that $|x| = -x$ for $x<0$ and $|x| = x$ for $x > 0$. Then,

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} (.5e^{-|x|})\,dx$$

$$= \int_{-\infty}^{0} e^{tx} (.5e^{x})\,dx + \int_0^{\infty} e^{tx} (.5e^{-x})\,dx$$

$$= \underbrace{.5\int_{-\infty}^{0} e^{x(t+1)}\,dx}_{I_1} + \underbrace{.5\int_0^{\infty} e^{x(t-1)}\,dx}_{I_2}$$

I_1 is convergent if $(t+1)$ is positive $(t > -1)$: $I_1 = .5\int_{-\infty}^{0} e^{x(t+1)}\,dx = \frac{.5}{t+1} e^{x(t+1)}\Big|_{-\infty}^{0} = \frac{.5}{t+1}$.

I_2 is convergent if $(t-1)$ is negative $(t<1)$: $I_2 = .5\int_0^{\infty} e^{x(t-1)}\,dx = \frac{.5}{t-1} e^{x(t+1)}\Big|_0^{\infty} = -\frac{.5}{t-1}$.

$M_X(t) = I_1 + I_2 = \frac{.5}{t+1} - \frac{.5}{t-1} = \frac{.5}{1+t} + \frac{.5}{1-t} = \frac{1}{1-t^2}$ for $-1 < t < 1$.

Example. The weekly demand for propane gas (in 1000s of gallons) from a particular facility is an rv X with PDF

$$f(x) = 2\left(1 - \frac{1}{x^2}\right) \quad 1 \le x \le 2$$

a) Compute the CDF of X.

b) Obtain an expression for the $(100p)$th percentile. What is the value of $\tilde{\mu}$?

c) Compute $E(X)$. How do the mean and median of this distribution compare?

Solution. a) For $1 \le x \le 2$, $F(x) = \int_1^x f(y)dy = 2\int_1^x \left(1 - \frac{1}{y^2}\right)dy = 2\left[y + \frac{1}{y}\right]_1^x = 2\left(x + \frac{1}{x}\right) - 4$

Then, the CDF is

$$F(x) = \begin{cases} 0 & , x < 1 \\ 2\left(x + \frac{1}{x}\right) - 4 & , 1 \le x \le 2 \\ 1 & , x > 2 \end{cases}$$

b) Set $F(x) = p$. Then,

$p = 2\left(x + \frac{1}{x}\right) - 4 \Rightarrow p - 2x - \frac{2}{x} + 4 = 0 \Rightarrow px - 2x^2 - 2 + 4x = 0 \Rightarrow 2x^2 - x(p+4) + 2 = 0$

and $\eta_p = x = \frac{p + 4 + \sqrt{(p+4)^2 - 4.2.2}}{2.2} = \frac{p + 4 + \sqrt{p^2 + 8p + 16 - 16}}{4} = \frac{p + 4 + \sqrt{p^2 + 8p}}{4}$.

The other root is not included in $[1, 2]$.

To find the median, $\tilde{\mu}$, set $p = .5$: $\eta_{.5} = \frac{.5 + 4 + \sqrt{.25 + 4}}{4} = \frac{.5 + 4 + \sqrt{4.25}}{4}$

$= \frac{4.5 + 2.05}{4}$

$= \frac{6.55}{4}$

$= \boxed{1.640}$ ... (1)

The median weekly demand for gas is 1640 gallons.

c) $E(X) = \int_1^2 x \cdot 2\left(1 - \frac{1}{x^2}\right)dx = 2\int_1^2 \left(x - \frac{1}{x}\right)dx = 2\left[\frac{x^2}{2} - \ln x\right]_1^2$

$= 2\left[(2 - \ln 2) - \frac{1}{2} - 0\right]$

$= \frac{7}{2} - \ln 2$

$\approx \boxed{1.614}$ ... (2)

Consider (1) and (2) together, we see that μ is lower than $\tilde{\mu}$ because the distribution is negatively-skewed.